Title:

Exposomics and Cardiovascular Diseases: A Scoping Review of Machine Learning Approaches

Short Running Head (≤ 50 characters):

ML in Exposomics and CVD

Authors and Affiliations:

Katerina D. Argyri[1], Ioannis K. Gallos[1], Angelos Amditis[1], Dimitra D. Dionysiou[1]

[1] Institute of Communication and Computer Systems, National Technical University of Athens, Zografos Campus, 15780 Athens, Greece

Corresponding Author:

Name: Katerina D. Argyri

Mailing Address:

Institute of Communication and Computer Systems (ICCS)

National Technical University of Athens

Zografos Campus, 15780 Athens, Greece

Email: katerina.argyri@iccs.gr

ABSTRACT

The principal objective served by this article is to identify key literature and provide an overview of the breadth of research in the field of machine learning applications on exposomics data with a focus on cardiovascular diseases. Secondarily, this study aimed at identifying common limitations and meaningful directives to be addressed in the future.

Most of the identified literature focuses on Disease Understanding, Prevention, and Management and the remaining on healthcare resource planning and management. Linear, non-linear, and ensemble machine learning methods have been applied, with tree-based methods generally being the most popular, and Random Forests specifically reported as being the best performer most of the time. Non-traditional CVD risk factors spanning behavioural/ lifestyle, socio-economic and a wide range of environmental factors have been investigated in the identified literature. According to our findings, the most reported category is the environmental risk factors which are often considered alone rather than in conjunction with the others.

Applications of machine Learning have been a substantial accelerating factor on the field, enabling the analysis of high dimensional data, improving accuracy, investigating novel risk factors and expanding our knowledge on the impact of exposome on cardiovascular diseases. However, several challenges persist, including data heterogeneity and the lack of standardized protocols, which introduces bias and hinders reproducibility and comparability across studies, as well as concerns about the validity of inference when applying machine learning methods to identify associations between exposure and health outcomes. Addressing these issues requires collaborative and interdisciplinary effort to integrate standardized exposome data frameworks akin to those of other –omics fields, apply causal inference methods to validate findings, and further expand the use of Explainable Artificial Intelligence to build insights and enable comparability and understanding.

Keywords: *machine learning, deep learning, exposomics, ExWAS, cardiovascular disease, environmental health*

Introduction

Since the mid-20th century, cardiovascular (CV) diseases (CVDs) have emerged as the leading cause of death globally. In 2021, CVDs caused approximately 20.5 million deaths, accounting for approximately one third of all deaths worldwide [1]. In addition to this significant epidemiological burden, CVDs are estimated to impose a substantial financial burden, with the World Health Organization (WHO) recommending that health expenditures reach at least 5% of a country's GDP [1]. They represent a large group of diseases attributed to a complex interplay between intrinsic risk factors, such as genetic predisposition, biological sex, and age, and lifetime exposure to environmental and behavioral risk factors which are considered at least partially modifiable [2]. Environmental exposures to ambient and indoor air pollution, noise, extreme temperatures, second-hand smoke, and chemicals, among other factors, have been recognized by many global health and environment agencies as significant contributors to the high burden of CVD. It is estimated that approximately 25% of deaths from Non-Communicable Diseases including primarily CVD-related deaths are attributable to environmental risks [2]. In recent years, there has been growing recognition of the importance of modifiable factors as a whole in efforts to alleviate the burden of disease [3].

While preventive interventions targeting traditional risk factors (e.g., blood pressure and cholesterol management) have aided in reducing CVD incidents, it remains a major global health issue, highlighting the need for new approaches. Unlike DNA, which remains unchangeable, there are primary modifiable contributors to CVD that are amenable to prevention and policy initiatives aimed at promoting cardiovascular health. The fact that environmental CV risk factors are inherently preventable leads to the actionable conclusion that reducing them is a key-step to alleviating the global burden of cardiovascular disease. In this context, investigations are increasingly directed towards non-traditional risk factors that are present in the built, natural, and socio-economic environments comprising the “Exposome” [3], [4].

Exposome, the youngest member of the widely acknowledged -ome family, was first coined in 2005 by Dr. Christopher Wild, then-director of the International Agency for Research on Cancer (IARC), to complement the human genome and address the limitations of genetic research in explaining chronic disease etiology [5]. Aiming to fill this critical knowledge gap, the exposome was conceptualized as a systematic approach to measuring the entirety of environmental exposures encountered by an individual from conception onwards, including chemical, physical, biological, and lifestyle factors. In 2014, Gary Miller and Dean Jones expanded the exposome so as to emphasize diet, behavior, and endogenous processes, particularly focusing on biological responses to these exposures [6]. From their perspective, the exposome captures the essence of “nurture” in one of the oldest philosophical discussions of “nature” vs “nurture”, representing the summation and integration of external forces acting upon our genome throughout our lifespan. This includes factors such as diet, living environment, air quality, social interactions, lifestyle choices like smoking and exercise, and inherent metabolic and cellular activities. Measuring a quantity for the exposome serves as a biological index of our “nurture”, contextualizing the impact of specific exposures on health. This expansion and refinement of exposomics led to the inclusion of metabolomics moving beyond solely exposure-focused approaches to capture biological endpoints accompanied with substantial changes [7]. By exploring all these factors that constitute the exposome, researchers aim to understand and pinpoint modifiable risk factors and devise targeted interventions through active personal measures, behavioral strategies, novel policies, urban landscape reforms etc. to promote health and prevent disease across lifespan.

Along the lines of Genome-Wide Association studies (GWAS) and the identification of genetic basis of many complex traits and diseases [8], there have also been efforts to identify the “environmental” risk factors in the so-called ‘Environment-Wide association studies’ [9]. Finally, the wider term "Exposome-Wide Association Study" (ExWAS) has been proposed as a standardization term and a method designed to systematically investigate the connections between phenotypes and various exposures beyond the classic understanding of environmental factors, in line with the expanded definition of Exposome [6]. This approach facilitates the identification of significant correlations while addressing the challenge of multiple comparisons, aimed at finding the exposomic basis of a disease or trait [10]. Examples focused on CVD can be found in literature [11].

Different types of environmental risk factors have been reported in literature. First, chemical pollution, which spans air, soil, water and occupational pollution and is currently acknowledged as the most significant environmental cause of disease and premature death in the world [12]. Air pollution's main health risks stem from particulate matter which is well known to be linked to CVDs [13], while the water pollution hazards arise from unsafe sources. The health impacts of soil pollution can be attributed mainly to heavy metals, deforestation, over-fertilization, and pesticides, with nano- and microplastics emerging as a threat. Although lead toxicity primarily results from water and soil pollution, it warrants separate consideration due to its widespread environmental presence. There is

a close link between water and soil pollution, as polluted soil can contaminate surface and groundwater. Heavy metals and metalloids are particularly concerning due to their contribution to cardiovascular issues through oxidative stress and inflammation [14]. Chemical pollutants, particularly ambient air pollution, have garnered significant attention from research organizations assessing their impacts [15], [16]. These pollutants were also considered in the Global Burden of Disease study [17], [18].

Second, non-chemical pollution, such as transportation noise, light pollution and lack of green spaces, has been also shown to have substantial impact on CVD [4]. On the one hand, as urban areas expand and the demand for transportation increases, noise pollution is expected to rise. Research has demonstrated a connection between noise pollution and heightened CV risk, driven by mechanisms such as stress, sleep disruption, and increased inflammation. Furthermore, studies have shown an association between noise pollution and elevated risks of arterial hypertension, dyslipidemia, obesity, and type II diabetes mellitus. Thus, noise pollution not only directly impacts the CV system but also indirectly elevates the risk of developing traditional CV risk factors [19]. On the other end, it is known that nocturnal light pollution is associated with abnormal changes in circadian rhythms, which in turn may be linked to an increased risk of CVD [20], [21], increased blood pressure and risk of hospitalization for CVD [22]. While there is extensive and robust evidence linking noise pollution to an increased risk of CVD, the role of nocturnal light pollution in CV pathology is less studied. To validate and substantiate this association, further research is needed to figure out potential thresholds of 'safe' or 'acceptable' artificial light levels [23]. Finally, latest metanalyses have highlighted strong evidence on the link of green spaces and cardiovascular health [24]. For a comprehensive review on the epidemiology and pathophysiology of environmental stressors with a focus on CVDs, the interested reader may refer to pertinent literature [3], [4], [14], [25].

Apart from the classic environmental risk factors such as pollution, the exposome encompasses a wide range of lifestyle and socioeconomic factors. Growing evidence suggests that the CVD is characterized by socioeconomic inequalities [26], [27], [28]. These inequalities, most often assessed in terms of income, occupation and education, appear to be closely related with dietary habits and harmful lifestyle choices such as smoking and alcohol consumption [26]. Lifestyle interventions at targeting nutrition and physical activity have been shown to benefit cardiometabolic risk factors such as Body Mass Index (BMI), triglycerides and Low-Density Lipoproteins (LDL) in individuals at risk [29]. The overwhelming evidence of the impact of lifestyle on traditional cardiometabolic biomarkers has led to the emergence of the framework of "lifestyle medicine" which leverages lifestyle interventions to maintain cardiovascular health [30]. Most often these interventions focus primarily on diet, physical activity, perceived stress and anxiety and mitigation of harmful habits such as tobacco use and alcohol consumption.

While traditional epidemiology, which relies on well-established study designs, is a valuable tool for investigating the relationship between environmental exposures and health outcomes, the heterogeneous and dynamic nature of exposomics data presents new challenges. The exposome considers many environmental stressors simultaneously, as opposed to the one-by-one approach typically used in epidemiological research. This necessity, along with the large number of exposures, poses challenges such as increased complexity, high dimensionality, strong correlations between variables, and the need to understand both the combined effects and the causal structure between exposure risk factors and health outcomes. The Machine Learning (ML) toolbox, including Deep Learning (DL) techniques, is particularly well-suited to address these challenges [31]. By leveraging this set of tools, researchers can efficiently reduce data dimensionality and identify complex patterns and interactions; integrate diverse data types; enhance causal inference capabilities and uncover intricate

relationships between exposomic factors and health outcomes; provide deeper insights into how combined environmental and behavioral factors influence CVD.

ML techniques, including DL, have been gaining popularity for quite a few years now in the analysis and integration of diverse kinds of -omics data (e.g., genomics, proteomics, metabolomics) especially within the context of precision medicine [32], [33]. However, applications on exposomics are still in early stages. This is partly because the field itself is relatively new (the term was only coined in 2005) but also due to the very nature and scale of the exposome which covers all exposures from conception to death. Thus, data is expected to exhibit significant heterogeneity (e.g., lifestyle factors, socioeconomic variables, biological responses etc.) as well as spatial and temporal variability, requiring integration from multiple sources and technologies. Due to these challenges, researchers proposed a roadmap for the use of federated approaches to accelerate research in the field [34]. Federated infrastructures and learning enable institutions to collaboratively analyze data while ensuring security and compliance with ethical, legal, and regulatory aspects. Leveraging federated technologies, researchers aim to utilize large, decentralized datasets and enable the sharing, access, and integration of diverse longitudinal exposomics data, supporting among others, ExWAS studies, functional exposomics, and precision health applications. Finally, DL, with its capacity for large-scale processing of complex and disparate multi-modal datasets, holds promise for advancing the understanding of the implications of the exposomics in the disease [35].

To date, to the best of our knowledge, the only review paper on the application of ML techniques on exposomics data for CVD-related investigations exclusively includes social determinants of health [36]. The primary objective of our work is to address this literature gap, by identifying key studies and providing an overview of the breadth of research in the field of ML applications on exposomics data with a focus on cardiovascular diseases. Additionally, we identify common limitations and propose meaningful directives for future research.

The remainder of the article is structured as follows: The Methods section reports the adopted literature search strategy and outlines the three research questions guiding our analysis. Subsequently, the Results section provides an overview of the relevant literature in the context of the three main axes of our work. Finally, we summarize our findings, discuss remarkable insights, and we identify and address the limitations of this study.

Methods

Search strategy

An extensive search of PubMed, IEEE Xplore and Google Scholar has been conducted to thoroughly identify articles on ML applications exploring potential associations of exposomics factors with CVD-related outcomes. In addition to the broad terms ‘exposome’ and ‘exposomics’, we derived key search terms based on the core aspects of the exposome as defined by the European Human Exposome Network [37]. The latter identifies environmental, lifestyle-related, and socio-economic factors as the key exposures constituting the exposome. These key factors have been adopted as the keywords for this review and searched within the titles and abstracts of published studies. Specifically, the body of work reported herein has been obtained using the following search query:

("exposomics"[Title/Abstract] OR "exposome"[Title/Abstract] OR "environment"[Title/Abstract] OR "socio-economic"[Title/Abstract] OR "lifestyle"[Title/Abstract]) AND (cardiovascular disease[Title/Abstract]) AND ("machine learning"[Title/Abstract])

Database searches have been supplemented with studies identified through manual searching. This review has been conducted using a systematic approach consistent with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Extension for Scoping Reviews (PRISMA-ScR) [38]. No temporal bounds have been imposed on the publication dates.

Study Exclusion Criteria

Overall, we have applied the following exclusion criteria to ensure that this review maintains focus and relevance to the desired scope:

- Non-English language articles,
- Non-peer reviewed items (e.g. gray literature such as pre-prints, technical reports, web-based guidelines etc.),
- Items for which the full text was not accessible (e.g., articles presented at conferences as abstracts),
- Articles exclusively focused on traditional statistical methods,
- Articles not exploiting exposomics data,
- Articles assessing intermediate risk factors such as blood pressure rather than direct CVD outcomes,
- Articles exploiting biomarkers as an input capturing endogenous metabolic changes that may result from various internal and external factors,
- Articles focusing on populations with pre-existing comorbidities (cancer, liver disease, etc.),
- Review articles,
- Duplicate articles.

Research questions

The analysis to be presented revolves around the following Research Questions (RQ).

RQ1. How are studies categorized based on their primary use case, and which are their key distinguishing features?

RQ2. Which machine learning algorithms are more popular for investigating exposomics impact on CVDs and which seem to be ranking among the top performers?

RQ3. Which exposomic factors have been investigated and which have been identified as potentially good predictors?

The first research question (RQ1) focuses on categorizing the studies with respect to the context of application of the developed system. The second research question (RQ2) explores the ML algorithms preferred by the research community in an attempt to rank specific categories of algorithms with respect to their popularity in retrieved literature. Finally, the third research question (RQ3) aspires to

shed light on specific categories of promising predictors and hopefully to identify new directions of exposomics variables investigations.

Results

A general overview

Firstly, an overview of the retrieved literature is provided, highlighting the temporal distribution of publications, the geographical distribution of utilized datasets, and the specific ML tasks addressed.

The timeline of publications identified by using the specified search criteria, as displayed in the bar plot below (Fig.1), reveals a pronounced increase of relevant studies from 2021 onwards.

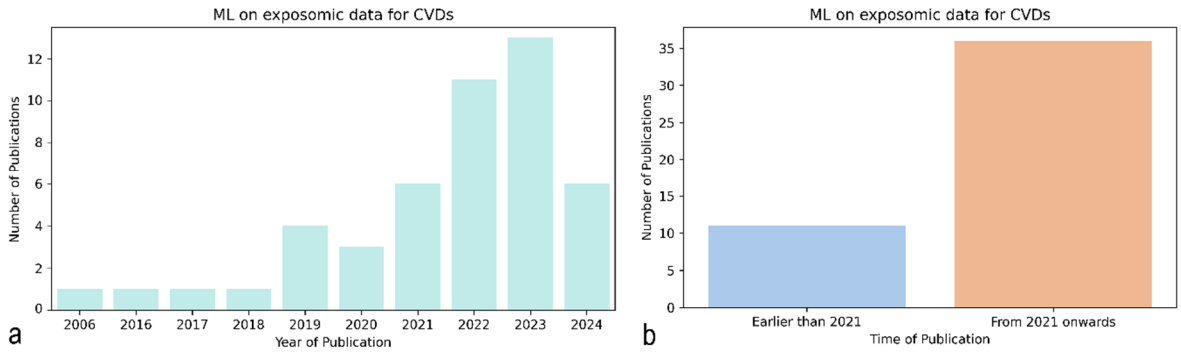

Figure 1. Time evolution of publications leveraging ML techniques to exploit exposomics datasets for CVD-related investigations a. Year of publication (x-axis) and number of identified studies per year (y-axis), b. Time of publication ('earlier than 2021' and 'from 2021 onwards')

Concerning the spatial distribution, the majority of identified studies reporting the origin of their datasets have used datasets collected within US or Asian territory (Fig. 2).

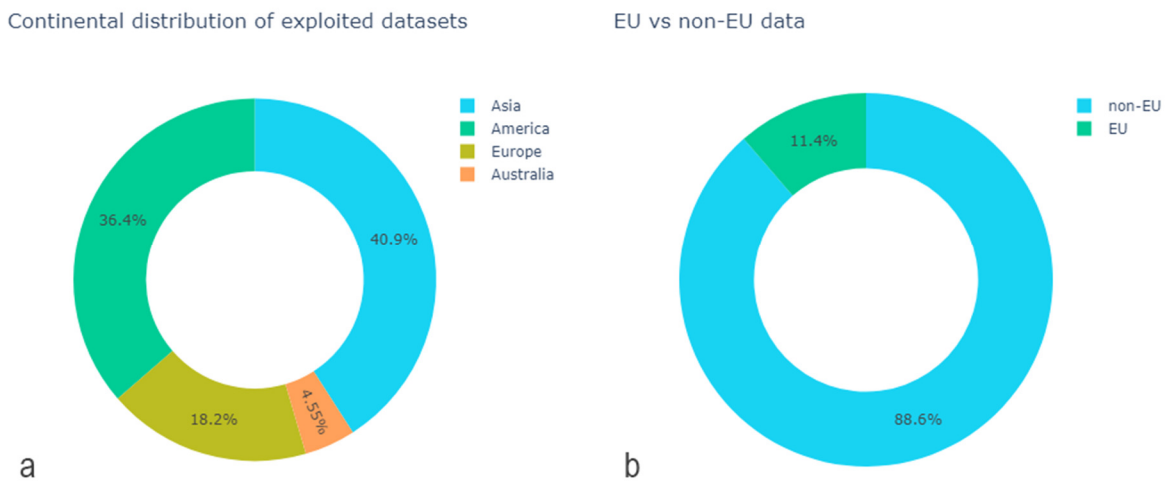

Figure 2. a. Continental distribution of exploited datasets, b. Data of EU vs non-EU origin.

Lastly, regarding the framing of the ML problems, an important observation is the pronounced, almost exclusive adoption of a supervised context. The most widely adopted problem framing is that of classification tasks, followed by regression tasks (Fig. 3).

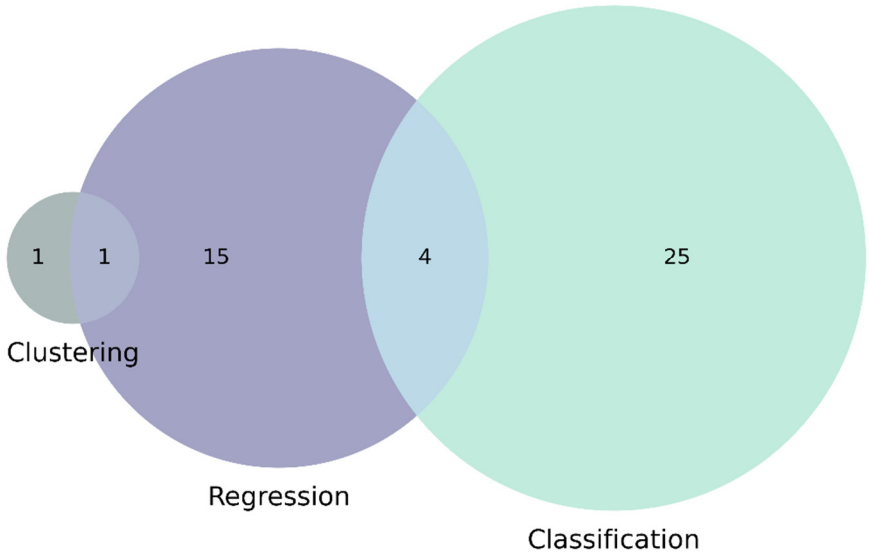

Figure 3. A Venn diagram of (core prediction) ML tasks categories encountered in literature focusing on exploring CVDs

Research questions

Proceeding with the analysis conducted, we elaborate on identified categories of studies with respect to the aspects corresponding to the research questions:

RQ1. How are studies categorized based on their primary use case, and what are their key distinguishing features?

Two main categories of studies have been identified based on the application context. The first category focuses on disease understanding, prevention and management. These studies either pertain to direct citizen/patient care or inform population-level interventions and policy measures. The second category aims at assisting healthcare resource planning and management. It includes studies focused on forecasting hospital admissions, healthcare costs, and optimizing resources to ensure better system efficiency, and commonly aspires to build and eventually deploy an early-warning system for medical resource allocation and management.

Disease understanding, prevention and management are the primary focus of the retrieved studies. In this context, CVD incidence/prevalence or CVD-induced mortality are commonly used as target outcomes. The overall goal of these studies is to provide insights for reducing disease burden and mortality. A significant body of literature focuses on the development of personalized CVD risk assessment tools focusing on individual-level associations [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52]. In [53] this perspective is extended by exploring the link between

maternal PM10 exposure during pregnancy and congenital heart defects in fetuses. Environmental Risk Score (ERS) measures (e.g. [54], [55]**)** have been proposed as a useful tool for characterizing cumulative risk from multiple exposures and are commonly used to estimate exposomic liability for a particular health outcome at an individual level. The multi-pollutant framework is gaining traction with the goal of better understanding the real-world health impacts of pollutant mixtures and facilitate risk stratification and targeted preventive interventions. It should be noted that several studies use the number of hospital admissions as a proxy to predict CVD incidence at a regional scale without considering aspects of healthcare system efficiency [56], [57], [58] and as such, we have considered them as part of the 'Disease understanding, prevention and management' category. Another portion of the literature deals with the prediction of CVD severity (e.g. [59], [60]) and CVD-induced mortality [61], [62], [63], [64]. As expected, significant focus is placed on the identification of novel risk factors for CVD incidence (e.g., [11], [65]) and CVD prevalence [35], [66] through the investigation of associations of variables of interest with selected outcomes. Another interesting research direction is the use of spatiotemporal data for spatially analyzing CVD prevalence [35], [66] or CVD-induced mortality (e.g. [60], [61]), and combining this knowledge with geographically targeted interventions, including urban-planning.

The second category, pertaining to healthcare resource planning and management, focuses on forecasting hospital admissions, healthcare costs, and optimizing resources to ensure better patient care and system efficiency. A key objective is the development of an early-warning system for healthcare resource allocation [67], [68], [69], [70]. Additionally, some studies address the financial sustainability of healthcare services, identifying key predictors of county-level healthcare costs [71] or high health-cost users among individuals with CVD [72].

RQ2. Which machine learning algorithms are found to be more popular for investigating exposomics impact on CVDs and which seem to be ranking among the top performers?

Tasks encountered in the retrieved literature have been predominantly addressed within classification and regression contexts, with clustering being much less common. Overall, a broad palette of ML algorithms spanning from linear to non-linear and ensemble algorithms has been exploited. In many cases different categories of algorithms are employed and compared against each other with respect to different evaluation metrics (e.g., accuracy in classification case, or MSE in regression). A detailed reporting of algorithms used in each study can be found in Table 1.

Research endeavors conducting comparative experiments among diverse algorithms account for nearly 70% of the identified studies. Starting from the most extensive investigations, a considerable part of the literature has explored a selection of linear, non-linear and ensemble methods [35], [41], [42], [44], [45], [47], [48], [52], [55], [64], [67], [68], [69], [70], [72], [73], [74], [75].

The second most commonly encountered comparative approach involves utilizing non-linear and ensemble methods [39], [50], [53], [54], [65], [74],have exploited linear and ensemble algorithms while [76] focused on linear and non-linear algorithms. Another comparative approach involves selecting more than one algorithm but within a specific family of models. In this context, [57] have focused on the use of non-linear models while a small portion of literature (e.g. [65], [66]) have exclusively investigated ensemble models.

Proceeding with the remaining 30% of the identified literature that does not conduct performance comparison whatsoever, a part of research opted for a non-linear algorithm [46], [56], [61], [62] while another part has focused on an ensemble algorithm of their selection [40], [51], [58], [60], [71].

As it becomes obvious from Figure 4 ensemble algorithms are the most popular choice [35], [39], [40], [41], [42], [43], [44], [45], [53], [54], [55], [59], [60], [63], [64], [65], [66], [67], [68], [69], [71], [72], [73] closely followed by non-linear algorithms [35], [39], [41], [42], [45], [53], [55], [55], [57], [58], [59], [61], [62], [63], [64], [65], [66], [69], [72], [73], [74], [76], [77], [78], [79]. The most widely adopted algorithms per category are Random Forest (RF) followed by Extreme Gradient Boosting (XGBoost) from the family of ensemble methods, Artificial Neural Networks (ANN) followed by Support Vector Machines (SVM) from the family of non-linear methods and linear/logistic regression as well as Lasso and Ridge variants from the linear algorithms.

As Table 2 shows, the majority of comparative studies identify different representatives of the family of ensemble models as top-performing algorithms. Among these, the RF algorithm is the most frequently highlighted as the best performer, followed by Stacking Ensemble.

Linear methods exploited in the relevant literature are linear/logistic regression [41], [42], [45], [48], [53], [63], [69], [76], [80], Linear Mixed Effects Models [35], SVM with linear kernel [67], Ridge regression [73], Lasso [54], [64], [72], Elastic Net (EN) [68], although the latter in a meta-learning context.

From the popular family of ensemble models, quite a few works have reported the use of Random Forests (RF) [51], [59], [60], [63], [65], [67], [69], [73], [74] with the latter using Random Survival Forest (RSF) i.e. an extension of the traditional Random Forest algorithm designed for survival analysis. Next widely adopted choices are XGBoost [40], [63], [65], [67] and Stacking ensemble model [68], [70], [75] followed by Light Gradient Boosting Machine (LGBM) [63], Gradient Boosting (GB) [73], Random Ferns [65], Bayesian Additive Regression Tree (BART) [55] .

Regarding non-linear methods, one of the most widely adopted methods is Support Vector Machines [41], [47], [48], [63], [69], [70], [74], [78]. Other algorithms commonly applied are kNN [47], [67], [72], [74], [75], Decision Trees [48], [50], Naive Bayes [44], [52] and Bayesian Kernel Machine Regression [49], [80]. Additionally, the use of artificial neural networks (ANN) and/or DL has been adopted in a considerable part of literature [35], [41], [42], [43], [46], [57], [59], [62], [69], [73], [76], [77]. It is noteworthy there is a single approach in the retrieved literature [52] where a GAN is also used model the joint distribution of risk factors, simulate possible disease progressions and to generate plausible alternative lifestyle scenarios and model uncertainties in disease progression.

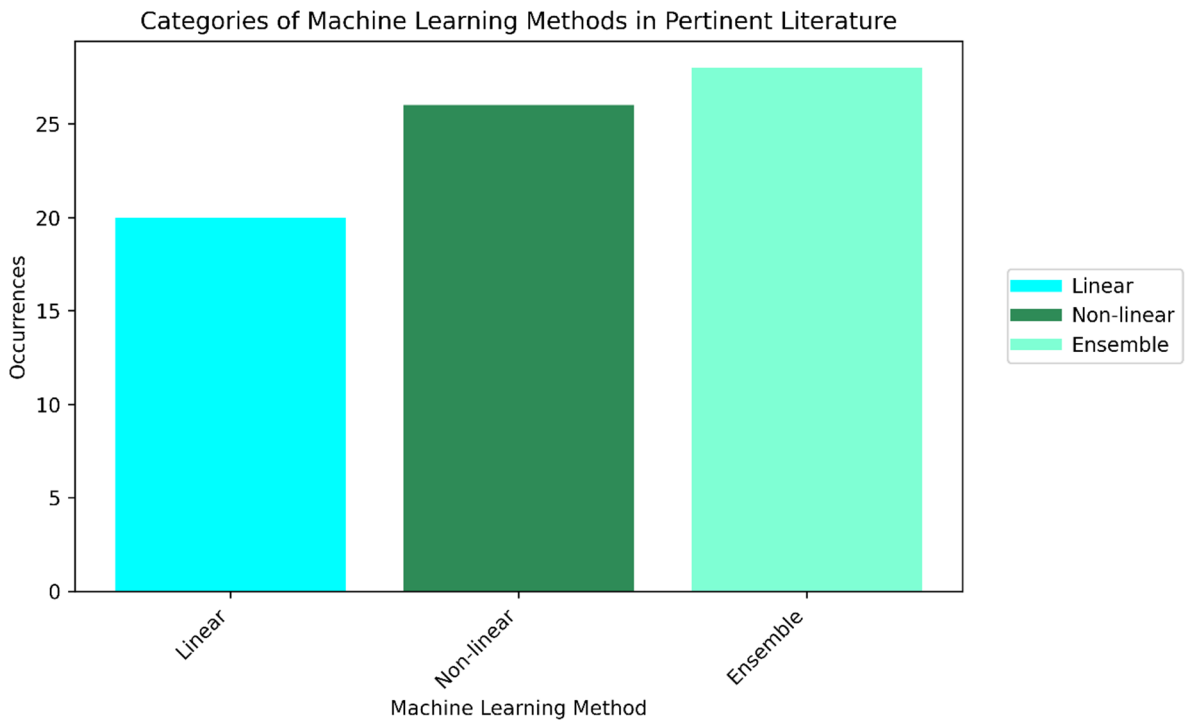

Figure 4. Popularity of categories of ML algorithms as reflected by the number of corresponding occurrences in identified literature.

RQ3. Which categories of exposomics factors have been investigated?

During the last years feature space investigations have broadened to include non-traditional CVD risk factors found in built, natural, and social environments, significantly contributing to the disease burden. The primary feature categories investigated in the retrieved literature, whether individually or in combination, are environmental exposure factors, lifestyle και socio-economic status-related factors. Most identified articles exploit exclusively environmental parameters while the remaining works consider either exclusively lifestyle factors or different combinations of these categories. For an overview see Fig.5 while for a more detailed reporting of the categories of exposomics features exploited in literature, see Table 3.

Environmental exposure is the most widely adopted category of exposome [46], [49], [53], [55], [56], [57], [58], [59], [62], [64], [67], [68], [69], [70], [73], [76], [77], [79], [81] with the bulk of relevant literature exploring a broad palette of factors, ranging from pollutants concentrations (e.g. in [40], [46], [53], [54], [56], [57], [60], [61], [64], [67], [68], [69], [73], [76], [77], [79]), meteorological parameters ([46], [53], [56], [60], [62], [64], [67], [69], [73], [76], [79], [81]) to biomarkers reflecting the extent of human exposure to heavy metals ([49], [55], [59], [74]) concentration based on human scalp hair analysis, blood or urine samples. Less often, parameters reflecting noise pollution (e.g. in [40], [54]), green spaces ([54], [73]) and traffic proximity/volume (e.g. [40], [52], [66]) are also included in the exploited feature space. Lastly, a novel approach worth mentioning consists in the exploitation of machine vision-enabled assessment of the built environment to investigate potential associations with CVD-related variables [35]. Features extracted from Google Street View (GSV) images could also generate activation maps enabling the identification of high-risk neighborhoods.

Following closely are lifestyle-related factors, mostly reflecting dietary/sleep patterns along with various health habits. In this category, the focus is predominantly on health-related habits and adopted dietary (e.g. [11], [39], [40], [44], [50], [78], [80]), physical activity or the lack of it (e.g. in [43], [48], [50], [52]) and/or sleep patterns/quality (e.g. in [40], [50], [51], [52]). The most commonly considered health habit-related parameters is smoking status (e.g. in [47], [50], [52], [75]) and/or alcohol consumption (e.g. in [50], [52], [75]). In population-focused studies, group-specific input may also be incorporated. For example, in [50], a study focusing on adolescents, total screen time has also been included as a feature in the model.
Regarding socio-economic factors, a wide range of variables are used, primarily reflecting income, education levels, economic status and/or the subject's occupation [11], [40], [42], [44], [45], [65], [72]. Less often, extra variables such as lack of private health insurance, home ownership, housing type or severe housing cost burden are also considered (e.g. in [61], [63], [66], [71], [78]).

A considerable part of literature explores combinations of all three categories i.e. environmental, lifestyle, and socio-economic factors [11], [39], [40], [43], [45], [61], [66]. Additionally, socio-economic and lifestyle-related parameters are frequently combined to investigate associations of exposomics data with CVD-related variables [44], [50], [65], [71], [78]. One last commonly encountered combination of categories consists in simultaneous investigation of socio-economic and environmental categories [35], [41], [60], [74]. Notably, socioeconomic are typically studied in combination with other categories.

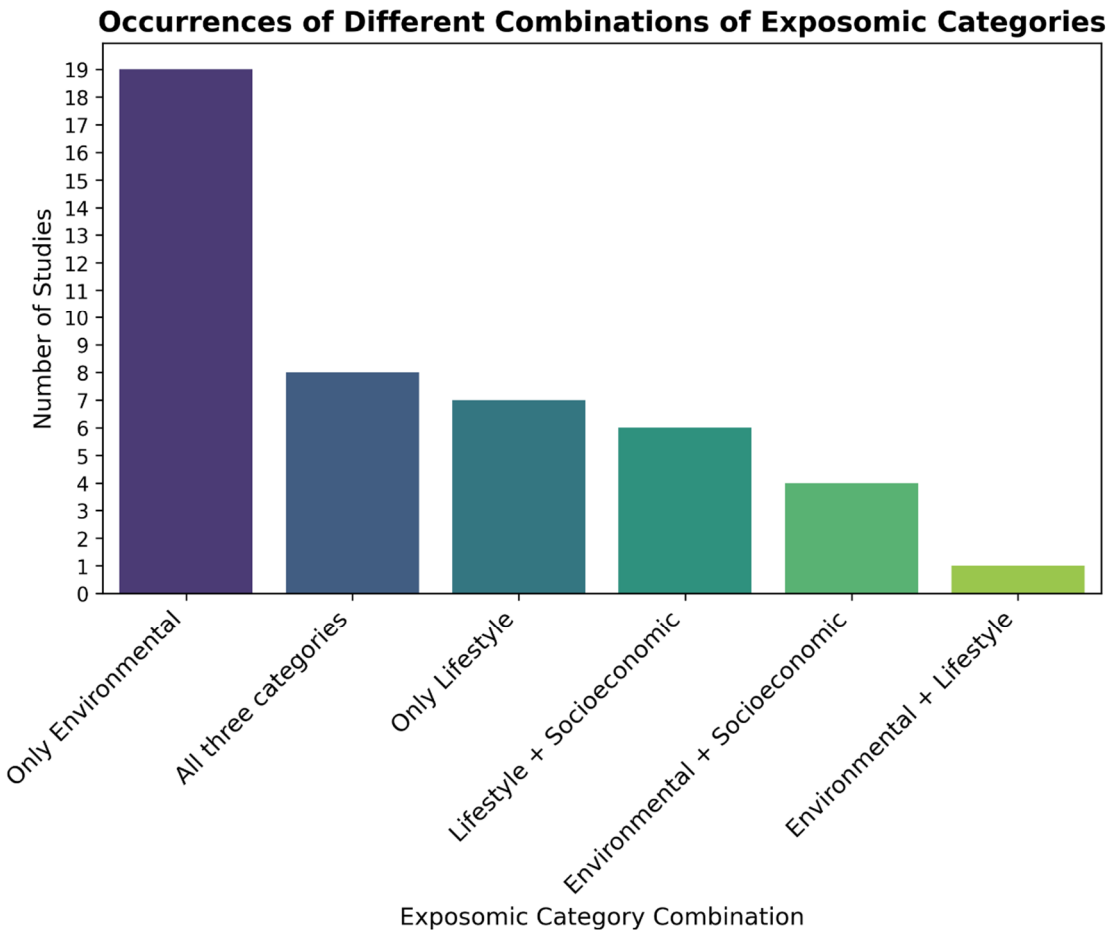

Figure 5. Occurrences of different combinations of exposomic categories used in retrieved literature.

Discussion

This scoping review aimed at presenting the state-of-the-art of ML applications on exposomics data, specifically focusing on CVDs. To this end, a substantial amount of current literature on the selected topic has been identified and analyzed with respect to diverse aspects i.e. the use case of each application and their distinguishing features, ML techniques employed and the categories of exposomics data exploited. Building on the insights from the previous sections, we will summarize findings and elaborate on the strengths, the challenges and the opportunities that ML-enabled exploitation of exposomics data holds for CVD-related variables.

As reflected by the timeline of identified publications, the field is constantly expanding, particularly from 2021 onwards, with the expansion in exploitation of ML on exposomics targeting CVDs being predominantly driven by research in the US and Asia. However, the past 4 years the EU has shown an increasing interest in the Human Exposome Project initiating multimillion funding and thus recognizing potential to enhance public health [85].

Overall, ML approaches have provided novel insights into the exposomic factors affecting CVD by overcoming the inherent limitations of traditional epidemiological models, which assume linear and independent relationships between factors and outcomes. Unlike conventional regression-based models (such as linear regression and LASSO), which can only produce accurate interpretations under the assumption of linearity, ML techniques can capture nonlinear, synergistic, and threshold effects among complex exposures while also improving accuracy (e.g. in [44], [63]). This is particularly important in exposomics where high dimensional data is involved, and multiple factors interact in intricate ways that cannot be easily disentangled using standard statistical methods. By leveraging algorithms that are more robust to multicollinearity (e.g., random forests) and capable of identifying hidden patterns in high-dimensional data, ML enables the detection of previously unrecognized interactions [55].

As presented in detail in the context of RQ1, two main categories of studies have been identified based on the application context: the first one focusing on disease management, prevention and understanding, and the second one addressing healthcare resource planning and management. Overall, the direction of research efforts falls within the scope of AI-driven health investigations aimed at effective risk stratification, public health surveillance and health policy planning and the ML assisted exploratory association analysis between novel risk factors and CVD outcomes, all of which are described to contribute to health-related sustainable development goals [86].

Regarding ML techniques, it is noteworthy that ML tasks have been primarily framed in a supervised context, leaving ample space for the exploration of under-researched unsupervised techniques. Overall, a variety of ML algorithms spanning from linear to non-linear and ensemble algorithms have been exploited. In many cases different categories of algorithms are employed and compared against each other. Ensemble algorithms seem to be the most widely adopted approach and especially Random Forest, Stacking Ensemble and XGBoost which frequently rank as the most efficient solutions in comparative experiments. In the context of unsupervised learning and data clustering, more methods could be explored imposing minimal assumptions on data, trying to identify patterns, simplify large datasets and enhance predictive modeling. A valuable addition to the toolbox of the well-established linear clustering methods could be the case of non-linear clustering algorithms and

specifically the so-called manifold learning, which aims to cluster data by identifying the intrinsic manifold upon which the data lies [87].

As presented in the context of RQ3, environmental exposure seems to be the most widely investigated aspect of the feature space followed by lifestyle-related exposure, with both of them often handled individually. In contrast, socioeconomic factors are typically studied in combination with other types of variables. Most environmental exposure studies are exploratory in nature, seeking to confirm associations of variables with health outcomes, instead of developing complex prediction models; they often aim to expand the understanding of the exposure-outcome relationships through identifying new or inadequately studied risk factors that may play an important role on disease prevalence and progression [11], [42], [65], [66]. Such efforts showcase the need for public health policies related to the environment [62].

A critical consideration in applying ML to exposomics research is the validity of inference. Many widely used ML algorithms, such as regression-based (e.g., LASSO) and tree-based (e.g., Random Forest and XGBoost) methods, are primarily designed for prediction rather than for modelling exposure-outcome associations (e.g., Bayesian Additive Regression Trees and Bayesian Kernel Machine Regression). Thus, they do not necessarily provide valid statistical inferences (e.g., they are often unable to quantify uncertainty around estimates) and/ or indicate true exposure-outcome associations. These algorithms are typically optimized with respect to prediction-oriented measures (e.g., AUC for classification, MAE for regression etc.) and then used to report associations - often relying on various feature importance metrics to identify key determinants [35], [39], [40], [41], [42], [44], [53], [55], [59], [60], [61], [62], [63], [65], [66], [68], [69], [71], [73], [74], [78], [81], [84], [88]. It is important to note that this approach raises concerns about post-selection bias, as feature selection is data-driven, and traditional statistical measures (e.g., p-values, confidence intervals) do not account for the fact that variables were selected based on observed data. For example, when conducting multiple statistical tests across many variables (e.g., in the context of linear regression tools), the likelihood of false positives increases, necessitating multiple testing adjustments (e.g., Bonferonni correction, False Discovery Rate etc.). Without proper correction, significance can be overstated resulting in potentially misleading conclusions. This is especially pertinent in exposomics, which frequently involve high-dimensional data with numerous, highly correlated exposures, and where identifying true predictors amidst correlated covariates is a significant challenge [89]. For example, in [62] the authors report the relative importance of each feature and compare the results with General Linear Model (GLM) where they consider the features significant based on p-value <0.05 without correcting for multiple comparisons.

Closely related to the above considerations is the issue of causal reasoning. When correlated variables are fed into ML models predicting CVD outcomes, their relative importance will be estimated higher when their impact on a predicted score is higher without necessarily implying causality. To enhance the reliability and generalizability of ML-driven exposomics research, a sound methodology should pay special attention to rigorous validation frameworks, including carefully chosen post-selection inference methods, such as cross-validation, permutation-based significance testing, stability selection, and, ultimately, external validation. However, the causality of the identified associations still needs to be validated independently within a causal inference framework [88] and/or through comparison with conventional association analysis [68]. When these aspects cannot be meticulously addressed, or are only partially addressed, authors should clearly acknowledge the limitations of their study. Expanding the use of causal inference methods (e.g., instrumental variable analysis, propensity score matching) within the ExWAS context, can help establish robust links between exposures and CVD outcomes, mitigating the impact of non-causal correlations and promoting the understanding of the underlying causal pathways.

Several additional limitations have been also reported across the identified literature. Firstly, the validity of the exploited ground truth cannot be ascertained, for example when using diagnostic codes (e.g., the international classification of diseases [65], [71]) corresponding to hospitalizations or death certificates [90]. Secondly, in the case of self-reported data categories such as lifestyle data [11], [39], [41], [42], [44], [45], [63], [78], [88], subjectivity and recall bias cannot be avoided. In addition, the bulk of conducted research only involved incidents that occurred in a single region and/ or the cohort was ethnically homogeneous [40], [44], [65], [71], [84]. Another important limitation stems from the lack of temporal alignment between the collection of medical data concerning CVD incidents and the collection of exposome-related data with many of the identified studies being observational in nature. Most of the time, the data sources were not comprehensive and/ or standardized [36]. Several works report limited availability of data sources [35], [44], [88], and finally, it is difficult to directly compare results from different studies since they are heterogeneous in terms of target, problem framing, study design, sample size and outcome assessment as reflected by the occasional lack of an external validation process but also by the diverse metric scores reported.

Most of these limitations are data-related, primarily arising from issues such as data heterogeneity, lack of standardization, inconsistent terminology, and incomplete, biased, or limited data sources. To address these challenges, it is essential to integrate standardized protocols for exposome data collection and harmonization across studies, reducing bias and enhancing reproducibility. Establishing a common language for exposomics is crucial for eliminating inconsistencies in terminology and classification. This would ultimately enable more effective comparison of findings across studies. It has been proposed by Schmitt et al. 2023 [34] that exposomics research would benefit significantly from a comprehensive, standardized, and interoperable data ecosystem framework, facilitating the integration of diverse longitudinal datasets across multidisciplinary [91] and international collaborations—similar to the Human Genome Project [92]. Such an ecosystem would link studies across continents, merging data from both existing and future longitudinal research. By establishing a federation of global data resources, this initiative would enhance the visibility, accessibility, and usability of exposomics data. Achieving this vision requires collaboration among multiple stakeholders, working together to form a global consortium that unifies frameworks and provides guidelines for an Exposome Study Design and a comprehensive exposome database [91]. In this context, it should be noted that the Environmental Exposure Assessment Research Infrastructure (EIRENE [93]) is the first European infrastructure dedicated to the human exposome.

The adoption of longitudinal study designs with synchronized environmental and health data collection can enhance the ability to track exposures over time and infer causality more reliably (e.g., the UK biobank utilized in [40], [84]). Additionally, integrating diverse data sources, including computerized and standardized data from wearables, can help mitigate recall bias, which is often introduced by self-reporting [11], [39], [40], [41], [42], [44], [45], [63], [65], [78], [84], [88]. Lastly, Explainable AI (XAI) techniques are valuable for enhancing model interpretability, enabling qualitative comparisons and ensuring that associations identified between exposomic factors and CVD-related outcomes are meaningful in real-world applications. It is noteworthy that many studies identified in this review implement XAI techniques in meaningful ways, attempting to interpret their models and explore potential associations (e.g., by using the model-agnostic SHapley Additive exPlanations) [11], [40], [42], [60], [63], [68], [74], [78], [81], [88].

All in all, the growing intersection of ML and CVD exposomics offers unprecedented opportunities to refine exposure assessment, improve early disease detection, and guide public health interventions. Expanding the feature space to include non-traditional CVD risk factors—spanning environmental, social, and lifestyle domains—has shown promise in enhancing the performance of conventional

models targeting CVD-related variables. Such findings may serve as a foundation for developing high-precision tools for personalized healthcare management. On a broader scale, this approach holds potential for prioritizing healthcare resource allocation, optimizing workflows, and implementing targeted prevention and intervention strategies to reduce CVD-related healthcare costs.

However, to translate these advancements into actionable insights, stronger methodological rigor and standardization efforts are needed. Even though ML algorithms effectively enhance predictive modeling, most of them are not inherently designed for causal inference, raising concerns about the validity of reported exposure-outcome associations. Addressing limitations requires rigorous validation frameworks, careful post-selection inference methods, and integration of causal inference techniques into ML pipelines. Additionally, inconsistencies in data sources, lack of standardization, and the heterogeneous nature of studies hinder comparability and reproducibility. Establishing standardized protocols and interoperable data-sharing frameworks, akin to initiatives in other healthcare domains, holds the potential to significantly advance the field.

Tables

Reference	ML task	ML algorithms	Validation	Evaluation Metric	Feature importance/selection
Chen et. al. 2024 [35]	Regression	CNN, ET, RF, LGBM, Linear Mixed Effects Model (LMEM)	10-fold cross-validation	R2, MAE, RMSE	sparse partial least squares (SPLS) regression
Hossain et al. 2024 [44]	Classification	LR, Naïve Bayes, DT, AdaBoost, RF, BAG, ensemble model combining the above as estimators	Data splitting with 80:20 ratio, 5-fold cross-validation	AUC-ROC, sensitivity, specificity, and accuracy	chi-square
Guimbaud et al. 2024 [54]	Regression	Lasso, RF, XGBoost	10-fold cross-validation	R2	SHAP
Khimani et al. 2024 [48]	Classification	LR, DTT, GB, RF, SVC	Data splitting with 70:30 ratio	accuracy, average precision, balanced accuracy, F1, neg log loss, precision, recall, and ROC-AUC	RF-based
Nissa et. al. 2024 [39]	Classification	DT, RF, GB, CatBoost, XGBoost, AdaBoost, LGBM	Data splitting with 60:40 ration + 5 and 10-fold cross-validation	accuracy, sensitivity PEC, Precision, NPV, FPR, FDR, FNR, F1, MCC	correlation-based

Sahriar et al. 2024 [47]	Classification	LR, RF, KNN, SVM, GB, XGBoost, Stacking		accuracy, recall, F1, AUC.	chi-square, recursive feature elimination (RFE)
Dogan et al. 2023 [52]	Classification, Generative modeling	RF, SVM, k-NN, naïve Bayes, XGBoost, ANN, LR + GAN	Data splitting with 50:50 ratio and 10-fold cross-validation	AUC-ROC	
Dritsas & Trigka al. 2023 [75]	Classification	NB, LR, Bagging, RotF, AdaBoostM1, RF, Voting, Stacking, MLP, kNN, GAN	10-fold cross-validation	accuracy, recall, precision and AUC	RF-based, Information Gain, Gain Ratio
Nghiem et et al. 2023 [72]	Classification	Lasso, DT, KNN, RF	Data splitting with 80:20 ratio, cross-validation	sensitivity, specificity, precision, F1 score, AUC	Gini index
Leiriao *et al*. 2023 [62]	Regression	ANN	cross-validation + external	MAPE	connection weight method
Li et. al. 2023 [74]	Classification	Adaboost, SVM, RF, DT, KNN	Data splitting 80:20	AUC, Sensitivity, Specificity, Average Precision, NPV, FPR, FNR, FDR, F1 Score, Brier score	SHAP
Lu and Qiu, 2023 [68]	Regression	Stacking ensemble (four base learners: Ridge, RF, GBDT, ANN) and a meta-learner (Elastic Net), LSTM	5-fold cross-validation + external	MAE, RMSE, MAPE, R2	SHAP
Atehortua *et al*. 2023 [40]	Classification	XGBoost	Internal + External	Sensitivity, Specificity, Precision, AUC-ROC	SHAP
Dong *et al*. 2023 [61]	Classification	CART	Internal + External		RF
Martin-Morales 2023 [63]	Classification	LR, SVM, RF, XGBoost, LGBM	5-fold cross-validation	Accuracy, Recall, Precision, F1, AUC-ROC	SHAP
Bhakta et al. 2023 [45]	Classification	DT, LR, naive Bayes (NB), voting, RF, GB, bagging, XGBoost, and AdaBoost	Data splitting with 80:20 ratio	Accuracy, Precision, Recall, and F1 Score	
Ohashi et al. 2023 [81]	Regression	LightGBM	10-fold cross-validation	RMSE, MAE	Boruta SHAP
Sun et al. 2023 [71]	Regression	XGBoost	-	-	Gini index
Wang et. al. 2023 [78]	Classification, Regression (Survival analysis)	LR, RF, SVC, MLP, Cox Survival Regression, Random survival	5-fold cross-validation	AUC, F1 score,	SHAP

		Forest (RSF), Fast Survival SVM		Precision, Recall	
Lee *et al*. 2022 [11]	Classification	LR	10-fold cross-validation+ test set	FDR	Lasso, knockoff boosted trees (KOBT), SHAP
Lee et al. 2022 [64]	Regression	XGBoost, RF, Lasso, Ridge, and EN, LSTM, stacked LSTM	10-fold cross-validation + external	MAE, RMSE	-
Marien *et al*. 2022 [73]	Regression	DT, RF, GBR, Ridge, MLP	10-fold cross-validation + external	R2, adjusted R2, ME, RMSE, BIC	DT, RF and GBR, PCA
Matheson et al. 2022 [51]	Regression (survival analysis)	RSF	Data splitting with 70:30 ratio	C-index, AUC (at different timepoints)	Minimum depth of the maximal subtree method
Salah and Srinivas 2022 [50]	Classification	DT, RF, XGBoost, DNN	Data splitting with 75:25 ratio, 10-fold cross-validation	AUC-ROC, AUC-PR, misclassification rate (MCR)	SHAP
Yao *et al*. 2022 [60]	Classification and regression	RF	-	Kappa coefficient, Precision, Recall, F1 score and AUC-ROC	SHAP
Li et al. 2022 [65]	Classification	RF, Random Ferns, and XGBoost	10-fold cross-validation + test set	AUC-ROC, AUC-PR	Boruta
Liu et al. 2022 [42]	Classification	LR, ANN, H2O Driverless AI + IF (for class imbalance handling)	cross-validation	AUC-ROC, AUC-PR, Categorical Net Reclassification Improvement (NRI), Integrated Discrimination Improvement (IDI)	BoostARoota, SHAP
Testa et al. 2022 [79]	Clustering and inferential analysis	Hierarchical Clustering, cluster selection based on dendrograms	n/a	n/a	n/a
Usmani *et al*. 2022 [77]	Regression	LSTM, ELSTM, DL	train + test	RMSE, MAE	-
Lin *et al*. 2021 [58]	Regression	LSTM		RMSE, MAPE	-
Monaco *et al.* 2021 [59]	Classification	RF, ANN, GLM	k=10,5,3 and Leave One Out	Sensitivity, Precision, F1 score	Boruta

			cross-validation (LOO-CV)		
Jalili et al. 2021 [76]	Classification	ANN, LR	cross-validation +testing (Data splitting 70:15:15)	MSE, ER, Pearson's r	
Sajid et. al. 2021 [41]	Classification	LR, ANN, SVM, RF	Data split (70:30) and 10-fold cross validation	Sensitivity, Specificity, Precision, ACC, AUC	Gini
Zhao et al. 2021 [80]	Regression	LR, BKMR (Gaussian Kernel)	-	-	PIPs
Usmani *et al*. 2021 [57]	Regression	LSTM, ELSTM, DL	Data split (70:30)	MAE, RMSE	
Qiu *et al*. 2020 [69]	Classification	ANN, SVM, LR, RF, XGBoost, LGBM	10-fold cross-validation	Accuracy, specificity, precision, F1 score, AUC, log-loss	LGBM
Hu *et al*. 2020 [66]	Regression	BART, GBM, XGBoost, RF	5-fold cross-validation	RMSE	BART, GBM, XGBoost, RF
Hu *et al*. 2020b [70]	Regression	Stacking Ensemble (base learners: LR, SVR, XGBoost, RF, GBDT), LR, SVR, XGBoost, RF, GBDT	5-fold cross-validation	MAE, RMSE, MAPE, R^2	RF-based
Chen *et al*. 2019 [67]	Classification	LR, XGBLinear, KNN, RF, XGBTree, and SVMLinear	10-fold cross-validation	AUC-ROC, sensitivity, specificity	Lasso
Domingo-Relloso et al. 2019 [49]	Classification, Regression (survival analysis)	BKMR-P, Cox Regression	-	-	Posterior inclusion probabilities (PIP)
Alaa et al. 2019 [43]	Classification	LR, RF, NN, AdaBoost, GBM (AutoPrognosis)	10-fold cross-validation	AUC-ROC	Lasso, PCA, RF
Ren et. al. 2018 [53]	Classification	RF, GB, LR	10-fold cross-validation	AUC-ROC, AUC-PRC, F1, Precision, Recall	Gini Coefficient and Relative Influence
Park 2017 [55]	Regression and classification	AENET-I, BART, BKMR, and Super Learner	5-fold cross-validation +	PRESS, MSE, MSPE AUC-ROC	custom, within Super Learner framework
Hsiao *et al*. 2016 [46]	Classification	Autoencoder and a Softmax classifier	-	-	-
Dominici et al. 2006 [56]	Clustering, Regression	K-means	n/a	n/a	n/a

Table 1. ML task addressed and key information on the adopted pipelines.

Category	Algorithm Reported as Best Per Study	Studies
Ensemble	*RF*	[41], [44], [47], [48], [52], [53], [59], [63], [65], [72], [74]
	Stacking Ensemble	[47], [68], [70], [75]
	XGBoost	[45], [50], [54]
	AdaBoost	[39]
	ET	[35]
	LGBM	[69]
Linear	*Ridge*	[64], [73]
	Cox survival regression (L1 penalty)	[78]
	EN	[64]
	LR	[67]
	AENET-I	[55]
Nom-linear	*ELSTM*	[77], [77]
	BART	[66]
	MLP	[73]
	ANN	[76]

Table 2. Best-performing algorithms per study

Reference	Exposomic categories included in feature space	Target outcome/variable
Chen et. al. 2024 [35]	Socio-economic and, built environment	CHD prevalence
Hossain et al. 2024 [44]	Socio-economic, lifestyle	CVD incidence
Guimbaud et al. 2024 [54]	Air quality (pollutant concentrations), meteorological factors, traffic noise,	Cardiometabolic risk

	traffic indicators, natural space, built environment, lifestyle, pollutant biomarkers	
Kimani et al. 2024 [48]	Lifestyle	CVD incidence
Nissa et. al. 2024 [39]	Environmental, socio-economic and lifestyle factors (including diet, physical activity and sleep hours)	Risk of heart attack
Sahriar et al. 2024 [47]	Lifestyle, socio-economic, environmental	Stroke incidence
Bhakta et al. 2023 [45]	Occupation, environment, lifestyle habits	Heart disease detection
Dogan et al. 2023 [52]	Lifestyle	Coronary Heart Disease, MI, Stroke, any type
Dritsas & Trigka, 2023 [75]	Lifestyle	CVD incidence
Leiriao *et al*. 2023 [62]	Air quality, meteorological parameters	Cardiorespiratory mortality
Li et al. 2023 [74]	Exposure to heavy metals, socioeconomic	CHD Risk
Lu and Qiu, 2023 [68]	Air quality (pollutants concentrations)	Daily counts of hospital admissions (CVDs)
Atehortua *et al*. 2023 [40]	Early-life, environmental (noise/ pollution levels), lifestyle, sociodemographics	Cardiometabolic risk estimation (coronary/ischaemic heart diseases, heart failure events, vascular dementia and cerebrovascular diseases)
Dong *et al*. 2023 [83]	Air quality (pollutants concentration), lifestyle, socioeconomic status	Premature CV mortality
Martin-Morales 2023 [63]	Dietary and non-diet-related health data	CVD mortality
Ohashi et. al. 2023 [81]	Daily weather data	CVD mortality
Sun et al. 2023 [71]	Demographic composition, risk factors, lifestyle and socioeconomic status	Total care costs per capita, inpatient care costs per capita, outpatient care costs per capita, CVD prevalence (%)
Wang et. al. 2023 [78]	Socioeconomic status, lifestyle (including dietary patterns)	CVD Mortality, IHD hospitalization

Salah and Srinivas 2022 [50]	Socioeconomic, lifestyle	CVD risk
Lee *et al*. 2022 [64]	Environmental (meteorological variables, air pollutants)	Non-accidental, cardiovascular, and respiratory mortality
Lee et al. 2022 [11]	Chemical and biological exposures, socioeconomic status, Lifestyle (food, diet and sleep)	Cardiac arrhythmia, congestive heart failure, coronary artery disease, heart attack, stroke, and a combined atherogenic-related outcome comprising angina, angioplasty, atherosclerosis, coronary artery disease, heart attack, and stroke incidents
Marien *et al*. 2022 [73]	Meteorological, air quality (pollutants concentration), vegetation index	Daily and annual CVD incidence (myocardial infarctions)
Matheson et al. 2022 [51]	Lifestyle	CVD incidence (CAD and ASCVD)
Yao *et al*. 2022 [60]	Urban multi-source spatio-temporal big data, road network, demographic, economic, meteorologic data and air pollutants	MI disease severity, MI mortality
Li et al. 2022 [65]	Socio-economic and lifestyle factors	CVD diagnoses
Liu et al 2022 [42]	Demographic and lifestyle Information (Dietary Intake, Health Behaviors)	CVD incidence (Stroke)
Testa et al. 2022 [79]	Air Pollutants and weather variables	CVD incidence (Acute cardiac or cerebrovascular events)
Sajid et. al. 2021 [41]	Non-clinical factors such as education, area of living, occupation, exposure to economic problems and the role in the family	CVD risk
Usmani *et al*. 2022 [77]	Air quality (pollutants concentration)	Cardiorespiratory mortality
Lin *et al*. 2021 [58]	Air quality, meteorological	Emergency department visits (CVD-related etiology)
Monaco *et al.* 2021 [59]	Heavy metal concentrations (extracted by means of TMA hair tests.)	CVD risk

Jalili et al. 2021 [76]	Air quality (pollutants concentration), meteorological	CVD rate (as reflected by hospital admissions)
Usmani *et al*. 2021 [57]	Air quality (pollutants concentration)	Monthly cardiorespiratory hospitalization
Zhao et al. 2021 [80]	Dietary	ASCVD risk score (PCE)
Qiu *et al*. 2020 [69]	Air quality (pollutants concentration), meteorological	Peak demand days of CVDs admissions
Hu *et al*. 2020 [66]	Air pollution, lifestyle, dietary, socio-economic	Stroke incidents (neighborhood-level)
Hu *et al*. 2020 [70]	Air quality, environment	Daily number of Hospital Admissions for CVD
Chen *et al*. 2019 [67]	Air quality (pollutants concentration), meteorological	Hospital admissions (hemorrhagic stroke)
Alaa et al 2019 [84]	Health and medical history, lifestyle and environment, physical activity, psychosocial factors, dietary and nutritional information, and sociodemographics	CVD risk
Domingo-Relloso et al. 2019 [49]	Metal concentrations in urine	CVD incidence
Ren et. al. 2018 [53]	Air quality (pollutants concentration), meteorological (humidity, temperature)	Risk of Congenital Heart Defects
Park 2017 [55]	Metal biomarkers in blood and urine	CVD mortality
Hsiao *et al*. 2016 [46]	Air quality (pollutants concentration), meteorological	CVD risk
Dominici et al. 2006 [56]	Air quality (pollutants concentration), meteorological	Daily counts of county-wide hospital admissions (for primary diagnosis of cerebrovascular, peripheral, and ischemic heart diseases, heart rhythm, heart failure, chronic obstructive pulmonary disease, and respiratory infection, and injuries as a control outcome)

Table 3. Exposome-related feature categories investigated and target outcomes addressed in pertinent literature.

References

[1] M. Di Cesare *et al.*, “The Heart of the World,” *Glob Heart*, vol. 19, no. 1, p. 11, doi: 10.5334/gh.1288.

[2] “World Heart Report 2023: Full Report,” World Heart Federation. Accessed: Mar. 28, 2025. [Online]. Available: https://world-heart-federation.org/resource/world-heart-report-2023/

[3] T. Münzel, M. R. Miller, M. Sørensen, J. Lelieveld, A. Daiber, and S. Rajagopalan, “Reduction of environmental pollutants for prevention of cardiovascular disease: it’s time to act,” *European Heart Journal*, vol. 41, no. 41, pp. 3989–3997, Nov. 2020, doi: 10.1093/eurheartj/ehaa745.

[4] A. Bonanni, M. Basile, R. A. Montone, and F. Crea, “Impact of the exposome on cardiovascular disease,” *European Heart Journal Supplements*, vol. 25, no. Supplement_B, pp. B60–B64, Apr. 2023, doi: 10.1093/eurheartjsupp/suad069.

[5] C. P. Wild, “Complementing the genome with an ‘exposome’: the outstanding challenge of environmental exposure measurement in molecular epidemiology,” *Cancer Epidemiol Biomarkers Prev*, vol. 14, no. 8, pp. 1847–1850, Aug. 2005, doi: 10.1158/1055-9965.EPI-05-0456.

[6] G. W. Miller and D. P. Jones, “The nature of nurture: refining the definition of the exposome,” *Toxicol Sci*, vol. 137, no. 1, pp. 1–2, Jan. 2014, doi: 10.1093/toxsci/kft251.

[7] R. Vermeulen, E. L. Schymanski, A.-L. Barabási, and G. W. Miller, “The exposome and health: Where chemistry meets biology,” *Science*, vol. 367, no. 6476, pp. 392–396, Jan. 2020, doi: 10.1126/science.aay3164.

[8] E. Uffelmann *et al.*, “Genome-wide association studies,” *Nat Rev Methods Primers*, vol. 1, no. 1, pp. 1–21, Aug. 2021, doi: 10.1038/s43586-021-00056-9.

[9] Y. Zheng, Z. Chen, T. Pearson, J. Zhao, H. Hu, and M. Prosperi, “Design and methodology challenges of environment-wide association studies: A systematic review,” *Environmental Research*, vol. 183, p. 109275, Apr. 2020, doi: 10.1016/j.envres.2020.109275.

[10] M. K. Chung *et al.*, “Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs),” *Exposome*, vol. 4, no. 1, p. osae001, Jan. 2024, doi: 10.1093/exposome/osae001.

[11] E. Y. Lee *et al.*, “Questionnaire-based exposome-wide association studies (ExWAS) reveal expected and novel risk factors associated with cardiovascular outcomes in the Personalized Environment and Genes Study,” *Environ Res*, vol. 212, no. Pt D, p. 113463, Sep. 2022, doi: 10.1016/j.envres.2022.113463.

[12] P. J. Landrigan *et al.*, “The Lancet Commission on pollution and health,” *The Lancet*, vol. 391, no. 10119, pp. 462–512, Feb. 2018, doi: 10.1016/S0140-6736(17)32345-0.

[13] B.-J. Lee, B. Kim, and K. Lee, “Air Pollution Exposure and Cardiovascular Disease,” *Toxicol Res*, vol. 30, no. 2, pp. 71–75, Jun. 2014, doi: 10.5487/TR.2014.30.2.071.

[14] T. Münzel *et al.*, “Transportation Noise Pollution and Cardiovascular Health,” *Circulation Research*, vol. 134, no. 9, pp. 1113–1135, Apr. 2024, doi: 10.1161/CIRCRESAHA.123.323584.

[15] "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide." Accessed: Jul. 01, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240034228
[16] "Air quality in Europe 2022," European Environment Agency. Accessed: Jul. 16, 2024. [Online]. Available: https://www.eea.europa.eu/publications/air-quality-in-europe-2022/air-quality-in-europe-2022
[17] R. Fuller *et al.*, "Pollution and health: a progress update," *The Lancet Planetary Health*, vol. 6, no. 6, pp. e535–e547, Jun. 2022, doi: 10.1016/S2542-5196(22)00090-0.
[18] S. Sang, C. Chu, T. Zhang, H. Chen, and X. Yang, "The global burden of disease attributable to ambient fine particulate matter in 204 countries and territories, 1990-2019: A systematic analysis of the Global Burden of Disease Study 2019," *Ecotoxicol Environ Saf*, vol. 238, p. 113588, Jun. 2022, doi: 10.1016/j.ecoenv.2022.113588.
[19] T. Münzel, M. Sørensen, and A. Daiber, "Transportation noise pollution and cardiovascular disease," *Nat Rev Cardiol*, vol. 18, no. 9, pp. 619–636, Sep. 2021, doi: 10.1038/s41569-021-00532-5.
[20] S. Crnko, B. C. Du Pré, J. P. G. Sluijter, and L. W. Van Laake, "Circadian rhythms and the molecular clock in cardiovascular biology and disease," *Nat Rev Cardiol*, vol. 16, no. 7, pp. 437–447, Jul. 2019, doi: 10.1038/s41569-019-0167-4.
[21] L. E. Laugsand, L. J. Vatten, C. Platou, and I. Janszky, "Insomnia and the risk of acute myocardial infarction: a population study," *Circulation*, vol. 124, no. 19, pp. 2073–2081, Nov. 2011, doi: 10.1161/CIRCULATIONAHA.111.025858.
[22] S. Sun *et al.*, "Outdoor light at night and risk of coronary heart disease among older adults: a prospective cohort study," *Eur Heart J*, vol. 42, no. 8, pp. 822–830, Feb. 2021, doi: 10.1093/eurheartj/ehaa846.
[23] S. Bará, F. Falchi, R. C. Lima, and M. Pawley, "Keeping light pollution at bay: A red-lines, target values, top-down approach," *Environmental Challenges*, vol. 5, p. 100212, Dec. 2021, doi: 10.1016/j.envc.2021.100212.
[24] X.-X. Liu *et al.*, "Green space and cardiovascular disease: A systematic review with meta-analysis," *Environmental Pollution*, vol. 301, p. 118990, May 2022, doi: 10.1016/j.envpol.2022.118990.
[25] T. Münzel, O. Hahad, A. Daiber, and P. J. Landrigan, "Soil and water pollution and human health: what should cardiologists worry about?," *Cardiovasc Res*, vol. 119, no. 2, pp. 440–449, Jun. 2022, doi: 10.1093/cvr/cvac082.
[26] C. Méjean *et al.*, "The contribution of diet and lifestyle to socioeconomic inequalities in cardiovascular morbidity and mortality," *Int J Cardiol*, vol. 168, no. 6, pp. 5190–5195, Oct. 2013, doi: 10.1016/j.ijcard.2013.07.188.
[27] T. Psaltopoulou, G. Hatzis, N. Papageorgiou, E. Androulakis, A. Briasoulis, and D. Tousoulis, "Socioeconomic status and risk factors for cardiovascular disease: Impact of dietary mediators," *Hellenic Journal of Cardiology*, vol. 58, no. 1, pp. 32–42, Jan. 2017, doi: 10.1016/j.hjc.2017.01.022.
[28] M. Davari, M. R. Maracy, and E. Khorasani, "Socioeconomic status, cardiac risk factors, and cardiovascular disease: A novel approach to determination of this association," *ARYA Atheroscler*, vol. 15, no. 6, pp. 260–266, Nov. 2019, doi: 10.22122/arya.v15i6.1595.
[29] F. Tsodikov *et al.*, "The effect of lifestyle intervention on cardiometabolic risk factors in mental health rehabilitation hostel residents at-risk: a cluster-randomized controlled 15-month trial," *Int J Obes*, vol. 46, no. 5, pp. 926–934, May 2022, doi: 10.1038/s41366-022-01063-w.

[30] J. M. Rippe, “Lifestyle Strategies for Risk Factor Reduction, Prevention, and Treatment of Cardiovascular Disease,” *American Journal of Lifestyle Medicine*, vol. 13, no. 2, p. 204, Apr. 2019, doi: 10.1177/1559827618812395.
[31] L. Maitre *et al.*, “State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event,” *Environment International*, vol. 168, p. 107422, Oct. 2022, doi: 10.1016/j.envint.2022.107422.
[32] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, “Using machine learning approaches for multi-omics data analysis: A review,” *Biotechnol Adv*, vol. 49, p. 107739, 2021, doi: 10.1016/j.biotechadv.2021.107739.
[33] M. Picard, M.-P. Scott-Boyer, A. Bodein, O. Périn, and A. Droit, “Integration strategies of multi-omics data for machine learning analysis,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 3735–3746, Jan. 2021, doi: 10.1016/j.csbj.2021.06.030.
[34] C. P. Schmitt *et al.*, “A roadmap to advance exposomics through federation of data,” *Exposome*, vol. 3, no. 1, p. osad010, Jan. 2023, doi: 10.1093/exposome/osad010.
[35] Z. Chen, J.-E. Dazard, Y. Khalifa, I. Motairek, S. Al-Kindi, and S. Rajagopalan, “Artificial intelligence–based assessment of built environment from Google Street View and coronary artery disease prevalence,” *European Heart Journal*, vol. 45, no. 17, pp. 1540–1549, May 2024, doi: 10.1093/eurheartj/ehae158.
[36] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, “Social Determinants in Maguimchine Learning Cardiovascular Disease Prediction Models: A Systematic Review,” *Am J Prev Med*, vol. 61, no. 4, pp. 596–605, Oct. 2021, doi: 10.1016/j.amepre.2021.04.016.
[37] “Home - The European Human Exposome Network (EHEN).” Accessed: Jul. 17, 2024. [Online]. Available: https://www.humanexposome.eu/
[38] A. C. Tricco *et al.*, “PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation,” *Ann Intern Med*, vol. 169, no. 7, pp. 467–473, Oct. 2018, doi: 10.7326/M18-0850.
[39] N. Nissa, S. Jamwal, and M. Neshat, “A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques,” *Computation*, vol. 12, no. 1, Art. no. 1, Jan. 2024, doi: 10.3390/computation12010015.
[40] A. Atehortúa *et al.*, “Cardiometabolic risk estimation using exposome data and machine learning,” *International Journal of Medical Informatics*, vol. 179, p. 105209, Nov. 2023, doi: 10.1016/j.ijmedinf.2023.105209.
[41] M. R. Sajid *et al.*, “Nonclinical Features in Predictive Modeling of Cardiovascular Diseases: A Machine Learning Approach,” *Interdiscip Sci Comput Life Sci*, vol. 13, no. 2, pp. 201–211, Jun. 2021, doi: 10.1007/s12539-021-00423-w.
[42] J. Liu, E. L. Chou, K. K. Lau, P. Y. M. Woo, J. Li, and K. H. K. Chan, “Machine learning algorithms identify demographics, dietary features, and blood biomarkers associated with stroke records,” *J Neurol Sci*, vol. 440, p. 120335, Sep. 2022, doi: 10.1016/j.jns.2022.120335.
[43] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, “Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants,” *PLoS One*, vol. 14, no. 5, p. e0213653, 2019, doi: 10.1371/journal.pone.0213653.
[44] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, R. Hossain, and K. Hossain, “Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023,” *BMC Cardiovascular Disorders*, vol. 24, no. 1, p. 214, Apr. 2024, doi: 10.1186/s12872-024-03883-2.

[45] S. S. Bhakta, B. Sadhukhan, and N. Das, “Enhancing Early Detection of Cardiovascular Diseases using Machine Learning Techniques: A Comparative Study,” in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Sep. 2023, pp. 875–880. doi: 10.1109/ICIMIA60377.2023.10426035.
[46] H. C. W. Hsiao, S. H. F. Chen, and J. J. P. Tsai, “Deep Learning for Risk Analysis of Specific Cardiovascular Diseases Using Environmental Data and Outpatient Records,” in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, Jul. 2016, pp. 369–372. doi: 10.1109/BIBE.2016.75.
[47] S. Sahriar *et al.*, “Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms,” *Heliyon*, vol. 10, no. 5, p. e27411, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27411.
[48] A. Khimani, A. Hornback, N. Jain, P. Avula, A. Jaishankar, and M. D. Wang, “Predicting Cardiovascular Disease Risk in Tobacco Users Using Machine Learning Algorithms,” in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2024, pp. 1–5. doi: 10.1109/EMBC53108.2024.10782885.
[49] A. Domingo-Relloso *et al.*, “The association of urine metals and metal mixtures with cardiovascular incidence in an adult population from Spain: the Hortega Follow-Up Study,” *Int J Epidemiol*, vol. 48, no. 6, pp. 1839–1849, Dec. 2019, doi: 10.1093/ije/dyz061.
[50] H. Salah and S. Srinivas, “Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents,” *Sci Rep*, vol. 12, no. 1, p. 21905, Dec. 2022, doi: 10.1038/s41598-022-25933-5.
[51] M. B. Matheson, Y. Kato, S. Baba, C. Cox, J. A. C. Lima, and B. Ambale-Venkatesh, “Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort,” *Circulation Reports*, vol. 4, no. 12, pp. 595–603, 2022, doi: 10.1253/circrep.CR-22-0101.
[52] A. Dogan, Y. Li, C. Peter Odo, K. Sonawane, Y. Lin, and C. Liu, “A utility-based machine learning-driven personalized lifestyle recommendation for cardiovascular disease prevention,” *J Biomed Inform*, vol. 141, p. 104342, Dec. 2023, doi: 10.1016/j.jbi.2023.104342.
[53] Z. Ren *et al.*, “Maternal exposure to ambient PM10 during pregnancy increases the risk of congenital heart defects: Evidence from machine learning models,” *Sci Total Environ*, vol. 630, pp. 1–10, Jul. 2018, doi: 10.1016/j.scitotenv.2018.02.181.
[54] .-B. Guimbaud *et al.*, “Machine learning-based health environmental-clinical risk scores in European children,” *Commun Med*, vol. 4, no. 1, pp. 1–14, May 2024, doi: 10.1038/s43856-024-00513-y.
[55] S. K. Park, Z. Zhao, and B. Mukherjee, “Construction of environmental risk score beyond standard linear models using machine learning methods: application to metal mixtures, oxidative stress and cardiovascular disease in NHANES,” *Environmental Health*, vol. 16, no. 1, p. 102, Sep. 2017, doi: 10.1186/s12940-017-0310-9.
[56] F. Dominici *et al.*, “Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases,” *JAMA*, vol. 295, no. 10, pp. 1127–1134, Mar. 2006, doi: 10.1001/jama.295.10.1127.
[57] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, M. Marjani, R. Shaharudin, and M. T. Latif, “Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques,” *Environ Sci Pollut Res*, vol. 28, no. 40, pp. 56759–56771, Oct. 2021, doi: 10.1007/s11356-021-14305-7.

[58] Y.-C. Lin, C.-H. Tsai, H.-T. Hsu, and C.-H. Lin, “Using Machine Learning to Analyze and Predict the Relations Between Cardiovascular Disease Incidence, Extreme Temperature and Air Pollution,” in *2021 IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, Feb. 2021, pp. 234–237. doi: 10.1109/ECBIOS51820.2021.9510479.
[59] A. Monaco *et al.*, “Random Forests Highlight the Combined Effect of Environmental Heavy Metals Exposure and Genetic Damages for Cardiovascular Diseases,” *Applied Sciences*, vol. 11, no. 18, Art. no. 18, Jan. 2021, doi: 10.3390/app11188405.
[60] Y. Yao *et al.*, “Assessing myocardial infarction severity from the urban environment perspective in Wuhan, China,” *Journal of Environmental Management*, vol. 317, p. 115438, Sep. 2022, doi: 10.1016/j.jenvman.2022.115438.
[61] W. Dong *et al.*, “Risk factors and geographic disparities in premature cardiovascular mortality in US counties: a machine learning approach,” *Sci Rep*, vol. 13, no. 1, p. 2978, Feb. 2023, doi: 10.1038/s41598-023-30188-9.
[62] L. Leirião, M. de Oliveira, T. Martins, and S. Miraglia, “A Multi-Pollutant and Meteorological Analysis of Cardiorespiratory Mortality among the Elderly in São Paulo, Brazil-An Artificial Neural Networks Approach,” *Int J Environ Res Public Health*, vol. 20, no. 8, p. 5458, Apr. 2023, doi: 10.3390/ijerph20085458.
[63] A. Martin-Morales, M. Yamamoto, M. Inoue, T. Vu, R. Dawadi, and M. Araki, “Predicting Cardiovascular Disease Mortality: Leveraging Machine Learning for Comprehensive Assessment of Health and Nutrition Variables,” *Nutrients*, vol. 15, no. 18, p. 3937, Sep. 2023, doi: 10.3390/nu15183937.
[64] W. Lee, Y.-H. Lim, E. Ha, Y. Kim, and W. K. Lee, “Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches,” *Environ Sci Pollut Res*, vol. 29, no. 58, pp. 88318–88329, Dec. 2022, doi: 10.1007/s11356-022-21768-9.
[65] J.-X. Li *et al.*, “Machine learning identifies prominent factors associated with cardiovascular disease: findings from two million adults in the Kashgar Prospective Cohort Study (KPCS),” *Glob Health Res Policy*, vol. 7, no. 1, p. 48, Dec. 2022, doi: 10.1186/s41256-022-00282-y.
[66] L. Hu, B. Liu, J. Ji, and Y. Li, “Tree-Based Machine Learning to Identify and Understand Major Determinants for Stroke at the Neighborhood Level,” *J Am Heart Assoc*, vol. 9, no. 22, p. e016745, Nov. 2020, doi: 10.1161/JAHA.120.016745.
[67] J. Chen *et al.*, “Machine Learning-Based Forecast of Hemorrhagic Stroke Healthcare Service Demand considering Air Pollution,” *Journal of Healthcare Engineering*, vol. 2019, no. 1, p. 7463242, 2019, doi: 10.1155/2019/7463242.
[68] X. Lu and H. Qiu, “Explainable prediction of daily hospitalizations for cerebrovascular disease using stacked ensemble learning,” *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 59, Apr. 2023, doi: 10.1186/s12911-023-02159-7.
[69] H. Qiu, L. Luo, Z. Su, L. Zhou, L. Wang, and Y. Chen, “Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 83, May 2020, doi: 10.1186/s12911-020-1101-8.
[70] Z. Hu, H. Qiu, Z. Su, M. Shen, and Z. Chen, “A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases,” *IEEE Access*, vol. 8, pp. 138719–138729, 2020, doi: 10.1109/ACCESS.2020.3012143.

[71] F. Sun *et al.*, “Social Determinants, Cardiovascular Disease, and Health Care Cost: A Nationwide Study in the United States Using Machine Learning,” *Journal of the American Heart Association*, vol. 12, no. 5, p. e027919, Mar. 2023, doi: 10.1161/JAHA.122.027919.
[72] N. Nghiem, J. Atkinson, B. P. Nguyen, A. Tran-Duy, and N. Wilson, “Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets,” *Health Economics Review*, vol. 13, no. 1, p. 9, Feb. 2023, doi: 10.1186/s13561-023-00422-1.
[73] L. Marien *et al.*, “Machine learning models to predict myocardial infarctions from past climatic and environmental conditions,” *Natural Hazards and Earth System Sciences*, vol. 22, no. 9, pp. 3015–3039, Sep. 2022, doi: 10.5194/nhess-22-3015-2022.
[74] X. Li *et al.*, “Development of an interpretable machine learning model associated with heavy metals’ exposure to identify coronary heart disease among US adults via SHAP: Findings of the US NHANES from 2003 to 2018,” *Chemosphere*, vol. 311, p. 137039, Jan. 2023, doi: 10.1016/j.chemosphere.2022.137039.
[75] E. Dritsas and M. Trigka, “Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction,” *Sensors*, vol. 23, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/s23031161.
[76] M. Jalili *et al.*, “Ambient air pollution and cardiovascular disease rate an ANN modeling: Yazd-Central of Iran,” *Sci Rep*, vol. 11, no. 1, p. 16937, Aug. 2021, doi: 10.1038/s41598-021-94925-8.
[77] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, M. Marjani, R. B. Shaharudin, and M. T. Latif, “Artificial intelligence techniques for predicting cardiorespiratory mortality caused by air pollution,” *Int. J. Environ. Sci. Technol.*, vol. 20, no. 3, pp. 2623–2634, Mar. 2023, doi: 10.1007/s13762-022-04149-0.
[78] H. Wang *et al.*, “Using machine learning to predict cardiovascular risk using self-reported questionnaires: Findings from the 45 and Up Study,” *Int J Cardiol*, vol. 386, pp. 149–156, Sep. 2023, doi: 10.1016/j.ijcard.2023.05.030.
[79] A. Testa *et al.*, “Cluster analysis of weather and pollution features and its role in predicting acute cardiac or cerebrovascular events,” *Minerva Med*, vol. 113, no. 5, pp. 825–832, Oct. 2022, doi: 10.23736/S0026-4806.22.08036-3.
[80] Y. Zhao, E. N. Naumova, J. F. Bobb, B. Claus Henn, and G. M. Singh, “Joint Associations of Multiple Dietary Components With Cardiovascular Disease Risk: A Machine-Learning Approach,” *American Journal of Epidemiology*, vol. 190, no. 7, pp. 1353–1365, Jul. 2021, doi: 10.1093/aje/kwab004.
[81] Y. Ohashi, T. Ihara, K. Oka, Y. Takane, and Y. Kikegawa, “Machine learning analysis and risk prediction of weather-sensitive mortality related to cardiovascular disease during summer in Tokyo, Japan,” *Sci Rep*, vol. 13, no. 1, p. 17020, Oct. 2023, doi: 10.1038/s41598-023-44181-9.
[82] W. Huang *et al.*, “Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction,” *Sci Rep*, vol. 12, no. 1, p. 1033, Jan. 2022, doi: 10.1038/s41598-021-04649-y.
[83] W. Dong *et al.*, “Risk factors and geographic disparities in premature cardiovascular mortality in US counties: a machine learning approach,” *Sci Rep*, vol. 13, no. 1, p. 2978, Feb. 2023, doi: 10.1038/s41598-023-30188-9.
[84] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, “Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants,” *PLoS One*, vol. 14, no. 5, p. e0213653, 2019, doi: 10.1371/journal.pone.0213653.

[85] Y. Fayet, T. Bonnin, S. Canali, and E. Giroux, “Putting the exposome into practice: an analysis of the promises, methods and outcomes of the European Human Exposome Network,” *Social Science & Medicine*, p. 117056, Jun. 2024, doi: 10.1016/j.socscimed.2024.117056.
[86] N. Schwalbe and B. Wahl, “Artificial intelligence and the future of global health,” *The Lancet*, vol. 395, no. 10236, pp. 1579–1586, May 2020, doi: 10.1016/S0140-6736(20)30226-9.
[87] M. Meilă and H. Zhang, “Manifold Learning: What, How, and Why,” *Annual Review of Statistics and Its Application*, vol. 11, no. Volume 11, 2024, pp. 393–417, Apr. 2024, doi: 10.1146/annurev-statistics-040522-115238.
[88] J.-B. Guimbaud *et al.*, “Machine learning-based health environmental-clinical risk scores in European children,” *Commun Med*, vol. 4, no. 1, pp. 1–14, May 2024, doi: 10.1038/s43856-024-00513-y.
[89] L. Agier *et al.*, “A Systematic Comparison of Linear Regression–Based Statistical Methods to Assess Exposome-Health Associations,” *Environmental Health Perspectives*, vol. 124, no. 12, pp. 1848–1856, Dec. 2016, doi: 10.1289/EHP172.
[90] R. A. Hoffman, J. Venugopalan, L. Qu, H. Wu, and M. D. Wang, “Improving Validity of Cause of Death on Death Certificates,” *ACM BCB*, vol. 2018, pp. 178–183, Aug. 2018, doi: 10.1145/3233547.3233581.
[91] “Defining the Scope of Exposome Studies and Research Needs from a Multidisciplinary Perspective | Environmental Science & Technology Letters.” Accessed: Jul. 15, 2024. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.estlett.1c00648
[92] R. A. Gibbs, “The Human Genome Project changed everything,” *Nat Rev Genet*, vol. 21, no. 10, pp. 575–576, Oct. 2020, doi: 10.1038/s41576-020-0275-3.
[93] M. University, “Vision and Mission,” EIRENE. Accessed: Apr. 01, 2025. [Online]. Available: https://eirene.eu/

Funding

No funding was received for conducting this study.

Ethics declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors

Abbreviations

AdaBoost, Adaptive Boosting

AENET-I, Adaptive Elastic-Net with main effects and pairwise interactions

AI, Artificial Intelligence

ANN, Artificial Neural Network

APS, Average Precision Score

ASCVD, Atherosclerotic Cardiovascular Disease

AUC-PR, Area Under the Precision Recall Curve

AUC-ROC, Area Under the Receiver Operating Characteristic Curve

AUC, Area Under the Curve

BAG, Bagging (regressor or classifier based on context)

BART, Bayesian additive regression tree

BKMR, Bayesian Kernel Machine Regression

BMI, Body Mass Index

CAD, Coronary Artery Disease

CART, Classification And Regression Tree

CatBoost, Categorical Boosting

CHD, Coronary Heart Disease

CNN, Convolutional Neural Network

CVD, Cardio-Vascular Disease

GB, Gradient Boosting

DL, Deep Learning

DT, Decision Tree

ELSTM, Enhanced Long Short-Term Memory Model

EN, Elastic Net

ERS, Environmental Risk Score

ExWAS, Exposome-Wide Association Study

FA, Factor Analysis

FDR, False Discovery Rate

FNR, False Negative Rate

FPR, False Positive Rate

GAN, Generative Adversarial Network

GGT, Gamma-Glutamyl Transferase

GSV, Google Street View

IDI, Integrated Discrimination Improvement

IF, Isolation Forest

KNN, k-nearest neighbors

KOBT, Knockoff Boosted Trees

LASSO, Least Absolute Shrinkage and Selection Operator

LDL, Low-Density Lipoproteins

LGBM, Light Gradient Boosting Machine

LMEM, Linear Mixed Effects Model

LOO-CV, Leave-One-Out Cross-Validation

LR, Logistic Regression

LSTM, Long Short-Term Memory Model

MAE, Mean Absolute Error

MAPE, Mean Absolute Percentage Error

MCC, Matthew's Correlation Coefficient

MI, Myocardial Infarction

ML, Machine Learning

MLP, Multi-Layer Perceptron

MSE, Mean-Squared Error

MSPE, Mean-Squared Prediction Error

NB, Naïve Bayes

NPV, Negative Predictive Value

NRI, Categorical Net Reclassification Improvement

PCA, Principal Component Analysis

PCE, Pooled Cohort Equations

PRESS

RF, Random Forest

RMSE, Root Mean Squared Error

RSF, Random Survival Forest

SHAP, SHapley Additive exPlanations

SVC, Support Vector Classification

SVM, Support Vector Machines

XAI, Explainable Artificial Intelligence

XGBoost, Extreme Gradient Boosting