Received: 29 January 2024 | Accepted: 27 May 2024

DOI: 10.1002/cdt3.137

Chronic Diseases® and Translational Medicine

REVIEW

Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges

Ebenezer Afrifa-Yamoah[1] | Eric Adua[2,3] | Emmanuel Peprah-Yamoah[4] | Enoch O. Anto[3,5] | Victor Opoku-Yamoah[6] | Emmanuel Acheampong[7] | Michael J. Macartney[8] | Rashid Hashmi[2]

[1]School of Science, Edith Cowan University, Joondalup, Western Australia, Australia

[2]Rural Clinical School, Medicine and Health, University of New South Wales, Sydney, New South Wales, Australia

[3]School of Medical and Health Sciences, Edith Cowan University, Joondalup, Western Australia, Australia

[4]Teva Pharmaceuticals, Salt Lake City, Utah, USA

[5]Department of Medical Diagnostics, Faculty of Allied Health Sciences, College of Health Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[6]School of Optometry and Vision Science, University of Waterloo, Waterloo, Ontario, Canada

[7]Department of Genetics and Genome Biology, Leicester Cancer Research Centre, University of Leicester, Leicester, UK

[8]Faculty of Science Medicine and Health, University of Wollongong, Wollongong, New South Wales, Australia

Correspondence

Ebenezer Afrifa-Yamoah, School of Science, Edith Cowan University, Joondalup, WA, Australia.
Email: e.afrifayamoah@ecu.edu.au

Eric Adua, Rural Clinical School, Medicine and Health, University of New South Wales, Sydney, NSW, Australia.
Email: e.adua@unsw.edu.au

Abstract

Chronic diseases such as heart disease, cancer, and diabetes are leading drivers of mortality worldwide, underscoring the need for improved efforts around early detection and prediction. The pathophysiology and management of chronic diseases have benefitted from emerging fields in molecular biology like genomics, transcriptomics, proteomics, glycomics, and lipidomics. The complex biomarker and mechanistic data from these "omics" studies present analytical and interpretive challenges, especially for traditional statistical methods. Machine learning (ML) techniques offer considerable promise in unlocking new pathways for data-driven chronic disease risk assessment and prognosis. This review provides a comprehensive overview of state-of-the-art applications of ML algorithms for chronic disease detection and prediction across datasets, including medical imaging, genomics, wearables, and electronic health records. Specifically, we review and synthesize key studies leveraging major ML approaches ranging from traditional techniques such as logistic regression and random forests to modern deep learning neural network architectures. We consolidate existing literature to date around ML for chronic disease prediction to synthesize major trends and trajectories that may inform both future research and clinical translation efforts in this growing field. While highlighting the critical innovations and successes emerging in this space, we identify the key challenges and limitations that remain to be addressed. Finally, we discuss pathways forward toward scalable, equitable, and clinically implementable ML solutions for transforming chronic disease screening and prevention.

KEYWORDS

big data, chronic diseases, disease prediction, machine learning algorithms, OMICs data

Highlights

- Machine learning (ML) shows promise for early detection and prediction of chronic diseases.

Ebenezer Afrifa-Yamoah and Eric Adua contributed equally to this study.

- Complex "omics" data from genomics, proteomics, and other fields inform ML models.
- The review covers key ML approaches applied to diverse datasets for chronic diseases.
- Innovations and successes are highlighted, along with challenges and limitations.
- Pathways toward scalable, equitable, and clinically implementable ML solutions are discussed.

1 | INTRODUCTION

Chronic diseases, including type II diabetes mellitus (T2DM), cardiovascular disease, cancer, and chronic respiratory diseases, are negatively impacting many lives worldwide. Defined as a lifelong condition that requires perpetual medical attention, chronic diseases lower the quality of life of individuals, and presently, according to the World Health Organization (WHO), kill an estimated 41 million people worldwide.[1] Moreover, chronic diseases increase the economic burden of sufferers, rendering them financially impoverished due to frequent hospitalization and rising costs of medications. More disconcerting are people living in resource-limited countries, where 77% of all chronic disease deaths occur. It is well established that these diseases are potentially preventable with lifestyle modification, which, when combined with medication, may prevent disease progression. However, these efforts appear to be inadequate, as the burden of disease is still rising, and the available therapies are not curative but only ameliorate disease-related symptoms.[2]

Research has advanced to shift concepts from reactive medicine to proactive personalized prevention, risk stratification, diagnosis, and treatment for individual patients.[3–5] This approach identifies those at risk of developing diseases by collecting and analyzing big data on lifestyle, demographics, anthropometrics, family history, and more.[4,5] In recent years, stratification and diagnosis have utilized emerging fields, including genomics, transcriptomics, proteomics, glycomics, and lipidomics. OMICS datasets have revealed potential biomarkers and provided insights into the aetiologies and progressions of several diseases. OMICS data include numbers, images, audio, video, and text, often collected, extracted, transcribed, tracked, and segmented. Despite the impact of multiomics for early diagnosis, significant bottlenecks exist. These data generation procedures are labor-intensive and require expensive equipment. Moreover, the quantity of unlabeled data generated from OMICS processes in our tech-driven era is enormous.

Unlabeled data must be preprocessed and generally has limited analytical utility. A key step in expanding the use of OMICS data for efficient modeling and prediction is accurate data labeling. Data labeling involves tagging or annotating attributes, properties, or classifications for easy identification and reference. While potentially expensive and time-consuming, it aims to ensure accurate, consistent, scalable, high-quality tagging. In all OMICS applications, such as studying complex biological systems and predicting biomarkers related to human health,[6] data labels serve as learning targets in machine learning (ML) and artificial intelligence (AI). For instance, deep neural networks (NNs) perform best for named entity recognition and semantic role labeling when data is labeled.[7] The various OMICS data must be classified through data labeling, and the precise interpretation of the subsequent complex data requires advanced computational algorithms.

ML offers promising techniques to unlock insights from heterogenous data sources that can enable earlier detection of chronic diseases and better predict progression and prognosis.[8] For example, techniques such as NNs can model complex nonlinear relationships in data, while random forest (RF) models provide additional transparency into important variables through feature importance score measures.[7–10] Over the past decade, applications of ML for chronic disease detection and prediction have rapidly expanded. The growth has been further fueled by the fourth and fifth industrial revolutions, which have emphasized the significance of cloud computing and AI, both powered by ML algorithms.[9–11] As the medical field continues to generate an increasing amount and variety of data, reliance on data scientists' expertise becomes crucial for making sense of this "big data." Collaborations between data scientists and medical professionals have resulted in substantial progress in developing tailored methodologies for analyzing different types of medical data.[9,10] ML has become an essential tool for finding optimized solutions in disease prognosis and diagnosis, playing a central role in most medical advancements of the past decade.[1–3,6] Based on the applications of ML algorithms in medicine, perspectives on the pathophysiology of chronic conditions have improved, leading to a better understanding of their initiation and progression (Figure 1). A common application is feature selection, where important biomarkers are linked to various chronic diseases, including cardiovascular, T2DM, respiratory diseases, and cancers.[4]

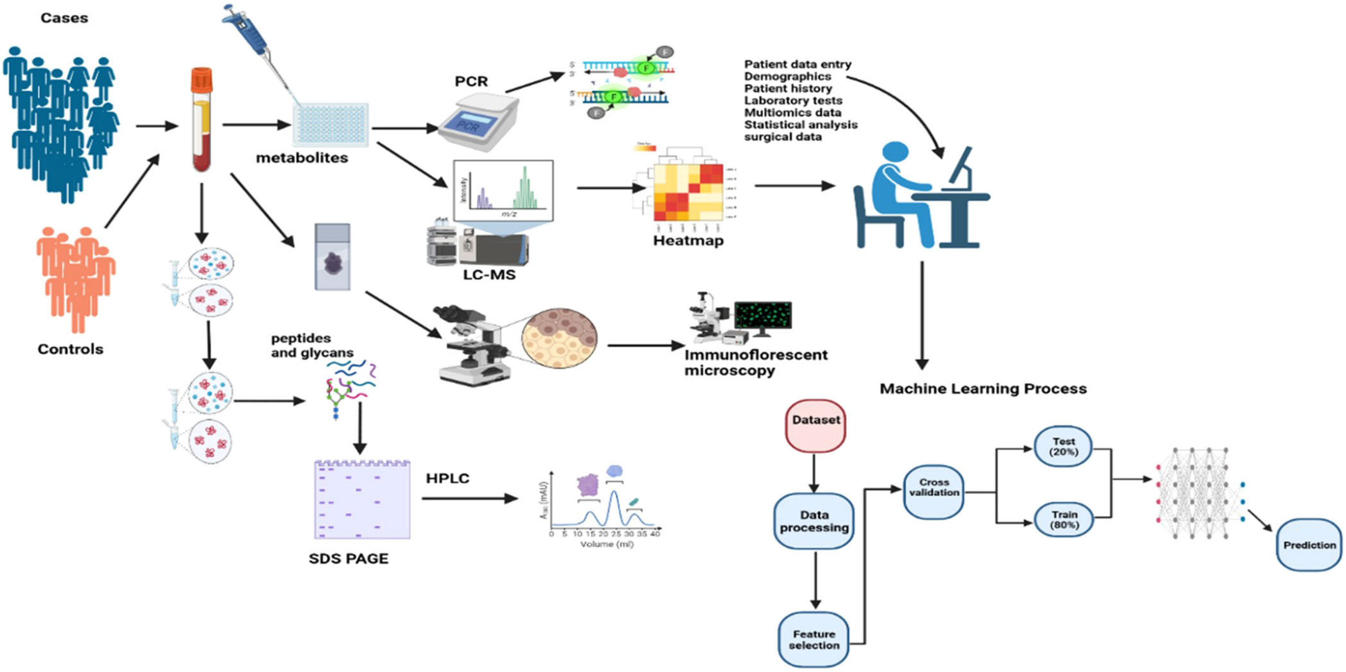

FIGURE 1 Laboratory methods and application of machine learning for disease prediction and diagnosis. Diagnosis of diseases begins with screening a population, sample collection, processing, and quantitation with molecular/analytical methods (e.g., polymerase chain reaction [PCR], liquid chromatography mass spectrophotometry [LC-MS], high-performance liquid chromatography [HPLC], sodium dodecyl sulfate polyacrylamide gel electrophoresis [SDS-PAGE]), and immunofluorescent macroscopy among others. Machine learning algorithms can discover patterns in the data generated from analytical methods and make predictions.

Although ML-powered AI holds promise to transform medical diagnosis and treatment by rapidly analyzing massive data, it faces ethical hurdles and practical constraints. This extensive overview study evaluates the applications of ML algorithms to predict and manage a range of chronic diseases by assessing the accuracy and utility of tasks such as risk stratification, personalized treatment plans, and disease progression monitoring while considering the unique challenges of the pathophysiological processes of these chronic conditions. We further discuss the ethical implications of utilizing ML algorithms for medical diagnosis, treatment, and risk assessment. We highlight the ethical issues around transparency and explainability of model decisions, potential biases in data and algorithms, accountability for outcomes, privacy trade-offs, and impacts on the doctor–patient relationship, and provide guidance on responsible ML implementation that upholds principles of beneficence, nonmaleficence, and fairness.

2 | BRIEF OVERVIEW OF ML ALGORITHMS

ML is a proponent of AI, and it is a process where computers adapt and learn from experience and build models without prior instructions or supervision.[3] The computer receives data and utilizes mathematical equations to predict outcomes.[3] ML algorithms can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning approaches.

Supervised learning algorithms train models to make predictions based on labeled input-output pairs in the training data. Popular supervised learning algorithms include linear regression, logistic regression (LR), support vector machines (SVMs), decision trees (DTs), RFs, and NNs.[11] Linear regression is used for predicting continuous values, while LR makes classifications based on logistic sigmoid outputs.[11] SVMs find optimal decision boundaries to categorize data points.[12] DTs make step-by-step binary splits on the features, while RFs improve performance by averaging predictions across many DTs.[13] NNs contain interconnected layers of nodes inspired by biological neurons, learning to make nonlinear transformations of the input.[14] The advantages and disadvantages of these ML algorithms are well discussed in studies.[11–14]

Unsupervised learning analyses unlabeled data to find hidden patterns. Key unsupervised learning techniques include clustering algorithms like k-means, hierarchical clustering, and Gaussian mixture models, which group data points based on similarity.[15] Dimensionality reduction techniques such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) identify important dimensions capturing most of the variance.[16,17]

Reinforcement learning trains models to optimize behaviors based on rewards and penalties. Instead of exact training labels, the model learns via trial-and-error interactions with a dynamic environment. Popular reinforcement learning techniques include Q-learning and policy gradient methods.[18]

In addition to the established algorithms above, cutting-edge ML techniques offer new horizons that are gaining traction in predictive modeling. Examples include generative models like generative adversarial networks (GANs) and variational autoencoders (VAEs) that can create synthetic, realistic data for domains with limited examples like medical images.[19] Causal ML techniques to determine causal predictors of disease and outcomes for better-guiding interventions.[20] Multimodal ensembles synthesizing insights across data types into unified models. Such emerging methods can unlock new capabilities around scarce data augmentation, optimized dynamic treatment regimens, and robust causal insights. Overall, ML encompasses a wide range of algorithms with different strengths suited for various data analysis tasks. Careful selection and tuning of algorithms are necessary to build effective ML systems.

3 | UNDERSTANDING THE PATHOPHYSIOLOGICAL PROCESSES OF CHRONIC DISEASES VIA ML ALGORITHMS

Chronic diseases often have complex pathophysiologies involving multiple risk factors and pathogenesis pathways.[21] These diseases develop gradually over many years and involve complex interactions between genetics, lifestyle, and environmental factors that influence disease pathogenesis.[22] For example, T2DM results from lifestyle factors such as diet and exercise interacting with genetic predispositions that lead to insulin resistance and impaired glucose metabolism over time.[3,23] Similarly, cardiovascular disease develops from atherosclerotic changes in arteries over decades. This process is driven by lipid disorders, hypertension, diabetes, smoking, and other risk factors.[24] Chronic disease pathophysiology unfolds through intricate molecular alterations over a protracted period, reflecting the interplay between genetic and nongenetic factors. Understanding these complex mechanistic pathways and interactions is key to enhancing early diagnosis, prediction, and personalized management of chronic conditions.

ML techniques offer opportunities to better understand these complex pathophysiological processes and improve disease prediction.[25] By analyzing large, multidimensional datasets, algorithms can identify patterns linking risk factors to early biological changes that presage chronic disease development.[26] For cardiovascular disease, ML models integrating genomic, proteomic, clinical, and imaging data may enable earlier detection of high-risk plaque features or subclasses at greater risk for myocardial infarction.[26] In oncology, algorithms correlating molecular tumor profiles with clinical outcomes could refine prognostic models for recurrence risk in breast cancer or disease progression in prostate cancer.[27] ML techniques like NNs and RFs can analyze large multivariate datasets to uncover predictive patterns in risk factors, biomarkers, and clinical signs that foreshadow chronic disease onset.[8] By modeling complex nonlinear relationships, these algorithms can provide personalized risk scores and identify high-risk subgroups for preventive interventions. Deep learning methods that extract high-level data representations also show promise for integrating diverse omics, physiological, and health record data for disease prediction.[8,24–27]

In the sections below, we discuss the complex pathophysiology underlying several major chronic diseases, including T2DM, chronic kidney disease (CKD), cardiovascular diseases (CVDs), cancers, chronic respiratory diseases, and inflammatory conditions. For each disease, we highlight how the intricate molecular changes unfolding over many years lead to pathogenesis and complications. We also examine recent applications of ML to analyze multidimensional data for these conditions. Specifically, we review how advanced algorithms integrating omics, imaging, electronic health record (EHR), and other disparate datasets have provided novel insights into prediction, prognosis, and personalized management. This discussion illustrates the growing complexity of biomedical data relevant to chronic diseases and demonstrates the utility of ML approaches for elucidating disease mechanisms, refining risk stratification, and guiding clinical decision-making. Overall, sophisticated modeling of heterogeneous data types promises to deepen our understanding of chronic disease trajectories and enable more precise, individualized care.

4 | PATHOPHYSIOLOGY OF T2DM

T2DM arises from the body's inability to metabolize glucose properly, leading to the accumulation of glucose in the blood.[28] In 2015 alone, over a million people died from the disease, as per the International Diabetes Federation. While medications and lifestyle changes have enhanced longevity and improved quality of life for many patients, the complexity and heterogeneity of T2DM continue to impede a cure.[28]

T2DM is caused by both insulin resistance in tissues and deficient insulin secretion from the pancreatic beta (β) cells.[29] Initially, beta cells compensate by secreting more insulin. However, over time, insulin production declines as the β cells fatigue and undergo apoptosis (Figure 2). Compared to controls, those with impaired

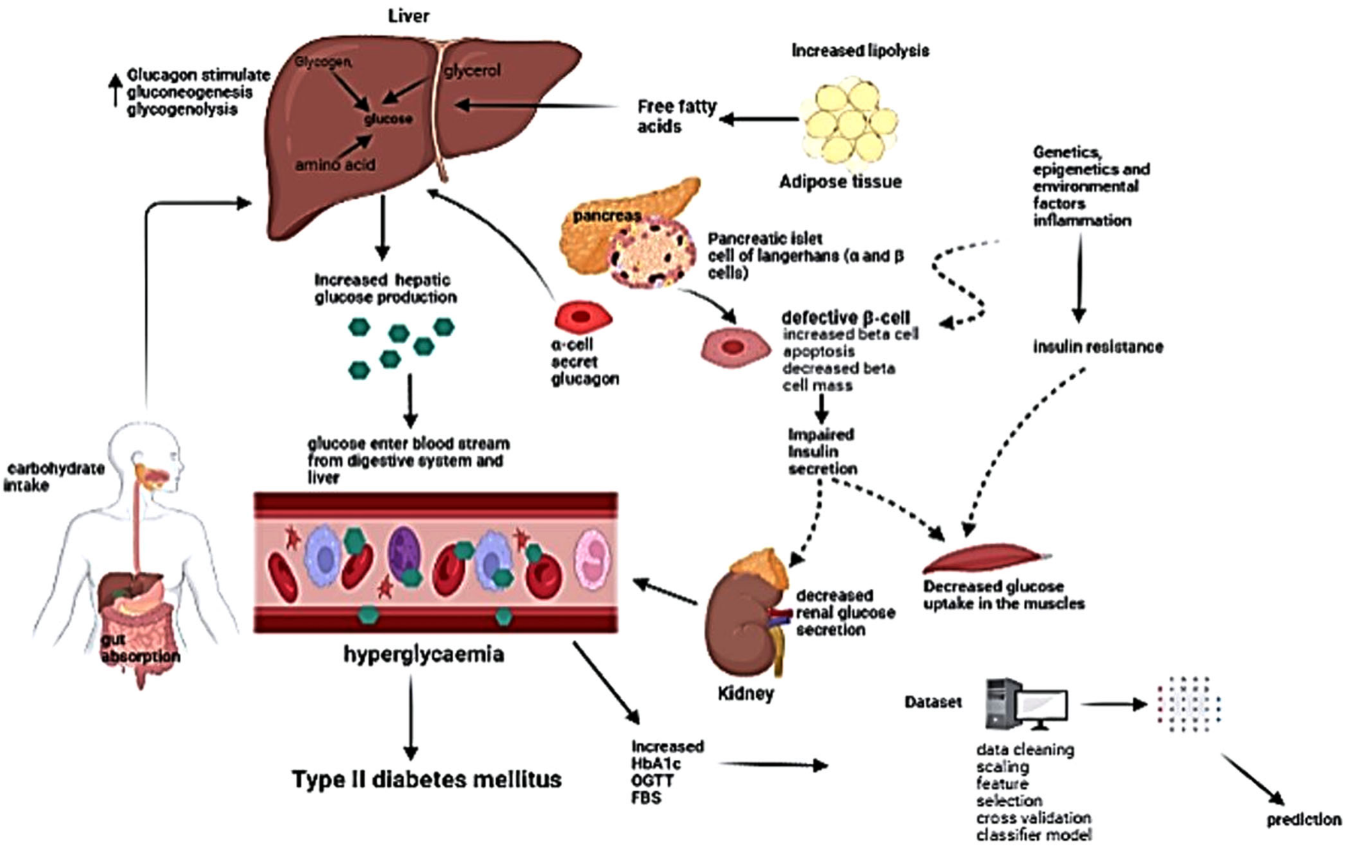

FIGURE 2 Pathophysiology of Type II diabetes. Alpha and beta cells of the pancreas secret insulin and glucagon, respectively. These two hormones interact with the liver to regulate blood sugar levels. While glucagon stimulates glycogenolysis and gluconeogenesis, leading to a rise in sugar levels in the blood, insulin promotes glycogenesis and glucose uptake in skeletal muscles and other tissues. In diabetes, there is either impaired insulin secretion or insulin resistance, resulting in elevated plasma concentration of sugar.

fasting glucose and T2DM have shown 40% and 63% loss of β-cell volume, respectively. Additionally, studies reveal that T2DM β cells frequently die by apoptosis.[29] While defective beta cells strongly associate with T2DM pathogenesis, alpha cell destruction also correlates with disease progression.[29–31] Approximately 30% of T2DM cases result from genetic abnormalities [32]. Genome-wide association studies have identified over 176 loci implicated in T2DM, including CDKAL1, SLC30A8, HHEX, CDKN2A/B, IGF2BP2, and others.[33–35] Mutations in genes coding for glucokinase, insulin receptors, and mitochondria also contribute.[32] For example, Unoki et al. identified a mutation in the KCNQ1 gene involved in insulin secretion.[36]

Insulin resistance results in decreased glucose uptake in skeletal muscles and reduced kidney reabsorption. This lack of glucose reabsorption leads to glucosuria, as glucose is excreted in the urine. Consequently, water follows the excreted glucose. In healthy individuals, pancreatic beta cells synthesize proinsulin and ensure proper insulin formation during posttranslational modification. These cells respond to hormonal and intracellular signals by stimulating nutrient-stimulating factor production. However, in diabetes, nutrient deprivation in beta cells increases hunger.[29]

T2DM complications include hypoglycemia, which can occur in those receiving sulfonylurea or insulin who engage in excessive physical activity. A less common complication is diabetic ketoacidosis, wherein triglycerides and amino acids are broken down for energy instead of glucose.[29] Glucagon stimulates ketone formation from free fatty acids, also utilizing free fatty acids and alanine as substrates for hepatic glucose production via gluconeogenesis. Specifically, this generates acetoacetic and beta-hydroxybutyric acids, causing abdominal pain, vomiting, and other symptoms. Insulin deficiency can also lead to a hyperglycaemic hyperosmotic state.[32,34] While insulin normally inhibits lipolysis, this is countered by catecholamines, cortisol, and growth hormones. The resulting high serum glucose concentration and increased diuresis raise osmotic pressure. As the kidneys eliminate free water and electrolytes, dehydration increases. Other microvascular and macrovascular T2DM complications include CVD, stroke, diabetic retinopathy, nephropathy, neuropathy, and nonalcoholic fatty liver disease,[34,35] but these are beyond the scope of this review.

5 | THE APPLICATION OF ML ALGORITHMS FOR DETECTION AND PREDICTION IN T2DM

Early detection of diseases is crucial for effective intervention, but often initial symptoms go unnoticed until it is too late. High-throughput computational techniques offer a promising avenue for fast-tracking target identification and developing improved therapies. In the context of diabetes, various ML approaches have been applied to distinguish between healthy individuals and those at risk.

Polat and Günes[37] utilized PCA to distinguish diabetes from healthy individuals. Farran et al.[38] applied a K-neural network (KNN) to identify increased risks for hypertension (94%) and diabetes (75%). Lai et al.[39] predicted diabetes using LR, RF, DTs, and a gradient boosting machine (GBM), highlighting GBM and LR as superior algorithms. Ravaut employed a gradient boosting DT model to predict 3-year complication outcomes due to diabetes,[40] while Perveen et al.[41] used a hidden Markov Model to identify individuals likely to develop T2DM in 8 years.

Wang et al.[42] compared multivariate LR and artificial neural network (ANN) on T2DM data, showing that ANN had a higher predictive value. However, another study favored RF over DT, NN, LR, and Naïve Bayes (NB) for predicting diabetes risk.[43]

In the realm of chronic diseases, RF has been employed to identify metabolite patterns. Dias-Audibert et al.[44] used RF to identify biomarkers indicative of weight gain and diabetes. Peddinti et al.[45] employed feature selection to reveal metabolites associated with T2DM. Sisodia et al. reported NB having maximum accuracy in predicting diabetes compared to DT and SVM algorithms.[46]

Several studies explored predictors of diabetes. For instance, Oh et al.[47] investigated the development of T2DM from hyperlipidemia, hypertension, and impaired fasting glucose. Marcos et al.[48] used RF and GBM to predict obesity in children based on 190 predictor variables of body mass index (BMI). Zhang et al.[49] identified urine glucose concentrations as the top predictor of T2DM among various features. Dinh et al.[50] documented age as a predictor, while Kopita et al.[51] identified hyperglycemia as the primary predictor when compared with age, HDL-c, TG, physical activity, and medical history.

The reviewed studies demonstrate the potential of ML approaches, such as LR, RF, and NNs, to identify individuals at risk for diabetes and predict future complications. ML innovations for diabetes include risk stratification, patient state assessment, and complication prediction using time series data. The key predictors emerging from these computational analyses align with known pathophysiological factors that contribute to diabetes progression. For example, elevated blood glucose, increased BMI, hypertension, and abnormal lipid levels were repeatedly found to be predictive of diabetes development and outcomes. These factors cause systemic inflammation, insulin resistance, and metabolic dysfunction—hallmarks of diabetes pathophysiology. ML models were also able to discern complex patterns among hundreds of variables to pinpoint novel biomarker signatures of prediabetes and diabetes-related complications. Overall, these techniques can complement the understanding of the molecular underpinnings of disease and enhance early risk profiling, prediction of long-term outcomes, and personalized therapy. When thoughtfully designed and validated, ML models hold promise to reveal complex diabetogenic pathways for more timely and targeted prevention and treatment.

6 | PATHOPHYSIOLOGY OF CKD

CKD progressively damages the structure and function of the kidneys.[52] It is a major contributor to global morbidity and mortality, with an estimated 1.2 million deaths and 35.8 million disability-adjusted life years (DALYs) attributed to CKD in 2017.[52] Survivors, especially those with end-stage renal disease (ESRD), face immense costs for medications and renal replacement therapies (RRT). Without financial capacity for ongoing dialysis or access to RRT, many succumb to premature death. Early detection and diagnosis of CKD risk factors like diabetes, hypertension, high sodium intake, and obesity can prompt timely interventions to prevent progression to ESRD (Figure 3). However, current diagnostic tests like measuring albuminuria, proteinuria, or urinary albumin-to-creatinine ratio have questionable sensitivity for detecting CKD.[53] Estimating glomerular filtration rate (eGFR) is increasingly used for staging CKD. Serum creatinine levels are measured, and eGFR is calculated using equations like Cockcroft-Gault or Modification of Diet in Renal Disease. CKD stages are as follows: eGFR ≥60 mL/min/1.73 m^2 (Stages 1 and 2); eGFR 30–59 mL/min/1.73 m^2 (Stage 3); eGFR 15–29 mL/min/1.73 m^2 (Stage 4); and eGFR <15 mL/min/1.73 m^2 (Stage 5).[52]

7 | THE APPLICATION OF ML ALGORITHMS FOR DETECTION AND PREDICTION IN CKD

Dovgan et al.[54] explored the application of ML algorithms, such as RF, SVM, LR, k-NN, XGBoost, and NN, to predict the optimal timing for a patient with CKD to initiate RRT, encompassing renal transplantation or dialysis. Another study employed an NN to forecast which individuals with acute kidney injury would necessitate dialysis within a 48-h window.[55] Assessing CKD severity, ML algorithms, including lasso regression,

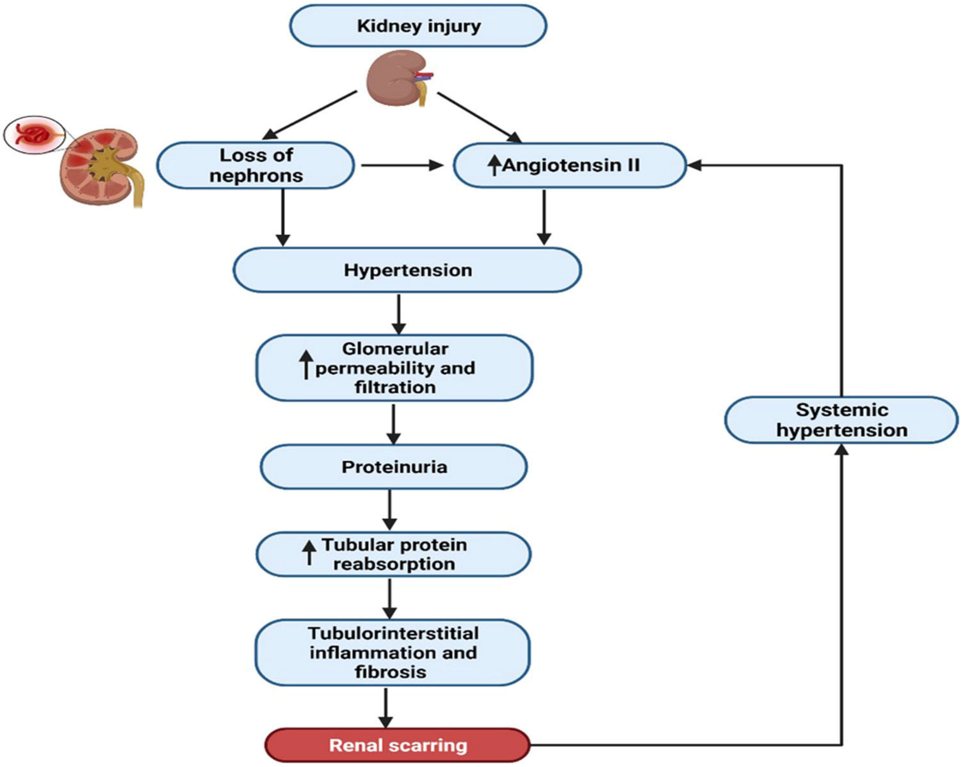
Kidney injury
Loss of nephrons
↑Angiotensin II
Hypertension
↑Glomerular permeability and filtration
Proteinuria
↑Tubular protein reabsorption
Tubulorinterstitial inflammation and fibrosis
Renal scarring
Systemic hypertension

FIGURE 3 Pathophysiology of chronic kidney disease.[52] Kidney injury results in loss of nephrons and increased levels of angiotensin II. Angiotensin II is a vasoconstrictor that triggers increased blood pressure. Alongside hypertension, there is an increase in glomerular permeability, tubular protein reabsorption, tubular/interstitial inflammation, and eventually renal scarring.

NN, ridge regression, k-NN, SVM, and XGBoost, achieved an average area under the curve (AUC) of 0.87 and a sensitivity of 0.8.[56] Proteomics data was subjected to ML algorithms to differentiate immune-mediated CKD from CKD caused by other factors.[57] Almansour et al.[58] utilized SVM, k-NN, and Soft Independent Modeling of Class Analogy on the UCI repository data set for chronic renal failure (CRF), achieving a 93% accuracy in distinguishing CKD patients from healthy individuals.

To predict the onset of CRF, NB, ANN, DT, and random subspace classification algorithms were applied, with DT identified as the best classifier among them.[59] Lee et al. employed an unsupervised bag-of-words model on histopathology images from kidney biopsies, revealing morphological features predicting CKD existence and outcomes with a 0.93 AUC for GFR and loss of function after 1 year.[60] Bueno et al.[61] demonstrated a sequential CNNs segmentation-classification strategy with 98% accuracy in detecting and classifying normal and sclerotic glomeruli. Predicting postoperative acute kidney injury risk in renal cell carcinoma patients, ML models, including SVM, RF, extreme gradient boosting, and light GBM (gightGBM), outperformed LR.[62]

Kim et al.[63] delved into predicting dialysis adequacy in chronic hemodialysis patients using ML algorithms, specifically RF, and XGBoost, alongside deep learning models such as CNN and gated recurrent unit. The XGBoost model emerged as the most accurate method, demonstrating superior performance compared to other algorithms. In summary, these studies showcase the versatility of ML in predicting various aspects of CKD, from onset and severity to treatment adequacy.

The potential of ML approaches like RF, NNs, and SVMs to predict different aspects of CKD progression and management has been established. The reviewed applications leverage longitudinal EHR data for prognosis forecasting and precision nephrology, while medical imaging and sensing data improve outcomes posttransplantation. For example, histology images and proteomics data were used to discern patterns predictive of immune-mediated kidney damage versus other CKD causes. ML models were also able to integrate diverse variables such as demographics, lab tests, and medications to forecast the onset of ESRD requiring dialysis. Optimal timing of interventions like transplantation was another application. Key predictors align with known drivers of CKD, including hypertension, diabetes, inflammation, and glomerular scarring. By uncovering complex interactions among hundreds of factors, ML can complement the understanding of renal decline mechanisms and support personalized prediction of disease trajectory. These techniques show promise in explaining CKD pathways for more timely and targeted prevention and treatment. However, model interpretability and causality warrant careful assessment when applying algorithms clinically.

8 | PATHOPHYSIOLOGY OF CVDs

CVDs, including coronary artery disease, stroke, rheumatic heart disease, and congenital heart defects (Figure 4), are major causes of mortality worldwide.[64] Key diagnostic tests for CVD risk evaluation include blood lipid measurements, electrocardiography (ECG), coronary artery calcium (CAC) scoring, and coronary CT angiography (CCTA). When strapped to a patient, ECG provides information about atrial depolarization (P-wave), ventricular depolarization (QRS complex), and ventricular repolarisation (T-wave). As a result, ECG reveals the electrical stability of the heart and can be used to detect whether the heart is experiencing normal rhythm, atrial fibrillation, heart block, premature ventricular contraction, and premature atrial contraction, among other arrhythmias. ML algorithms like ANNs and SVMs can optimize ECG interpretation. CAC scoring quantifies calcification in coronary arteries via computed tomography (CT) and independently predicts CVD risk.[65] CCTA visualizes coronary plaque morphology and vessel structure through noninvasive CT imaging. However, CCTA interpretation is subjective with inter-reader variability in diagnoses.[66] Advanced computational analysis of CCTA scans could enable more accurate and consistent CVD risk stratification.

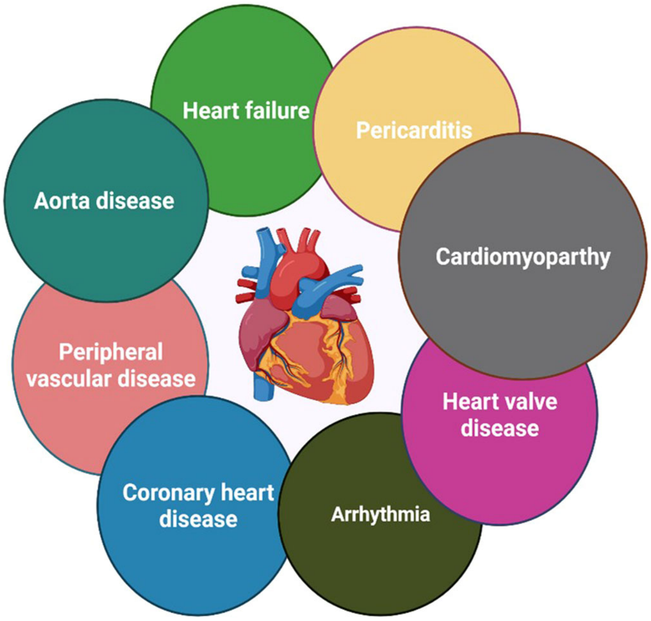

FIGURE 4 Different types of cardiovascular diseases.

9 | THE APPLICATION OF ML ALGORITHMS FOR THE DETECTION AND PREDICTION OF CVDs

ML algorithms have proven effective in enhancing the reliability and accuracy of scoring systems in medical imaging, particularly in CCTA automation. Applying SVM and nearest neighbor algorithms in CCTA automation demonstrated improved interpretability and quantification of image phenotypes.[67] Additionally, Bruse et al.[68] successfully employed characteristic cardiac magnetic resonance imaging (MRI) and hierarchical clustering analysis to subdivide aortic anatomical models into healthy cohorts and individuals with congenital heart disease. Similarly, cluster analysis applied to CT data allowed the subdivision of individuals with a bicuspid aortic valve into three distinct phenotypes.[69]

In predicting atherosclerotic CVD, ML methods such as RF, GBM, XGBoost, and LR were shown to be comparable or even superior to pooled cohort equations.[70] Yang et al.[71] reported the superiority of RF over other ML methods like classification and regression trees (CART), NB, Bagged Trees, and Ada Boost in predicting CVDs. Quesada et al.[72] found that 10 out of 15 ML algorithms adequately predicted CVD in a large cohort. AutoPrognosis, an automated framework, was demonstrated to improve the risk prediction of CVD compared to traditional methods such as Framingham Score and the Cox proportional hazard model.[73] Panaretos et al.[74] highlighted the superior predictive potential of ML methods, specifically RF and k-NN, in assessing the relationship between dietary patterns and cardiovascular risk compared to linear regression.

In the context of congenital heart defects, Yu et al.[75] applied PCA and RF to identify potential biomarkers using DNA methylation data. Three genes (MIR663, FGF3, and FAM64A) were identified with RF, achieving an average sensitivity and specificity of 85% and 98%, respectively. Principal components explained over 70% of the variance and effectively classified samples with congenital heart defects and control groups (Figure 5).

The reviewed studies demonstrate the potential of ML approaches like RFs, SVMs, and clustering algorithms to enhance diagnosis and risk prediction for CVDs. ML techniques for cardiovascular risk prediction, electrocardiogram (ECG) analysis, medical imaging analysis, electronic phenotyping from EHRs, and multi-omics integrative modeling have been established. By automating and improving the interpretation of complex diagnostic tests like CT angiography, ECG, and MRI, these techniques can enable earlier detection of conditions like atherosclerosis, heart arrhythmias, and congenital defects. For instance, ML models were able to identify patterns in CT angiography scans that distinguish normal anatomy from congenital defects and progressive atherosclerosis. In addition, ML models

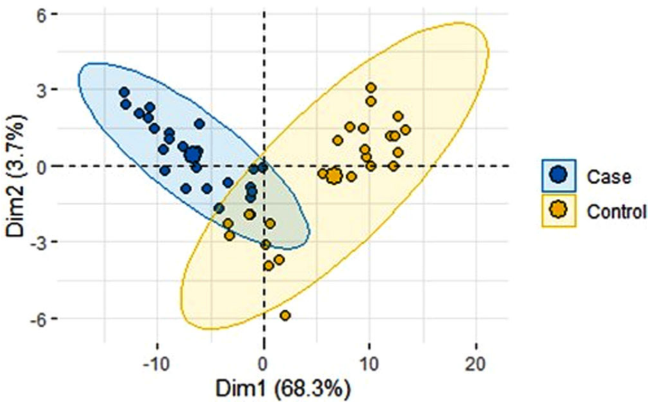

FIGURE 5 Principal component analysis (PCA) (with 95% confidence ellipses) showcases the clustering patterns and reasonable discriminatory power for samples with congenital heart defects (case) and control using the top 20 expressive genes identified via RF based on orthogonal linear combinations of the features.

were able to integrate diverse factors from genetics, diet, and clinical data to predict individual risk of developing CVD. Key features identified by the models provide insight into the underlying biology. For example, DNA methylation patterns were linked to congenital heart defects. Overall, these computational methods show promise in elucidating the intricate molecular pathways driving cardiovascular pathologies. When thoughtfully designed and validated, they could guide personalized prevention and treatment strategies. However, close collaboration between data scientists and clinicians is crucial to ensure clinical validity and utility.

10 | PATHOPHYSIOLOGY OF CANCER

Cancer, characterized by abnormal gene functioning, arises from spontaneous DNA mutations, either inherited or induced by environmental agents such as radiation, chemicals, and viruses. These alterations include point mutations, translocations, deletions, and gene amplifications, affecting crucial genes like proto-oncogenes (e.g., K-ras, epidermal growth factors [EGFR]) and tumor suppressor genes (e.g., TP53, retinoblastoma).[76–78] The resulting modified cells evade immune responses and growth inhibitory signals, leading to uncontrolled proliferation and immortality. Tumor cells must navigate through various stages, interacting with the extracellular matrix, penetrating vascular barriers, entering circulation, and ultimately colonizing distant organs.[79,80]

The mechanism underlying cancer cell proliferation involves the secretion of growth factors and their corresponding receptors, such as nerve growth factor, epidermal growth factor (EGF), transforming growth factor (TGF-α and TGF-β), and platelet-derived growth factor (PDGF). These factors, in conjunction with cytokines and hormones, trigger intracellular biochemical signaling that activates or represses genes. Lack of growth factor stimulation can result in programmed cell death and apoptosis.[81–84] It is increasingly evident that cancer results from both genetic alterations and epigenetic changes, with cancer cells exhibiting distinct epigenomic profiles compared to healthy cells. Epigenetics, defined as heritable changes in gene expression without alterations to the DNA sequence, involves processes including DNA methylation, nucleosome remodeling, and histone modification. In cancer, CpG islands in promoter regions become hypermethylated, silencing crucial genes like tumor suppressors.[85,86]

Aberrant posttranslational modifications of histones and altered functions of histone-modifying enzymes contribute to certain cancers. For instance, missense mutations in p300 histone acetyltransferase have been identified in colorectal and breast cancers. Changes in histone modification patterns have been linked to predicting prostate cancer recurrence.[87–89]

The pathophysiology of cancer involves a complex interplay of genetic alterations and epigenetic changes that disrupt normal cellular processes (Figure 6). The transition from controlled cellular growth to uncontrolled proliferation and invasion is orchestrated by a combination of mutations in critical genes and modifications in epigenetic mechanisms. Understanding these underlying molecular events is crucial for developing targeted therapeutic interventions and improving cancer management strategies.

11 | THE APPLICATION OF ML ALGORITHMS FOR DETECTION AND PREDICTION OF CANCER

The detection of various cancers poses a challenge for clinicians, who traditionally rely on biopsies for diagnosis. However, this approach may underutilize significant information. MRI has emerged as a powerful tool for visualizing tumors, aiding in biopsy selection, assessing tumor severity and stage, and facilitating the development of targeted therapies. While MRI is valuable for tumor localization and patient stratification, interpretation can be subjective, leading to interreader variability among radiologists.[90–93]

Laboratory technologies such as polymerase chain reaction (PCR) and next generation sequencing (NGS) enable quantitative detection of gene expressions in various tissues. Tests like in situ hybridization and DNA microarrays identify cancer biomarkers, allowing the classification of multiple cancers.[94] The data generated by these technologies, rich and complex, benefits from advanced computational approaches for interpretation. ML plays a vital role in comprehensively analyzing data, identifying patterns, and improving patient diagnosis.

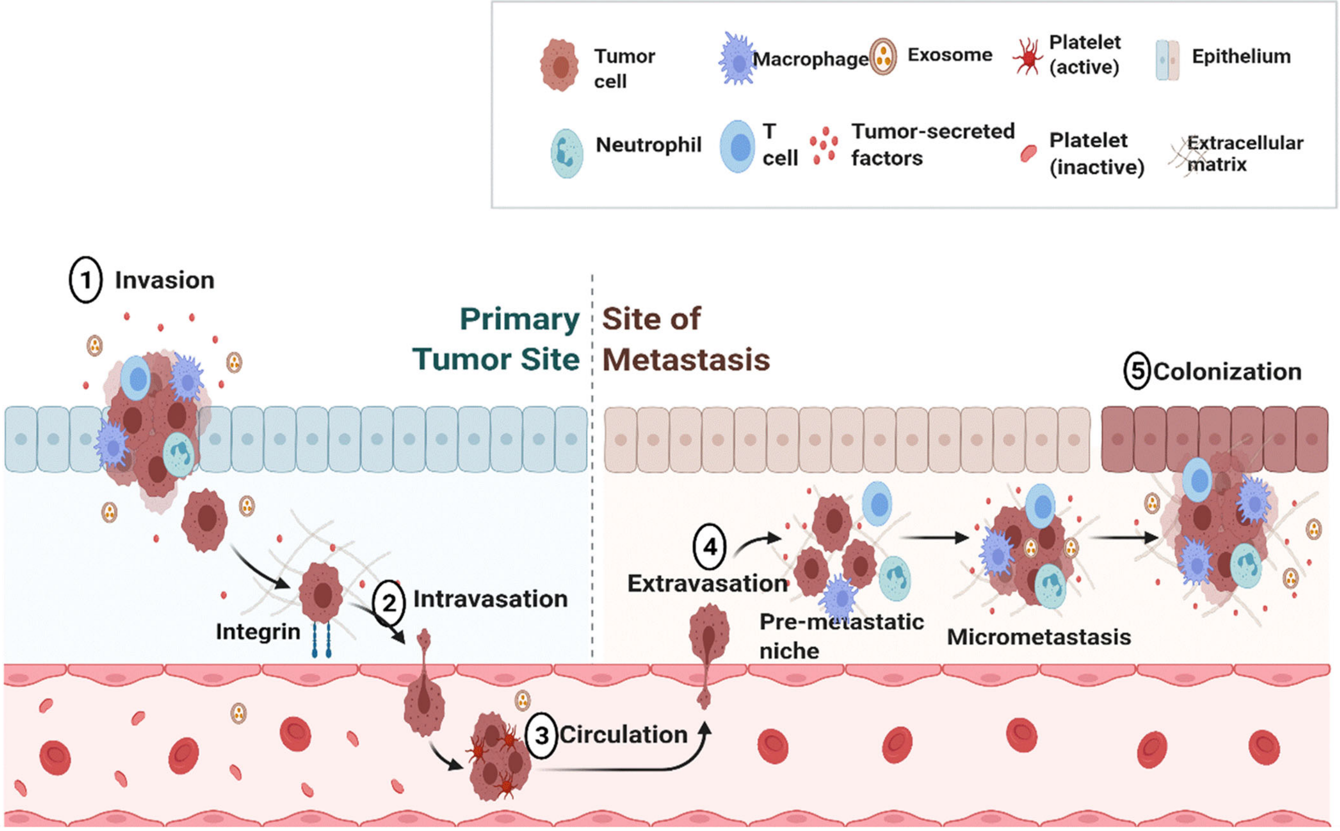

FIGURE 6 The metastatic cascade. The spread of a tumor is characterized by a sequence of events. These events are local invasion, intravasation, circulation through the vasculature, extravasation, formation of micrometastasis, and colonization.

ML methods have demonstrated effectiveness in validating presumptive diagnoses, comparable to experienced human radiologists.[95]

ML extends its application to genomics, enhancing the understanding of cellular processes affected by genomic alterations. ML algorithms analyze gene sequences, expression profiles, histone modifications, and RNA-seq data to differentiate disease phenotypes and make predictions. CNN integrated into DNA/RNA sequence data predicts binding scores and DNA-RNA-protein interactions. Algorithms are developed based on patterns in genetic data to build models, unraveling the impact of genomic alterations on cellular processes such as metabolism and cell growth.[96–98]

Moreover, ML algorithms excel in detecting genomic variants associated with diseases, whether in coding or noncoding regions. They rank genomic variants based on pathogenicity, providing valuable insights into disease mechanisms. ML can also be employed to identify missing heritability and genetic variants in rare diseases. For instance, Yin et al.[99] demonstrated that CNN revealed a link between promoter regions and amyotrophic lateral sclerosis.

In summary, we have highlighted key applications of ML methods for cancer detection, diagnosis, prognosis, and treatment prediction including medical imaging analysis, molecular profiling integration, and patient trajectory modeling from EHR data. The utility of ML approaches, including CNN, in advancing cancer detection, diagnosis, and treatment has been emphasized. By automating the analysis of complex diagnostic imaging and genomic data, these computational techniques can uncover novel biomarkers and molecular patterns associated with cancer subtypes and disease trajectories. Key applications include improving radiologist interpretation of MRI scans for tumor characterization, as well as discerning complex gene expression signatures that provide insights into dysregulated cellular pathways driving malignancy.[95–98] For example, models were able to link promoter region mutations to neurodegenerative disease.[99] Such computational findings can complement understanding of the intricate molecular events underlying tumor initiation, progression, and metastasis.

12 | PATHOPHYSIOLOGY OF CHRONIC RESPIRATORY DISEASES

Chronic respiratory diseases, including asthma, occupational lung diseases, chronic obstructive pulmonary disease (COPD; emphysema, chronic bronchitis, chronic asthma), and pulmonary hypertension, are a group of progressive diseases that block the airway and affect multiple sites of the lung, such as the alveolar and perialveolar tissues. This group of diseases may permanently deprive a person of normal breathing and, when unmanaged, can lead to death. In 2019 alone, asthma killed 46,100 people and affected an estimated 262 million people.[100] These diseases are characterized physiologically by obstruction of airflow into and out of the lungs and airway remodeling due to chronic inflammation or infections. For example, emphysema is a permanent enlargement of alveolar spaces, and some pathological factors that have been implicated in its development include protease‐antiprotease activity or increased activity of matrix metalloproteinase (MMPs) that degrade lung matrix (MMP‐8 and ‐9), histone deacetylase inhibition, and oxidant injury.[101–103] Like many other lung diseases, a common environmental factor that is linked to emphysema is tobacco smoking. Cigarette smoke has been documented to unleash elastases, MMPs, proteinases, and cathepsins from macrophages. Moreover, cigarette smoke can trigger the release of TGF‐β, EGF, and PDGF, as well as cytokines such as IL‐8 and interferon‐gamma (IFN‐γ). Combined, these factors lead to airway remodeling and damage of elastic fiber that subsequently results in airflow limitation.[103] Spirometry, lung volume determinations, and diffusing capacity assessments are the common pulmonary function tests. The goals of these tests are to measure the extent of the lung anomaly and to describe the physiological abnormality.[104]

For many years, distinguishing restrictive pulmonary disease from obstructive has been done by quantifying the ratio of the 1 s forced expiratory volume (FEV1) to forced vital capacity (FVC) and calculating its percentage. In a normal individual, 75%–80% of the total volume of exchangeable air in the lungs ([vital capacity, i.e., VC = tidal volume [TV] + inspiratory reserve volume [IRV] + expiratory reserve volume [ERV]) is exhaled in 1 s. An FEV1/FVC < 70% indicates an obstructive lung disease, whereas restrictive lung disorders generally have an FEV1 to FVC ratio unchanged or greater than 70%.[105] Making a correct diagnosis is a critical step to optimizing therapy for COPDs. For example, Aaron et al. reported that 20%–73% of cases of asthma are undiagnosed, whereas overdiagnosis has been documented in 30% to 61% of cases.[106] Also, underdiagnosis and overdiagnosis have been documented in 70% of COPD cases. Whereas underdiagnosis of respiratory diseases may result in reduced quality of life and frequent hospitalization. On the contrary, overdiagnosis may potentially expose patients to detrimental side effects from medications that are unlikely to give a clinical benefit.[107]

13 | THE APPLICATION OF ML ALGORITHMS FOR DETECTION AND PREDICTION IN CHRONIC RESPIRATORY DISEASES

The application of high‐resolution CT and bronchoalveolar lavage[108] has become increasingly prevalent in detecting interstitial lung diseases like idiopathic pulmonary fibrosis and sarcoidosis.[109] Surgical lung biopsy, such as transbronchial lung cryobiopsy, may be necessary for a definitive diagnosis in some cases.[110] Advancements in medical technology have ushered in a new era where ML plays a pivotal role in diagnosing lung diseases. ML algorithms analyze patterns in diagnostic tests, automatically interpret information, and predict outcomes.[111] For instance, Finamore et al. utilized K‐means cluster analysis to assess the association between end‐stage COPD, chronic heart failure, and CRF with mobility, care dependency, health status, and life‐sustaining treatment preferences, achieving a classification accuracy of 94.22%.[112] Mostafaei et al. explored the link between smoking gene expression and COPD using various ML algorithms such as AdaBoost Classification Trees, DT, GBM, NB, NN, RF, SVM, adaptive LASSO, Elastic‐Net, and Ridge LR. Their analysis identified 44 candidate genes, including PRKAR2B, GAD1, LINC00930, and SLITRK6, implicated in COPD pathogenesis.[113]

ML algorithms have also been instrumental in early exacerbation detection and triage in COPD. SVM, LR, NB, k‐NN, gradient boosted, and ensemble DT methods outperformed individual pulmonologists in determining the likelihood of exacerbation and consensus triage, offering transparent and consistent decision‐making.[114] Similarly, NN could detect cancerous lung nodules as accurately as experienced radiologists.[115] In the domain of emphysema and interstitial lung disease, ML algorithms have enhanced visual scoring with ad hoc designed image‐based features, allowing the characterization of 10 novel emphysema radiological subtypes.[116] Moreover, these algorithms have facilitated the cost‐effective classification of fibrotic lung disease in a highly reproducible manner.[117] Combining ANN and spirometric measurements, Ioachimescu et al. distinguished between normal, obstruction, restriction, and mixed impairments more accurately than traditional methods.[118] CNN with CT scans as output surpassed traditional spirometry parameters and RF classifiers in discriminating predominant emphysema/airway phenotypes in COPD.[119]

In the context of COVID‐19, Ali et al. classified pneumonia severity (mild, progressive, and severe) in patients using SVM, DT, k‐NN, and CNN, achieving high accuracy rates ranging from 87.5% to 95.622%.[120] ML algorithms have also been applied to predict and explain inflammation in Crohn's disease, with GBMs demonstrating accurate predictions of inflammation severity.[121] Regularized regression and LR performed comparatively well, showcasing their utility in disease prediction.[121]

We have highlighted the potential of ML approaches, such as NNs, SVMs, and cluster analysis, to advance diagnosis and prognosis in chronic respiratory illnesses like COPD, pulmonary fibrosis, and COVID-19 pneumonia.[108–121] These techniques could discern distinct disease subtypes and molecular patterns associated with pathogenesis via analysis of complex imaging data and gene expression profiles. For example, models identified candidate genes like PRKAR2B and SLITRK6 linked to smoking-related lung damage.[113] Radiological emphysema subtypes were also revealed. Such computational insights can reveal the intricate inflammatory, fibrotic, and dysfunctional respiratory pathways underlying these conditions. Accurate classification of disease severity and exacerbation risk using ML could also guide triage and timely interventions. These analytical approaches hold promise to unravel the complex molecular underpinnings of respiratory diseases for improved early detection, personalized treatments, and prognostic information to enhance patient outcomes. However, close collaboration between data scientists and pulmonologists is key to clinically validate and implement these emerging technologies responsibly.

14 | PATHOPHYSIOLOGY OF INFLAMMATORY DISEASES

Inflammation serves as an innate defense mechanism involving various components such as immune cells (e.g., leukocytes, macrophages), cell-derived mediators (histamines, prostaglandins, leukotrienes, cytokines, nitric oxides), and plasma-derived mediators (complement proteins, kinins). This collective response aims to combat infections, eliminate dead cells or tissues, and initiate repair and recovery processes.[122,123] The inflammatory process can manifest as either acute or chronic.

Acute inflammation is short-lived, with an elevated inflammatory response that subsides once the cause of cell injury is removed, inactivated, or degraded.[122,123] Characterized by an efflux of neutrophils to the site of injury, acute inflammation involves vascular changes leading to increased vascular permeability, resulting in the leakage of plasma proteins and the formation of edema. Prostaglandins, leukotrienes, and thromboxane A2, synthesized from arachidonic acid, contribute to increased vascular permeability.[80]

Chronic inflammation, influenced by psychological, biological, and environmental factors, is characterized by prolonged leukocyte and macrophage accumulation, concurrent tissue injury, or repair. Recovery in chronic inflammation can occur through neutralization of the irritant, destruction by leukocytes, or antibiotic use. Alternatively, damage may be repaired through the formation of granulation tissue (fibroblasts, blood vessels, and macrophages), leading to fibrosis or scarring.

Both acute and chronic inflammation share a common mechanism, though the initial stimuli may differ. Acute inflammation is triggered by pathogenic molecules, known as pathogen-associated molecular patterns (PAMPs), binding to pattern recognition receptors (PRRs) on immune cells. PRRs include toll-like receptors, NOD-like receptors, C-type lectin receptors, RIG-1-like receptors, and nucleotide-binding oligomerization domain-like receptors.[124] Conversely, chronic inflammation is primarily caused by damage-associated molecular patterns (DAMPs) and is not dependent on PAMPs or the initial stimulus.

Cytokines play a crucial role in regulating the immune response during inflammation, classified as pro-inflammatory (e.g., IL-1, IL-6, and TNF-α) and anti-inflammatory (TGF-β, IL-10, and IL-4). The homeostatic regulation of cytokine expression in normal cells is essential.[125] Uncontrolled inflammation, such as a strong response against self-antigens or harmless environmental antigens, can lead to tissue damage, impaired immune cell function, hindered repair mechanisms, and increased susceptibility to diseases. Abnormal inflammatory responses are associated with pathological events such as diabetes, inflammatory bowel disease (IBD), psoriasis, Crohn's disease, and arthritis.[124] (See Figure 7 for a visual representation of cytokine regulation in a normal cell).

15 | THE APPLICATION OF ML ALGORITHMS FOR DETECTION AND PREDICTION IN INFLAMMATORY DISEASES

In clinical practice, markers of inflammation, including C-reactive protein (CRP), serum amyloid A, and fibrinogen, play a crucial role.[124,126] Inflammatory disorders exhibit diverse presentations and localizations and can affect multiple organs. This heterogeneity poses significant challenges for diagnosis and management. While histological and endoscopic evaluations can be useful, they may not be sufficient for accurate diagnosis and can be labor-intensive, requiring expertise. Determining the extent and severity of tissue damage and its localization is essential for deciding the most appropriate treatment approach. ML algorithms offer a promising avenue for classifying various inflammatory conditions and predicting responses to therapy.

Mossotto et al.[127] utilized unsupervised ML algorithms (PCA, multidimensional scaling) and supervised ML algorithms such as ensemble learners, linear discriminant analysis, and SVM to classify children with pediatric inflammatory bowel disease (PIBD). Applying these algorithms to endoscopic and histologic data enabled the classification of 83.3% of individuals.[127] Similarly, in another study, SVM, XGBoost, dense neural network, RF,

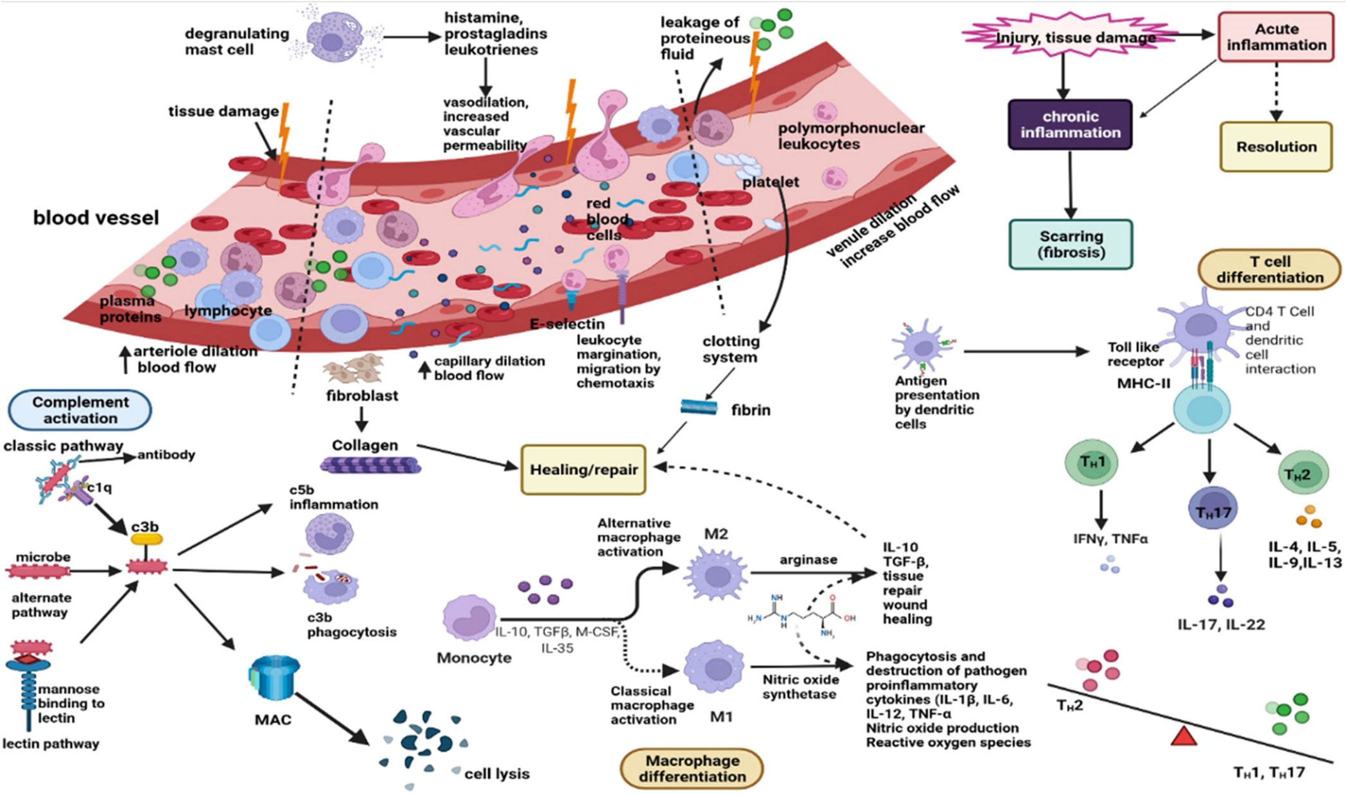
degranulating mast cell
histamine, prostagladins leukotrienes
vasodilation, increased vascular permeability
leakage of proteineous fluid
tissue damage
blood vessel
polymorphonuclear leukocytes
platelet
red blood cells
venule dilation increase blood flow
plasma proteins
lymphocyte
E-selectin
arteriole dilation blood flow
capillary dilation blood flow
leukocyte margination, migration by chemotaxis
clotting system
fibroblast
fibrin
Complement activation
classic pathway
antibody
c1q
c3b
microbe
alternate pathway
mannose binding to lectin
lectin pathway
Collagen
Healing/repair
c5b inflammation
c3b phagocytosis
MAC
cell lysis
Monocyte
IL-10, TGFβ, M-CSF, IL-35
Alternative macrophage activation
M2
arginase
Classical macrophage activation
M1
Nitric oxide synthetase
Macrophage differentiation
IL-10 TGF-β, tissue repair wound healing
Phagocytosis and destruction of pathogen proinflammatory cytokines (IL-1β, IL-6, IL-12, TNF-α Nitric oxide production Reactive oxygen species
Injury, tissue damage
Acute inflammation
chronic inflammation
Resolution
Scarring (fibrosis)
T cell differentiation
Antigen presentation by dendritic cells
Toll like receptor
MHC-II
CD4 T Cell and dendritic cell interaction
T_H1
T_H17
T_H2
IFNγ, TNFα
IL-17, IL-22
IL-4, IL-5, IL-9,IL-13
T_H2
T_H1, T_H17

FIGURE 7 Response to inflammation. The components of inflammation include vascular changes—vasodilation that results in increased blood flow and increased vascular permeability that allows fluids to reach the infected site. Immune cells, including macrophages and leukocytes, are also activated to phagocytose infectious agents. Monocytes are also recruited, which become macrophages. These macrophages can differentiate into M1 and M2 when stimulated by cytokines. M1 releases the anti-bactericidal compound NO, whereas M2 macrophages promote wound healing and tissue repair. Protein systems, including the complement, coagulation, and kinin, are all activated inflammatory responses. Depending on the stimulus, T-cells may be differentiated into TH1, TH2, and TH17, and all these are involved in various stages of inflammation.

and CNN were employed to classify PIBD using historical data. CNN emerged as the best performer, achieving an accuracy of 90.57% with a standard deviation of 3.45.[128]

Among patients with IBD receiving intravenous glucocorticosteroids, RF identified higher levels of CRP and longer disease duration as predictors of their hyperglycaemic status.[129] Additionally, RF accurately predicted IBD (AUC > 0.9) based on 117 differential bacterial taxa.[130] Kraszewski et al. developed ML algorithms to predict Crohn's disease and ulcerative colitis using historical data from fecal, urine, and blood tests. RF emerged as the best classifier for Crohn's disease and ulcerative colitis with a precision of 97% and 91%, respectively.[131]

To distinguish between IBD and alimentary lymphoma in cats, Awaysheh et al.[132] created three ML algorithms, including NB, DT, and ANN. Models trained on data from complete blood count and serum chemistry tests showed that NB and ANN were better classifiers (sensitivities of 70.8% and 69.2%, respectively) than DT (63%, $p < 0.0001$). ML algorithms were also employed for risk prediction and early diagnostic disease of IBD associated with arthropathy, achieving promising results with an ROC of 0.90 (95% confidence interval [CI] 0.80–0.99; accuracy 96%).[132]

ML techniques, including RF, NNs, and SVM, offer the capability to analyze complex patterns in biomarkers, histology, microbiome, and other clinical data. By applying these techniques, distinct molecular subtypes of inflammation affecting specific organs could be identified. For instance, models have successfully differentiated Crohn's disease from ulcerative colitis based on data from fecal and blood tests.[131] Some of the key variables identified by algorithms, such as CRP levels and bacterial composition, provide valuable insights into the drivers of chronic inflammation.[129] ML approaches can also integrate diverse factors to predict steroid response and anticipate complications, such as arthropathy.

16 | CHALLENGES OF ML APPLICATIONS IN CHRONIC DISEASE DETECTION AND PREDICTION

Currently, there are a limited number of internationally approved products specifically designed for chronic disease detection and prediction using ML techniques. However, there are several promising initiatives and ongoing efforts to bring such products to market (see https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices). One notable example is the FDA-approved IDx-DR, an AI-based digital diagnostic system for detecting diabetic retinopathy, a common complication of diabetes that can lead to blindness. IDx-DR, developed by Digital Diagnostics, uses ML algorithms to analyze images of the retina and identify signs of diabetic retinopathy, enabling early detection and intervention. Another example is the CE-marked KardioScreen™, an ML-based tool developed by Kardiolytics for predicting the risk of CVD. KardioScreen analyses various patient data points, including demographics, medical history, and lifestyle factors, to provide personalized risk assessments and recommendations for preventive care.

While these examples demonstrate the potential of ML in chronic disease detection and prediction, it is important to note that the regulatory landscape for AI-based medical products is still evolving. ML holds immense potential to transform chronic disease detection and management, yet thoughtful implementation is key to realizing benefits while navigating pitfalls, including model transparency, causality, bias, validation, data quality and EHR limitations and privacy trade-offs, and impacts on the doctor–patient relationship.

16.1 | Model transparency and interpretability

Model transparency and interpretability have emerged as major challenges in the application of ML, particularly deep learning, to healthcare.[133] Complex algorithms like deep NNs comprise multiple hidden layers and intricate connections between computational nodes. This complexity enables powerful pattern recognition from data but also leads these models to behave as inscrutable "black boxes", obscuring the basis for their predictions and diagnoses. Without transparency into the reasoning behind AI system outputs, barriers are posed to clinical acceptance and real-world deployment of such tools. Moreover, the opaqueness prevents accountability in auditing model decisions that impact patient care.

Some strategies have been proposed to open the black box and improve model interpretability. For example, Ribeiro et al.[134] proposed the novel Local Interpretable Model-Agnostic Explanations technique that can help provide local explanations about the contributions of individual inputs to model outputs.[134] However, fundamental trade-offs remain between accuracy and explainability. Simpler and more interpretable models like DTs often achieve lower performance than black-box models, and explanations of model functioning may not fully capture the intricate interactions occurring within multilayer NNs. Interpretability techniques also make assumptions or approximations about the model to generate explanations rather than revealing the true reasoning process. Significant research is still needed to make deep learning models more intrinsically understandable without sacrificing predictive power. For now, pragmatic approaches that provide post hoc explanations of model behaviors may present a middle ground. However, ultimately, reasonable interpretation of the causal mechanisms linking input features to outputs is key to engendering appropriate trust in ML algorithms and enabling responsible clinical adoption.

16.2 | Inference on causality

Determining causality versus correlation is another fundamental challenge in applying ML to healthcare.[135] ML models are efficient at finding predictive patterns and correlations in data. However, just because a model identifies a correlation between variables does not necessarily mean that one causes the other. For instance, an algorithm may predict disease outcomes based on detected biomarkers, but those markers may simply be correlates rather than true causal factors driving pathogenesis. While predictive correlations can still be clinically useful, revealing underlying causative mechanisms is critical for developing effective interventions.

Most ML models are statistical rather than causal by nature—they uncover mathematical relationships but cannot confirm causal mechanisms like biological experiments can. Combining ML with biological knowledge and experimental validation in a feedback loop is key to moving from correlation to causation. For example, putative causal genes identified by algorithms can be experimentally perturbed to validate effects on disease processes. Promising new techniques are also emerging from the field of causal inference that integrate causal assumptions into ML models.[136] However, causal inference from observational data remains challenging. Access to large datasets from interventional studies, where variables are manipulated rather than just observed, can enhance the ability to discern causation. Overall, ML offers a valuable hypothesis-generating tool, but collaborations with domain experts and experimentalists are crucial to substantiate causal discoveries, explain disease mechanisms, and guide therapeutic development.

16.3 | Algorithm biases

Algorithmic biases pose a major challenge in applying ML to healthcare, as models can perpetuate disparities if trained on skewed or incomplete data.[137] Real-world datasets often reflect societal biases and lack diverse representation. If certain populations are underrepresented in training data, models may learn patterns that disproportionately benefit the majority groups. For example, an algorithm to predict disease risk could be less accurate for minorities if developed using data from predominantly white patients.

Several strategies exist to help mitigate algorithmic bias and promote fair representation.[138] Careful sampling and data augmentation techniques can improve model training with balanced, representative data. Algorithms can also be constrained to satisfy mathematical definitions of fairness, like producing equal false positive/negative rates across groups. However, completely eliminating bias is enormously challenging, given the complexities of real-world data. Bias can be introduced from many sources, including incomplete knowledge of confounding variables, and equitable outcomes may require trade-offs between fairness constraints and model performance.

A multidimensional approach is, therefore, essential. In addition to technical bias mitigation techniques, diversity among data scientists and interdisciplinary teams incorporating domain experts can improve consideration of biases throughout development. Open communication, standardized evaluation metrics, and careful scrutiny of model behavior across user groups are imperative. However, understanding the limitations of available data and inevitable gaps in knowledge is key. The continuous evolution of best practices for responsible design is crucial to fulfilling the full potential of ML in supporting chronic disease prediction and management. Above all, these technologies should augment human intelligence and not replace human accountability for equitable care.

16.4 | Model validation

Rigorous validation on heterogeneous datasets is critical to assess the real-world generalizability of ML models and avoid biases from overfitting to limited data.[139] Common practices like cross-validation on held-out test sets, while useful, may not adequately evaluate model performance across diverse settings. Algorithms tuned on data from a single institution, for instance, may not generalize well to other populations and care environments.

Robust validation requires testing models on data that differ meaningfully from the original training distribution. Strategies include evaluating models across multiple datasets from distinct institutions or locations and on patient subgroups with differing demographics, risk factors, and disease stages. Challenging models with synthesized data containing novel combinations of features can also assess generalizability. Furthermore, standardized reporting guidelines have been proposed to improve reliability and transparency in describing model development, evaluation, and real-world testing.[140]

However, major barriers to rigorous validation remain around data availability, quality, and interoperability across settings. Thoughtful regulatory guidance can help stimulate the generation of high-quality, heterogeneous datasets.[140] But ultimately, cooperation across institutions and the healthcare ecosystem is imperative to enable robust model evaluation and sustain trust in ML technologies. Ongoing monitoring of performance across populations and care settings should be embedded in clinical deployment. Although resource-intensive, a dedicated focus on generalizability can ensure ML lives up to its potential while avoiding the pitfalls of overfitting.

16.5 | Data quality

Data quality issues like incompleteness, noise, biases, and variability pose major obstacles to developing accurate and robust ML models in healthcare.[141] Real-world health data is often messy, with challenges like missingness, duplication, and errors. Missing data can stem from skipped measurements or documentation lapses. Strategies like imputation methods have been proposed to address incomplete datasets. But there are limits to techniques for handling missing data, and lowered data quality can hamper model performance and validity.

More fundamentally, success with ML hinges on access to comprehensive, high-quality, standardized, and well-curated data. This necessitates dedicated efforts to properly generate, record, structure, validate, integrate, and prepare data for computational use. Standardized terminologies and ontologies are key to unifying data from disparate sources. But thoughtfully designed EHR, transparent data management protocols, and governance policies are equally crucial to engender trust and enable effective sharing.

The FAIR principles provide useful guidelines, emphasizing the findability, accessibility, interoperability, and reusability of data.[142] Adherence to such principles can fuel ML by facilitating aggregation of diverse, high-fidelity data. But this requires cross-institutional and cross-stakeholder collaboration. While data curation demands resources, the investments can pay long-term dividends for developing and responsibly implementing AI in medicine.

16.6 | Quality, structure, and fragmentation of EHRs

EHRs offer a valuable source of real-world clinical data. However, major challenges exist around the quality, structure, and fragmentation of EHR data, which can limit utility for ML.[143] Records often contain irregular sampling, redundant entries, missing values, and coding errors. Lack of standardization across healthcare systems also hinders the integration of records. Finally, EHRs are designed for billing and clinical workflows—not computational analysis.

However, advances in transfer learning and generative modeling show promise in leveraging EHR data despite limitations.[144,145] Transfer learning adapts models trained on rich datasets to new tasks where data is sparse, enabling learning from imperfect EHR data. GANs can simulate realistic synthetic EHR records to augment training data, while federated learning also allows collaborative model development across systems without centralized data sharing. Yet gaps remain in effectively harnessing EHR data, and the key priorities include improving data quality through error correction techniques and enhanced documentation practices. Clinical and computational experts collaborating to optimize data collection, structure records suitably for analysis, and apply privacy-preserving analytic techniques will further enhance the promise of EHR mining.

16.7 | Privacy trade-offs and impacts on the doctor–patient trust

The increasing use of ML algorithms in medicine has the potential to alter the traditional doctor–patient relationship and responsibility structures, potentially leading to tensions between privacy and utility trade-offs.[146] To train accurate models, ML systems require large, high-quality datasets, which may involve collecting and analyzing sensitive patient information without explicit consent.[147] Moreover, the opacity of ML models challenges informed consent and makes it difficult for clinicians to explain ML-aided diagnoses, disrupting doctor-patient trust and shared decision-making. This creates an ethical dilemma regarding patient privacy versus the potential public health benefits.[146,147]

To ensure responsible implementation, data collection and ML model development should adhere to principles of data minimization, transparency, and patient consent.[148] While ML can provide diagnostic and treatment recommendations, the clinician remains responsible for the final judgment and care of the patient. To uphold ethical doctor–patient relationships as ML adoption accelerates, clinicians must communicate the role and limitations of ML in a transparent manner to patients, invite patient participation in ML-related decisions, advocate for explainable ML systems, audit ML-aided diagnoses for biases, and retain responsibility for all clinical judgments, advocating against fully autonomous ML diagnosticians.[149,150] To this end, explainable AI systems should be favored over black-box models when reasonable, and ML-aided decisions must be audited for biases that could exacerbate healthcare disparities.[149] Additionally, guidelines and policies are needed to clarify liabilities and preserve clinician accountability as the human linchpin in healthcare.

17 | CONCLUSION AND OUTLOOK

The exponential growth of complex biomedical data from sources like genomics, medical imaging, EHRs, and wearables has outpaced traditional statistical analysis methodologies. ML offers a promising avenue to help researchers, clinicians, and healthcare professionals exploit the rich insights of multimodal data. A key advantage of ML is the ability to rapidly process expansive, high-dimensional datasets to identify patterns and make predictions with reasonably high accuracy. While the full potential of ML in biomedicine remains untapped, we are already witnessing revolutions in data-driven fields ranging from genomics to medical imaging enabled by its application.

However, realizing the promise of ML in the chronic disease space requires thoughtful consideration of both its practical implementation and theoretical underpinnings. Practically, issues around data quality, algorithmic bias, model interpretation, validation, and integration into clinical workflows need to be proactively addressed. Furthermore, while ML excels at pattern recognition, incorporating causal domain knowledge is imperative for elucidating biological mechanisms and guiding therapeutic development rather than just making black-box predictions. Theoretically, intensive research is still required to develop tailored techniques that handle the distinct properties of biomedical data structures while ensuring generalizability across diverse cohorts. Promising directions include hybrid ML pipelines, explainable AI, and deep NNs capable of learning multimodal feature representations. Overall, the future is promising for ML in augmenting human intelligence for discovery and improved decision-making. Nurturing collaborations between data scientists and clinical researchers and pursuing responsible and ethical ML development will be key to successfully harnessing its potential for combating chronic diseases.

AUTHOR CONTRIBUTIONS

Ebenezer Afrifa-Yamoah and Eric Adua conceived and designed of study, methodology, performed the literature search and visualization, and wrote the first draft of manuscript. Emmanuel Peprah-Yamoah, Enoch O. Anto, Victor Opoku-Yamoah, Emmanuel Acheampong, Michael J. Macartney, and Rashid Hashmi screened the literature, critically reviewed, and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all the individuals who have contributed to this research. We extend our special thanks to the colleagues from the School of Science, Edith Cowan University, Perth, Australia, and the Rural Clinical School, Medicine and Health, University of New South Wales, Wagga Wagga, Australia, for their valuable insights and expertise that greatly assisted our work. Their contributions have been instrumental in shaping the direction and quality of this research. We would also like to extend our appreciation to the anonymous reviewers for their thorough and constructive feedback, which has significantly improved the quality of this manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Not applicable.

ETHICS STATEMENT

Not applicable.

ORCID

Ebenezer Afrifa-Yamoah http://orcid.org/0000-0003-1741-9249
Eric Adua http://orcid.org/0000-0002-6865-3812
Emmanuel Peprah-Yamoah http://orcid.org/0000-0001-5199-2829
Enoch O. Anto https://orcid.org/0000-0001-9023-6612
Victor Opoku-Yamoah http://orcid.org/0000-0002-7608-8040
Emmanuel Acheampong http://orcid.org/0000-0002-5338-3258
Michael J. Macartney http://orcid.org/0000-0001-7265-5725
Rashid Hashmi http://orcid.org/0000-0002-5792-1156

REFERENCES

1. World Health Organisation (WHO). *Non-Communicable Diseases*, WHO; 2021. Accessed January 2, 2024. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
2. Adua E, Roberts P, Sakyi SA, et al. Profiling of cardio-metabolic risk factors and medication utilisation among type II diabetes patients in Ghana: a prospective cohort study. *Clin Transl Med*. 2017;6:e32. doi:10.1186/s40169-017-0162-5
3. Adua E, Kolog EA, Afrifa-Yamoah E, et al. Predictive model and feature importance for early detection of type II diabetes mellitus. *Transl Med Commun*. 2021;6:17. doi:10.1186/s41231-021-00096-z
4. Adua E, Afrifa-Yamoah E, Kolog EA. Leveraging supervised machine learning for determining the link between suboptimal health status and the prognosis of chronic diseases. In: Wang W, ed., *All Around Suboptimal Health. Advances in Predictive, Preventive and Personalised Medicine*. Springer; 2024:18. doi:10.1007/978-3-031-46891-9_9
5. Golubnitschaja O, Baban B, Boniolo G, et al. Medicine in the early twenty-first century: paradigm and anticipation-EPMA position paper 2016. *EPMA J*. 2016;7:23. doi:10.1186/s13167-016-0072-4
6. Adua E, Afrifa-Yamoah E, Peprah-Yamoah E, et al. Multi-block data integration analysis for identifying and validating targeted N-glycans as biomarkers for type II diabetes mellitus. *Sci Rep*. 2022;12(1):10974. doi:10.1038/s41598-022-15172-z
7. Ruder S, Plank B. Strong baselines for neural semi-supervised learning under domain shift. In: Gurevych I, Miyeo Y eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), *Melbourne Australia*. Association for Computational Linguistics; 2018: 1044-1054. https://aclanthology.org/P18-1096
8. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236-1246. doi:10.1093/bib/bbx044
9. Clifton DA, Niehaus KE, Charlton P, Colopy GW. Health informatics via machine learning for the clinical management of patients. *Yearb Med Inform*. 2015;24:38-43. doi:10.15265/IY-2015-014
10. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347-1358. doi:10.1056/NEJMra1814259
11. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2009.
12. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018
13. Breiman L. *Classification and Regression Trees*. Routledge; 2017.
14. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Net*. 2015;61:85-117. doi:10.1016/j.neunet.2014.09.003
15. Jain AK. Data clustering: 50 years beyond K-means. In: Daelemans W, Goethals B, Morik K, eds., Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. *Lecture Notes in Computer Science*. 5211. Springer; 2008:3-4. doi:10.1007/978-3-540-87479-9_3
16. Jolliffe IT. *Principal Component Analysis*. Springer; 2002.
17. Maaten LVD, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(86):2579-2605. http://jmlr.org/papers/v9/vandermaaten08a.html
18. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT Press; 2018.
19. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. 2018;321:321-331. doi:10.1016/j.neucom.2018.09.013
20. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intel*. 2020;2:369-375. doi:10.1038/s42256-020-0197-y
21. Hotamisligil GS. Inflammation, metaflammation and immunometabolic disorders. *Nature*. 2017;542(7640):177-185. doi:10.1038/nature21363
22. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. *Lancet*. 2014;384(99357):45-52. doi:10.1016/S0140-6736(14)60648-6
23. Anto EO, Boadu WIO, Korsah EE, et al. Unrecognized hypertension among a general adult Ghanaian population: an urban community-based cross-sectional study of prevalence and putative risk factors of lifestyle and obesity indices. *PLoS Global Public Health*. 2023;3(5):e0001973. doi:10.1371/journal.pgph.0001973
24. Anto EO, Frimpong J, Boadu WIO, et al. Cardiometabolic syndrome among general adult population in Ghana: the role of lipid accumulation product, waist circumference-triglyceride index, and triglyceride-glucose index as surrogate indicators. *Health Sci Rep*. 2023;6(7):e1419. doi:10.1002/hsr2.1419
25. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine

clinical data? *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944

26. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol*. 2017;69(21):2657-2664. doi:10.1016/j.jacc.2017.03.571
27. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005
28. Adua E, Memarian E, Afrifa-Yamoah E, et al. N-glycosylation profiling of type 2 diabetes mellitus from baseline to follow-up: an observational study in a Ghanaian population. *Biomark Med*. 2021;15(7):467-480. doi:10.2217/bmm-2020-0615
29. Butler AE, Janson J, Bonner-Weir S, Ritzel R, Rizza RA, Butler PC. β-cell deficit and increased β-cell apoptosis in humans with type 2 diabetes. *Diabetes*. 2003;52:102-110. doi:10.2337/diabetes.52.1.102
30. Sandoval DA, D'Alessio DA. Physiology of proglucagon peptides: role of glucagon and GLP-1 in health and disease. *Physiol Rev*. 2015;95:513-548. doi:10.1152/physrev.00013.2014
31. Gromada J, Chabosseau P, Rutter GA. The α-cell in diabetes mellitus. *Nat Rev Endocrinol*. 2018;14:694-704. doi:10.1038/s41574-018-0097-y
32. Kohei K. Pathophysiology of type 2 diabetes and its treatment policy. *JMAJ*. 2010;53:41-46.
33. Flannick J, Thorleifsson G, Beer NL, et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet*. 2014;46:357-363. doi:10.1038/ng.2915
34. Obirikorang C, Adu EA, Anto EO, et al. Association between transcription factor 7-like-2 polymorphisms and type 2 diabetes mellitus in a Ghanaian population. *Sci*. 2021;3(4):40. doi:10.3390/sci3040040
35. Ng MCY, Park KS, Oh B, et al. Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians. *Diabetes*. 2008;57:2226-2233. doi:10.2337/db07-1583
36. Unoki H, Takahashi A, Kawaguchi T, et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet*. 2008;40:1098-1102. doi:10.1038/ng.208
37. Polat K, Güneş S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Dig Sig Process*. 2007;17:702-710. doi:10.1016/j.dsp.2006.09.005
38. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open*. 2013;3:e002457. doi:10.1136/bmjopen-2012-002457
39. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord*. 2019;19(1):101. doi:10.1186/s12902-019-0436-6
40. Ravaut M, Sadeghi H, Leung KK, et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *npj Dig Med*. 2021;4(1):24. doi:10.1038/s41746-021-00394-8
41. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci Rep*. 2019;9(1):13805. doi:10.1038/s41598-019-49563-6
42. Wang C, Li L, Wang L, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. *Diab Res Clin Pract*. 2013;100(1):111-118. doi:10.1016/j.diabres.2013.01.023
43. Nai-arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. *Proc Comp Sci*. 2015;69:132-142. doi:10.1016/j.procs.2015.10.014
44. Dias-Audibert FL, Navarro LC, de Oliveira DN, et al. Combining machine learning and metabolomics to identify weight gain biomarkers. *Front Bioeng Biotechnol*. 2020;8:6. doi:10.3389/fbioe.2020.00006
45. Peddinti G, Cobb J, Yengo L, et al. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia*. 2017;60(9):1740-1750. doi:10.1007/s00125-017-4325-0
46. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Proc Comp Sci*. 2018;132:1578-1585. doi:10.1016/j.procs.2018.05.122
47. Oh W, Kim E, Castro MR, et al. Type 2 diabetes mellitus trajectories and associated risks. *Big Data*. 2016;4(1):25-30. doi:10.1089/big.2015.0029
48. Marcos-Pasero H, Colmenarejo G, Aguilar-Aguilar E, Ramírez de Molina A, Reglero G, Loria-Kohen V. Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques. *Sci Rep*. 2021;11(1):1910. doi:10.1038/s41598-021-81205-8
49. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the henan rural cohort study. *Sci Rep*. 2020;10(1):4406. doi:10.1038/s41598-020-61123-x
50. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019;19(1):211. doi:10.1186/s12911-019-0918-5
51. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep*. 2020;10(1):11981. doi:10.1038/s41598-020-68771-z
52. Bikbov B, Purcell CA, Levey AS, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2020;395(10225):709-733. doi:10.1016/S0140-6736(20)30045-3
53. Baumgarten M, Gehr T. Chronic kidney disease: detection and evaluation. *Am Fam Physician*. 2011;84(10):1138-1148.
54. Dovgan E, Gradišek A, Luštrek M, et al. Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. *PLoS One*. 2020;15(6):e0233976. doi:10.1371/journal.pone.0233976
55. 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. doi:10.1038/s41586-019-1390-1
56. Xiao J, Ding R, Xu X, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J Transl Med*. 2019;17(1):119. doi:10.1186/s12967-019-1860-0
57. Glazyrin YE, Veprintsev DV, Ler IA, et al. Proteomics-based machine learning approach as an alternative to conventional biomarkers for differential diagnosis of chronic kidney diseases. *Int J Mol Sci*. 2020;21(13):4802. doi:10.3390/ijms21134802
58. Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *Int Urol Nephrol*. 2016;48(12):2069-2075. doi:10.1007/s11255-016-1346-4
59. Al-Hyari AY, Al-Taee AM, Al-Taee MA Clinical decision support system for diagnosis and management of Chronic Renal Failure. Paper presented at: *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*; 2013.
60. Lee J, Warner E, Shaikhouni S, et al. Unsupervised machine learning for identifying important visual features through bag-of-words using histopathology data from chronic kidney disease. *Sci Rep*. 2022;12(1):4832. doi:10.1038/s41598-022-08974-8
61. Bueno G, Fernandez-Carrobles MM, Gonzalez-Lopez L, Deniz O. Glomerulosclerosis identification in whole slide

images using semantic segmentation. *Comput Methods Programs Biomed*. 2020;184:105273. doi:10.1016/j.cmpb.2019.105273

62. Lee Y, Ryu J, Kang MW, et al. Machine learning-based prediction of acute kidney injury after nephrectomy in patients with renal cell carcinoma. *Sci Rep*. 2021;11(1):15704. doi:10.1038/s41598-021-95019-1
63. Kim HW, Heo SJ, Kim JY, Kim A, Nam CM, Kim BS. Dialysis adequacy predictions using a machine learning method. *Sci Rep*. 2021;11(1):15417. doi:10.1038/s41598-021-94964-1
64. Afrifa-Yamoah E, Adua E, Anto EO, et al. Conceptualised psycho-medical footprint for health status outcomes and the potential impacts for early detection and prevention of chronic diseases in the context of 3P medicine. *EPMA J*. 2023;14:585-599. doi:10.1007/s13167-023-00344-2
65. Mets OM, Vliegenthart R, Gondrie MJ, et al. Lung cancer screening CT-based prediction of cardiovascular events. *JACC Cardiovasc Imaging*. 2013;6(8):899-907. doi:10.1016/j.jcmg.2013.02.008
66. Maroules CD, Hamilton-Craig C, Branch K, et al. Coronary artery disease reporting and data system (CAD-RADSTM): inter-observer agreement for assessment categories and modifiers. *J Cardiovasc Comp Tomogr*. 2018;12(2):125-130. doi:10.1016/j.jcct.2017.11.014
67. Takx RAP, de Jong PA, Leiner T, et al. Automated coronary artery calcification scoring in non-gated chest CT: agreement and reliability. *PLoS One*. 2014;9(3):e91239. doi:10.1371/journal.pone.0091239
68. Bruse JL, Schievano S, Zuluaga MA, et al. Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches. *IEEE Trans Biomed Eng*. 2017;64(10):2373-2383. doi:10.1109/TBME.2017.2655364
69. Wojnarski CM, Roselli EE, Idrees JJ, et al. Machine-learning phenotypic classification of bicuspid aortopathy. *J Thorac Cardiovasc Surg*. 2018;155(2):461-469.e4. doi:10.1016/j.jtcvs.2017.08.123
70. Ward A, Sarraju A, Chung S, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *npj Dig Med*. 2020;3(1):125. doi:10.1038/s41746-020-00331-1
71. Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. 2020;10(1):5245. doi:10.1038/s41598-020-62133-5
72. Quesada JA, Lopez-Pineda A, Gil-Guillén VF, et al. Machine learning to predict cardiovascular risk. *Int J Clin Pract*. 2019;73(10):e13389. doi:10.1111/ijcp.13389
73. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14(5):e0213653. doi:10.1371/journal.pone.0213653
74. Panaretos D, Koloverou E, Dimopoulos AC, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the ATTICA study. *Br J Nutr*. 2018;120(3):326-334. doi:10.1017/S0007114518001150
75. Yu K, Lee KH, Afrifa-Yamoah E, et al. Identification of candidate congenital heart defects biomarkers by applying a random forest approach on DNA methylation data. *Atherosclerosis*. 2021;331:e218-e219. doi:10.1016/j.atherosclerosis.2021.06.670
76. Brink M. K-ras oncogene mutations in sporadic colorectal cancer in the Netherlands Cohort Study. *Carcinogenesis*. 2003;24(4):703-710. doi:10.1093/carcin/bgg009
77. Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. *Proc Natl Acad Sci USA*. 2003;100(3):776-781. doi:10.1073/pnas.0334858100
78. Muller PAJ, Vousden KH. p53 mutations in cancer. *Nat Cell Biol*. 2013;15(1):2-8. doi:10.1038/ncb2641
79. Hirohashi S. Inactivation of the E-cadherin-mediated cell adhesion system in human cancers. *Am J Pathol*. 1998;153(2):333-339. doi:10.1016/S0002-9440(10)65575-7
80. Kumar V, Abbas AK, Aster JC. *Robbins Basic Pathology e-Book*. Elsevier Health Sciences; 2017.
81. Cohen S, Levi-Montalcini R, Hamburger V. A nerve growth-stimulating factor isolated from sarcom as 37 and 180. *Proc Natl Acad Sci USA*. 1954;40(10):1014-1018. doi:10.1073/pnas.40.10.1014
82. Roberts AB, Anzano MA, Lamb LC, et al. Isolation from murine sarcoma cells of novel transforming growth factors potentiated by EGF. *Nature*. 1982;295(5849):417-419. doi:10.1038/295417a0
83. Waterfield MD, Scrace GT, Whittle N, et al. Platelet-derived growth factor is structurally related to the putative transforming protein p28 sis of simian sarcoma virus. *Nature*. 1983;304(5921):35-39. doi:10.1038/304035a0
84. Aaronson SA. Growth factors and cancer. *Science*. 1991;254(5035):1146-1153. doi:10.1126/science.1659742
85. Hassler MR, Egger G. Epigenomics of cancer—emerging new concepts. *Biochimie*. 2012;94(11):2219-2230. doi:10.1016/j.biochi.2012.05.007
86. Jones PA, Baylin SB. The epigenomics of cancer. *Cell*. 2007;128(4):683-692. doi:10.1016/j.cell.2007.01.029
87. Giles RH, Peters DJM, Breuning MH. Conjunction dysfunction: CBP/p300 in human disease. *TIG*. 1998;14(5):178-183. doi:10.1016/S0168-9525(98)01438-3
88. Seligson DB, Horvath S, Shi T, et al. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*. 2005;435(7046):1262-1266. doi:10.1038/nature03672
89. Fraga MF, Ballestar E, Villar-Garea A, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet*. 2005;37(4):391-400. doi:10.1038/ng1531
90. Savage N. How AI is improving cancer diagnostics. *Nature*. 2020;579(7800):S14-S16. doi:10.1038/d41586-020-00847-2
91. Zhen L, Liu X, Yegang C, et al. Accuracy of multiparametric magnetic resonance imaging for diagnosing prostate cancer: a systematic review and meta-analysis. *BMC Cancer*. 2019;19(1):1244. doi:10.1186/s12885-019-6434-2
92. Kim JY, Kim SH, Kim YH, Lee HJ, Kim MJ, Choi MS. Low-risk prostate cancer: the accuracy of multiparametric MR imaging for detection. *Radiology*. 2014;271(2):435-444. doi:10.1148/radiol.13130801
93. Maruvada P, Wang W, Wagner PD, Srivastava S. Biomarkers in molecular medicine: cancer detection and diagnosis. *Biotechniques*. 2005;38(4S):9-15. doi:10.2144/05384SU04
94. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med*. 2008;359(16):1675-1684. doi:10.1056/NEJMoa0803545
95. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920
96. He B, Bergenstråhle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng*. 2020;4(8):827-834. doi:10.1038/s41551-020-0578-x
97. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*. 2015;33(8):831-838. doi:10.1038/nbt.3300
98. Huang K, Xiao C, Glass LM, Critchlow CW, Gibson G, Sun J. Machine learning applications for therapeutic tasks with genomics data. *Patterns*. 2021;2(10):100328. doi:10.1016/j.patter.2021.100328
99. Yin B, Balvert M, van der Spek RAA, et al. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. *Bioinformatics*. 2019;35(14):i538-i547. doi:10.1093/bioinformatics/btz369

100. World Health Organisation. *Asthma*. WHO; 2021. Accessed January 1, 2024. https://www.who.int/news-room/fact-sheets/detail/asthma
101. Thurlbeck WM, Müller NL. Emphysema: definition, imaging, and quantification. *Am J Roentgenol*. 1994;163(5):1017-1025. doi:10.2214/ajr.163.5.7976869
102. Taraseviciene-Stewart L, Voelkel NF. Molecular pathogenesis of emphysema. *J Clin Invest*. 2008;118(2):394-402. doi:10.1172/JCI31811
103. Sharafkhaneh A, Hanania NA, Kim V. Pathogenesis of emphysema: from the bench to the bedside. *Proc Am Thorac Soc*. 2008;5(4):475-477. doi:10.1513/pats.200708-126ET
104. Giri PC, Chowdhury AM, Bedoya A, et al. Application of machine learning in pulmonary function assessment where are we now and where are we going? *Front Physiol*. 2021;12:678540. doi:10.3389/fphys.2021.678540
105. Gold WM, Koth LL. Pulmonary function testing. *Murray Nadel's Textbook of Respiratory Medicine*. 2016:407-435.e18. doi:10.1016/B978-1-4557-3383-5.00025-7
106. Aaron SD, Boulet LP, Reddel HK, Gershon AS. Underdiagnosis and overdiagnosis of asthma. *Am J Respir Crit Care Med*. 2018;198(8):1012-1020. doi:10.1164/rccm.201804-0682CI
107. Kaplan A, Cao H, FitzGerald JM, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *J Allergy Clin Immunol Pract*. 2021;9(6):2255-2261. doi:10.1016/j.jaip.2021.02.014
108. Efared B, Ebang-Atsame G, Rabiou S, et al. The diagnostic value of the bronchoalveolar lavage in interstitial lung diseases. *J Negat Results Biomed*. 2017;16(1):4. doi:10.1186/s12952-017-0069-0
109. Ryu JH, Daniels CE, Hartman TE, Yi ES. Diagnosis of interstitial lung diseases. *Mayo Clin Proc*. 2007;82(8):976-986. doi:10.4065/82.8.976
110. Casoni GL, Tomassetti S, Cavazza A, et al. Transbronchial lung cryobiopsy in the diagnosis of fibrotic interstitial lung diseases. *PLoS One*. 2014;9(2):e86716. doi:10.1371/journal.pone.0086716
111. Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. *Curr Opin Pulm Med*. 2018;24(2):117-123. doi:10.1097/MCP.0000000000000459
112. Finamore P, Spruit MA, Schols JMGA, Antonelli Incalzi R, Wouters EFM, Janssen DJA. Clustering of patients with end-stage chronic diseases by symptoms: a new approach to identify health needs. *Aging Clin Exp Res*. 2021;33(2):407-417. doi:10.1007/s40520-020-01549-5
113. Mostafaei S, Kazemnejad A, Azimzadeh Jamalkandi S, et al. Identification of novel genes in human airway epithelial cells associated with chronic obstructive pulmonary disease (COPD) using machine-based learning algorithms. *Sci Rep*. 2018;8(1):15775. doi:10.1038/s41598-018-33986-8
114. Swaminathan S, Qirko K, Smith T, et al. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS One*. 2017;12(11):e0188532. doi:10.1371/journal.pone.0188532
115. Angelini E, Dahan S, Shah A. Unravelling machine learning: insights in respiratory medicine. *Eur Respir J*. 2019;54(4):1901216. doi:10.1183/13993003.01216-2019
116. Tanabe N, Sato S, Oguma T, et al. Associations of airway tree to lung volume ratio on computed tomography with lung function and symptoms in chronic obstructive pulmonary disease. *Respir Res*. 2019;20(1):77. doi:10.1186/s12931-019-1047-5
117. Jacob J, Bartholmai BJ, Rajagopalan S, et al. Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based CT analysis with conventional severity measures. *Eur Respir J*. 2017;49(1):1601011. doi:10.1183/13993003.01011-2016
118. Ioachimescu OC, Stoller JK. An alternative spirometric measurement. area under the expiratory flow-volume curve. *Ann Am Thorac Soc*. 2020;17(5):582-588. doi:10.1513/AnnalsATS.201908-613OC
119. Bodduluri S, Nakhmani A, Reinhardt JM, et al. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. *JCI Insight*. 2020;5(10):e132781. doi:10.1172/jci.insight.132781
120. Kelly JT, Campbell KL, Gong E, et al. The Internet of Things: impact and implications for Health Care Delivery. *J Med Internet Res*. 2020;22(11):e20135. doi:10.2196/20135
121. Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. *Health Informatics J*. 2019;25(4):1201-1218. doi:10.1177/1460458217751015
122. Straub RH. The brain and immune system prompt energy shortage in chronic inflammation and ageing. *Nat Rev Rheumatol*. 2017;13(12):743-751. doi:10.1038/nrrheum.2017.172
123. Straub RH, Schradin C. Chronic inflammatory systemic diseases: an evolutionary trade-off between acutely beneficial but chronically harmful programs. *Evol Med Public Health*. 2016;2016(1):eow001. doi:10.1093/emph/eow001
124. Chen L, Deng H, Cui H, et al. Inflammatory responses and inflammation-associated diseases in organs. *Oncotarget*. 2018;9(6):7204-7218. doi:10.18632/oncotarget.23208
125. Furman D, Campisi J, Verdin E, et al. Chronic inflammation in the etiology of disease across the life span. *Nat Med*. 2019;25(12):1822-1832. doi:10.1038/s41591-019-0675-0
126. Hu FB, Meigs JB, Li TY, Rifai N, Manson JE. Inflammatory markers and risk of developing type 2 diabetes in women. *Diabetes*. 2004;53(3):693-700. doi:10.2337/diabetes.53.3.693
127. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of paediatric inflammatory bowel disease using machine learning. *Sci Rep*. 2017;7(1):2427. doi:10.1038/s41598-017-02606-2
128. Schneider N, Sohrabi K, Schneider H, Zimmer KP, Fischer P, de Laffolie J. Machine learning classification of inflammatory bowel disease in children based on a large real-world pediatric cohort CEDATA-GPGE® registry. *Front Med*. 2021;8:666190. doi:10.3389/fmed.2021.666190
129. McDonnell M, Harris RJ, Borca F, et al. High incidence of glucocorticoid-induced hyperglycaemia in inflammatory bowel disease: metabolic and clinical predictors identified by machine learning. *BMJ Open Gastroenterol*. 2020;7(1):e000532. doi:10.1136/bmjgast-2020-000532
130. Manandhar I, Alimadadi A, Aryal S, Munroe PB, Joe B, Cheng X. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. *Am J Physiol Gastroint Liver Physiol*. 2021;320(3):G328-G337. doi:10.1152/ajpgi.00360.2020
131. Kraszewski S, Szczurek W, Szymczak J, Reguła M, Neubauer K. Machine learning prediction model for inflammatory bowel disease based on laboratory markers. working model in a discovery cohort study. *J Clin Med*. 2021;10(20):4745. doi:10.3390/jcm10204745
132. Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats. *J Vet Diagn Invest*. 2016;28(6):679-687. doi:10.1177/1040638716657377
133. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626
134. Ribeiro MT, Singh S, Guestrin C "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*; 2016:1135-1144.
135. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

136. Mitra N, Roy J, Small D. The future of causal inference. *Am J Epidemiol*. 2022;191(10):1671-1676. doi:10.1093/aje/kwac108
137. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care. *AMA J Ethics*. 2019;21(2):E167-E179. doi:10.1001/amajethics.2019.167
138. Dancy CL, Saucier PK. AI and blackness: toward moving beyond bias and representation. *IEEE Trans Technol Soc*. 2022;3(1):31-40. doi:10.1109/TTS.2021.3125998
139. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. 2019;20(3):405-410. doi:10.3348/kjr.2019.0025
140. Yang WH, Shao Y, Xu YW. Guidelines on clinical research evaluation of artificial intelligence in ophthalmology (2023). *Int J Ophthalmol*. 2023;16(9):1361-1372. doi:10.18240/ijo.2023.09.02
141. Stevens CA, Lyons AR, Dharmayat KI, et al. Ensemble machine learning methods in screening electronic health records: a scoping review. *Dig Health*. 2023;9:205520762311732. doi:10.1177/20552076231173225
142. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18
143. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Ann Rev Biomed Data Sci*. 2018;1:53-68. doi:10.1146/annurev-biodatasci-080917-013315
144. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc*. 2017;24(2):361-370. doi:10.1093/jamia/ocw112
145. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *J Biomed Inf*. 2020;101:103337. doi:10.1016/j.jbi.2019.103337
146. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z
147. Char DS, Shah NH, Magnus D. Implementing machine learning in health care- addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229
148. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med*. 2020;260:113172. doi:10.1016/j.socscimed.2020.113172
149. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med*. 2020;26(1):16-17. doi:10.1038/s41591-019-0649-2
150. Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics*. 2020;26(5):2749-2767. doi:10.1007/s11948-020-00228-y

How to cite this article: Afrifa-Yamoah E, Adua E, Peprah-Yamoah E, et al. Pathways to chronic disease detection and prediction: mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges. *Chronic Dis Transl Med*. 2025;11:1-21. doi:10.1002/cdt3.137