Exposomics and Cardiovascular Diseases:

A Scoping Review of Machine Learning Approaches

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

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

katerina.argyri@iccs.gr, ioannis.gallos@iccs.gr, a.amditis@iccs.gr, dimitra.dionysiou@iccs.gr

ABSTRACT

Cardiovascular disease has been established as the world's number one killer**,** causing over 20 million deaths per year. This fact, along with the growing awareness of the impact of exposomic risk factors on cardiovascular diseases, has led the scientific community to leverage machine learning strategies as a complementary approach to traditional statistical epidemiological studies that are challenged by the highly heterogeneous and dynamic nature of exposomics data. The principal objective served by this work is to identify key pertinent 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, we aimed at identifying common limitations and meaningful directives to be addressed in the future. Overall, this work shows that, despite the fact that machine learning on exposomics data is under-researched compared to its application on other members of the -omics family, it is increasingly adopted to investigate different aspects of cardiovascular diseases.

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

Statements and Declarations: *The authors declare that they have no conflicts of interest regarding the publication of this manuscript*.

Introduction

Since the mid-20th century, cardiovascular diseases (CVDs) have emerged as the leading cause of death globally. Focusing on Europe, CVDs have been reported to account for 3.9 million deaths annually and over 1.8 million deaths within the European Union (EU) [1]. In addition to this significant epidemiological burden, CVDs are estimated to impose a financial cost of 210 billion euros per year on the EU economy [1]. They represent a large group of diseases attributed to a complex interplay between intrinsic risk factors, such as genetic predisposition, biological sex, 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 the European Environment Agency as significant contributors to the high burden of CVD. It is estimated that over 18% of CVD-related deaths in Europe 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 problem at a global scale highlighting the need for new approaches. Unlike one's 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 risks factors are inherently preventable leads to the actionable conclusion that reducing them is a key-step to alleviating the burden of cardiovascular disease in Europe. 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]. According to them, 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, rather than solely exposure-focused approaches, aiming 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 by means such as active personal measures, behavioral strategies, novel policies, urban landscape reforms etc. in an effort 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 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].

There are different types of environmental risk factors 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 come from particulate matter which is well known to be linked with CVDs [13] while the water pollution hazards stem from unsafe sources. Soil pollution's health impacts can be attributed mainly to heavy metals, deforestation, over-fertilization, and pesticides, with nano and micro plastics 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's a close link between water and soil pollution, as polluted soils can contaminate surface and groundwater. Heavy metals and metalloids are particularly concerning for their contribution to cardiovascular issues through oxidative stress and inflammation [14]. Chemical pollutants and 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 have 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 cardiovascular 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 cardiovascular (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].[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. In order 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 meta analyses 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. There is growing evidence that the CVD is characterized by socioeconomic inequalities [26], [27], [28]. These inequalities that are most frequently assessed in terms of income, occupation and education seem to be closely related with dietary habits and harmful lifestyle choices such as smoking and alcohol consumption [26]. The effect of lifestyle intervention aiming at nutrition and physical activity has shown to benefit cardiometabolic risk factors such as Body Mass Index (BMI), triglycerides and Low-Density Lipoproteins (LDL) of 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 also mitigation of harmful habits such as tobacco use and alcohol consumption.

While traditional epidemiology relying on well-established study designs is a valuable tool to investigate the relationship between environmental exposures and health outcomes, new challenges have been posed by the heterogeneous and dynamic nature of exposomics data. The exposome concept aims at considering 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 pose challenges such as increased complexity, high dimensionality, high correlation 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 (i.e. the term 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 technologies to accelerate research in the field [34]. 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, and to the best of the authors' knowledge, the only review paper on the application of machine learning techniques on exposomics data for CVD-related investigations exclusively includes social determinants of health [36]. The principal objective served by this work is to address this literature gap, by identifying key studies and providing an overview of the breadth of research in the field of machine learning applications on exposomics data with a focus on cardiovascular diseases. Common limitations have also been identified and meaningful directives to be addressed in the future have been suggested.

The rest of the article is structured as follows: we start with the Methods section reporting the adopted literature search strategy and outlining the three research questions with respect to which the presented analysis has been conducted and then we proceed with the Results section providing the reader with an overview of the relevant literature in the context of the three main axes of our work. Finally and as a result of this review, we summarize findings, discuss remarkable insights, and also identify and address the limitations of this study.

Methods

Search strategy

An extensive search of PubMed, IEEE Xplore and ACM Digital Library has been conducted in order to thoroughly identify articles on machine learning applications exploring potential associations of exposomics factors with CVD-related outcomes and published in English. Relevant key terms have been extracted in line with the main aspects of the exposome recognized 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, searched in the titles and abstracts of published studies. Specifically, the body of work reported herein has been obtained with the following keywords:

- 'environment' AND 'machine learning' AND 'cardiovascular diseases'
- 'socio-economic' AND 'machine learning' AND 'cardiovascular diseases'
- 'lifestyle' AND 'machine learning' AND 'cardiovascular diseases'

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) statement [38]. No temporal bounds have been imposed on the publication dates.

Study Exclusion Criteria

The following exclusion criteria have been applied in order to ensure that this review maintains focus and relevance to the desired scope:

- Non-English 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,
- Review articles,
- Duplicate articles.

Research questions

The analysis to be presented revolves around the following research questions.

RQ1. Which are the main identified categories of studies with respect to their objective?

RQ2. Which machine learning algorithms have been applied 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 aspect investigated in the context of the first research question (RQ1) focuses on classifying the selected studies into two main categories based on the selected target variables combined with the context of potential 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 pertinent 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 pertinent literature is provided, highlighting the temporal distribution of publications, the geographical distribution of utilized datasets, and the specific machine learning tasks addressed.

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

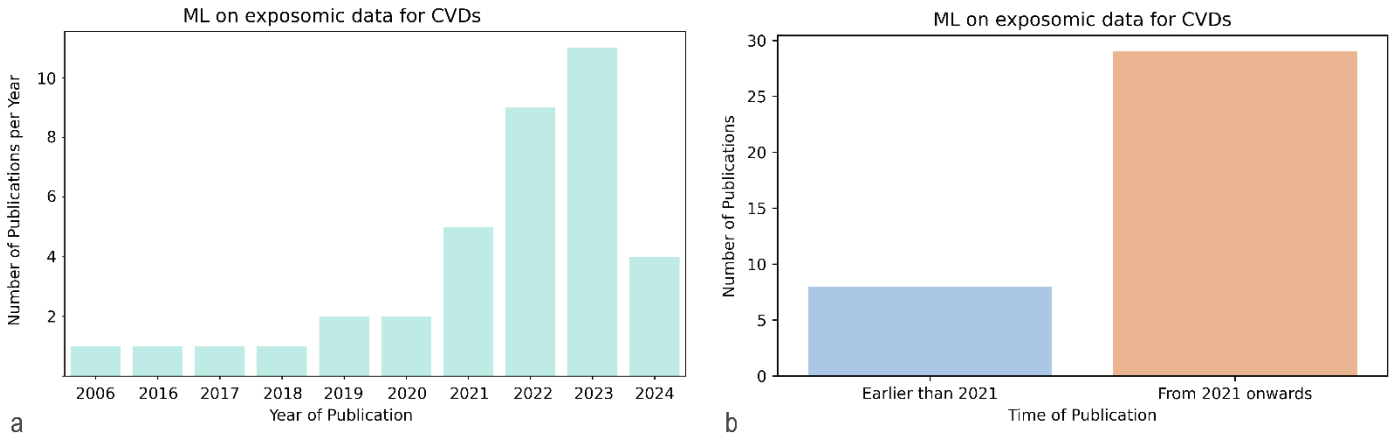

Figure 1. Time evolution of publications leveraging ML techniques to exploit exposomic datasets for CVD-related investigations a. Year of publication (x-axis) and number of identified studies per year (y-axis), b. Time of publication ('earlier than 2021' and 'from 2021 onwards')

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

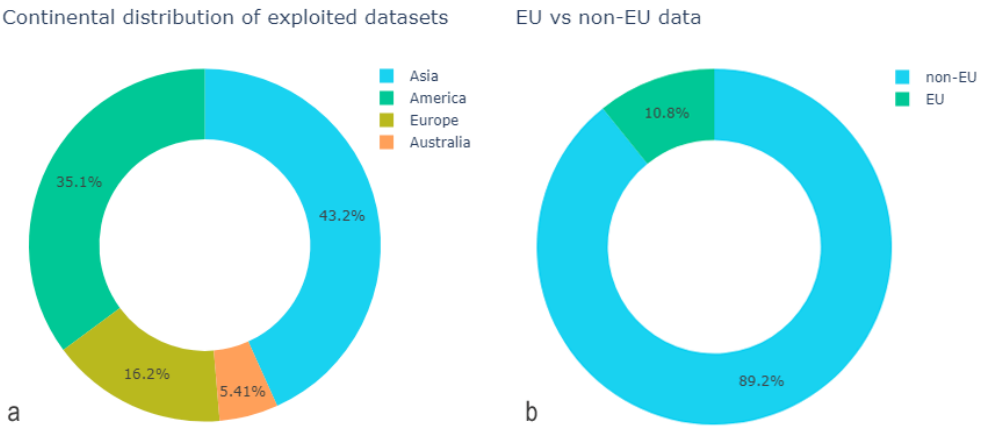

Figure 2. a. Continental distribution of exploited datasets, b. Data of EU vs non-EU origin.

Lastly, regarding the framing of the ML problems addressed in the identified literature, 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).

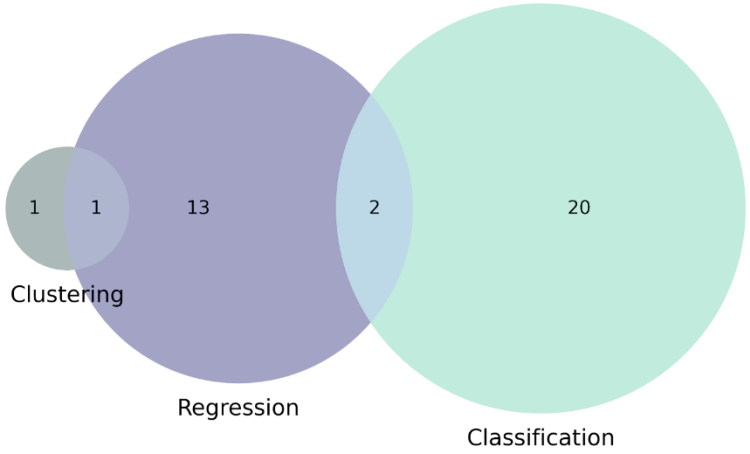

Figure 3. A Venn diagram of ML tasks categories encountered in literature focusing on exploring CVDs

Research questions

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

RQ1. Which are the main identified categories of studies with respect to their objective?

Two main categories of studies have been identified based on the selected target variable combined with the application context of the system under development. The first category aims at forecasting crucial cardiovascular outcomes or associated risk level and often focuses on identifying key determinants of the target outcomes and ranking them in terms of feature importance. The latter category opts for resource-related target variables, typically reflecting healthcare demand such as number of CVD-induced hospital admissions or identification of peak demand days of hospitalizations.

This kind of approach usually aspires to build and eventually deploy an early-warning system for medical resource allocation and management.

Starting from the articles focusing on the prediction/ forecast of cardiovascular outcomes, a general observation would be that most of the approaches either forecast cardiovascular disease incidence/ prevalence or investigate cardiovascular cause-specific mortality. Less often, disease risk stratification or disease severity prediction is conducted. Guimbaud et al. [39] have built early-life environmental risk scores (ERS) to assess the cumulative impact of environmental exposures on diverse health outcomes, including cardiometabolic health. The authors aimed at informing practitioners about actionable factors towards prevention measures in healthcare for high-risk children. In Chen et al. the association between machine vision–based built environment and prevalence of cardiometabolic disease at the neighborhood level has been investigated [35]. The authors aspired to identify at-risk neighborhoods, thereby contributing to a more informed and targeted public health strategy for mitigating CV risks associated with specific environmental factors. In Hossain et al. the authors focused on identifying crucial factors in predicting CVD risk [40]. As the authors claimed, the eventual goal of this work has been to enhance clinical practice by providing doctors with a new instrument to determine a patient's CVD prognosis. The primary objective in the study by Nissa et al. has been to identify individuals at risk of developing CVDs and eventually enable healthcare professionals to foster proactive healthcare measures [41]. County phenotypes associated with premature CV mortality have been identified in Dong et al. and their geographic distributions have been investigated using machine learning approaches as well as geographic information systems [42]. The authors conclude that interventions to reduce premature cardiovascular mortality should be targeted to geographic areas with high-risk phenotypes of premature cardiovascular mortality. Leirião et al. have generated forecasts of cardiorespiratory mortality in the elderly aiming at providing decision-makers with a powerful tool for the evaluation of environmental public health-related policies [43]. Martin-Morales et al. predicted CVD mortality with a focus on identifying associated risk factors from a pool of significant nutritional variables [44]. Lee et al. have generated forecasts of mortality from cardiovascular, respiratory, and non-accidental diseases [45]. Proceeding with the studies attempting to predict cardiovascular incidents, Marien et al. have forecasted the number of daily incidents of myocardial infarctions based on case-only data and claimed that the suggested ML approach provides a promising basis to model future MI under changing environmental conditions, as projected by scenarios for climate and other environmental changes [46]. Bhakta et al. have demonstrated the potential of ML models for the early detection of CVD, ultimately enabling medical professionals to implement timely interventions and achieve improved patient outcomes [47]. Liu et al. developed a detection system of stroke survivors with a focus on identifying key factors associated with stroke records [48]. Monaco et al. have evaluated the severity of CVD incidents and identified clinical features mostly associated with CVD risk [49]. Yao et al. have exploited multi-source spatio-temporal data to predict MI severity and to spatially analyze associated risk factors [50]. The authors also proposed urban planning-related directives aiming at reducing the risk and mortality. Atehortua et al. have attempted to demonstrate the potential of exposome-based machine learning as a risk assessment tool by developing an ML model for CVD risk prediction performing comparably to a more integrative model requiring clinical information [51]. Alaa et al. leveraged an automated ML framework applied on non-traditional variables to increase the accuracy of CVD risk predictions in asymptomatic individuals compared to a well-established risk prediction algorithm based on conventional CVD risk factors (Framingham score) and other baseline models [52]. Ren et al. have identified maternal exposure to PM10 as the primary risk factor for congenital heart defects based on two machine learning models [53]. Park et al. [54] developed several updated Environmental Risk Score (ERS) measures constructed to predict GGT, an indicator of oxidative stress, which has been exploited as a summary measure to examine the risk of exposure to multi-pollutants. Subsequently, associations between ERS and cardiovascular endpoints (blood pressure, hypertension and total and cardiovascular disease (CVD) mortality) have been evaluated. Lee et al. [11] performed an exposome-wide association study (ExWAS) on a selection of cardiovascular outcomes (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) using statistics and machine learning and claim to have identified novel risk factors for CVD. Li et al. [55] have identified and ranked prominent factors associated with CVD in a supervised context using CVD diagnoses. In Hsiao et al. [56] environmental and outpatient

records have been utilized for detecting the incidence of four specific categories of cardiovascular diseases. The authors claim that the proposed model can be further developed as a tool for personalized healthcare management. Lastly, Dominici et al. [57] have explored the association between short-term exposure to airborne particles and daily hospitalizations with respiratory or CV etiology on a national scale highlighting this ongoing threat to the health of the elderly population.

Proceeding with the category that focuses on predicting/ forecasting medical resources-related variables, the identified attempts aspiring the development of a healthcare resources management tool will be reported. Sun et al. [58] exploited ML techniques to identify the determinants of population-based CVD outcomes, such as the prevalence of CVD and its related health care costs. The authors focused on county level since counties have departments of health that are equipped with public health personnel and have the capacity to report and monitor real-world cardiovascular prevalence and costs. The findings have important policy implications for controlling the health care costs from CVD. Lin et al. [59] selected emergency department visits with CVD-related etiology as a target variable, emphasizing the critical need for ML-enabled decision support in clinical manpower and resource management. By predicting the incidence of emergent CVD events, they aimed at providing a basis for the development of a management system to address this need. Sajid et al. [60] demonstrated satisfactory performance in predicting CVD incidence with the use of nonclinical features readily available in any healthcare system. The authors aimed at reducing the potential burden of disease on already overburdened health systems in low-middle-income countries, which have limited access to health facilities by enabling the implementation of preventive strategies. Chen et al.[61] focused on providing an early-warning system upon potentially excessive numbers of hemorrhagic stroke admissions to medical institutions by forecasting the demand for hemorrhagic stroke healthcare services. Qiu et al. [62] aimed at the implementation of a decision-making tool for medical resource management by forecasting peak demand days of CVD admissions. The authors have also identified the main weather-related and air quality-related contributors to prediction accuracy. Along similar lines, Lu and Qiu have focused on forecasting daily hospital admissions due to cerebrovascular disease and claimed that their approach offers practical value for hospital management teams in early warning and healthcare resource allocation [63]. In Usmani et al. [64] prediction of the trends of daily and monthly hospitalization has been conducted and along similar lines Jalili et al. [65] forecast the number of hospital admissions of CVD patients without however elaborating on a specific application context. Hu et al. [66] identified major determinants of stroke incidence at the neighborhood level potentially useful to prioritize and allocate resources to optimize community-level interventions for stroke prevention. Nghiem et al. [67] employed a selection of ML algorithms to predict high health-cost users among individuals with CVD, aiming to advance the application of preventive measures for population health improvement and ultimately the optimization of health services planning.

RQ2. Which machine learning algorithms have been applied for investigating exposomics impact on CVDs and which seem to be ranking among the top performers?

Tasks encountered in pertinent literature have been predominantly addressed within classification and regression contexts, with clustering being much less common. Overall, a broad palette of machine learning 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. 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], [40], [45], [46], [47], [48], [54], [60], [61], [62], [63], [67], [68]. The second most commonly encountered comparative approach involves utilizing non-linear and ensemble methods [41], [55], [68].Guimbaud et al. [39] have exploited linear and ensemble algorithms and Jalili et al. [65] have 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, Usmani et al. [64], have focused on the use of non-linear models while Hu et al. [66] have exclusively investigated ensemble models.

Proceeding with the remaining 30% of pertinent literature that does not conduct performance comparison whatsoever, a part of research opted for a non-linear algorithm [42], [43], [56], [57], while another part has focused on an ensemble algorithm of their selection [50], [51], [58], [59].

As it becomes obvious from Figure 4 ensemble algorithms are the most popular choice encountered in pertinent literature [35], [39], [40], [41], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [58], [60], [61], [62], [63], [66], [67], [68], closely followed by non-linear algorithms [35], [41], [42], [43], [44], [45], [46], [47], [48], [49], [54], [59], [60], [62], [64], [65], [66], [67], [68], [69], [70], [71], [72]. 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 XGBoost.

Linear methods exploited in pertinent literature are logistic regression [48], [62],SVM with linear kernel [61], Ridge regression [46],Elastic Net (EN) [63], 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) [44], [46], [49], [50], [55], [61], [62], [68], followed by XGBoost [44], [51], [55], [61], Light Gradient Boosting Machine (LGBM)[44], Gradient Boosting (GB) [46], Random Ferns [55], Bayesian additive regression tree (BART) [54] and Stacking ensemble model [63].

Regarding non-linear methods, the use of artificial neural networks (ANN) and/or DL has been adopted in a considerable part of literature [35], [43], [46], [48], [49], [52], [56], [60], [62], [64], [65], [69].

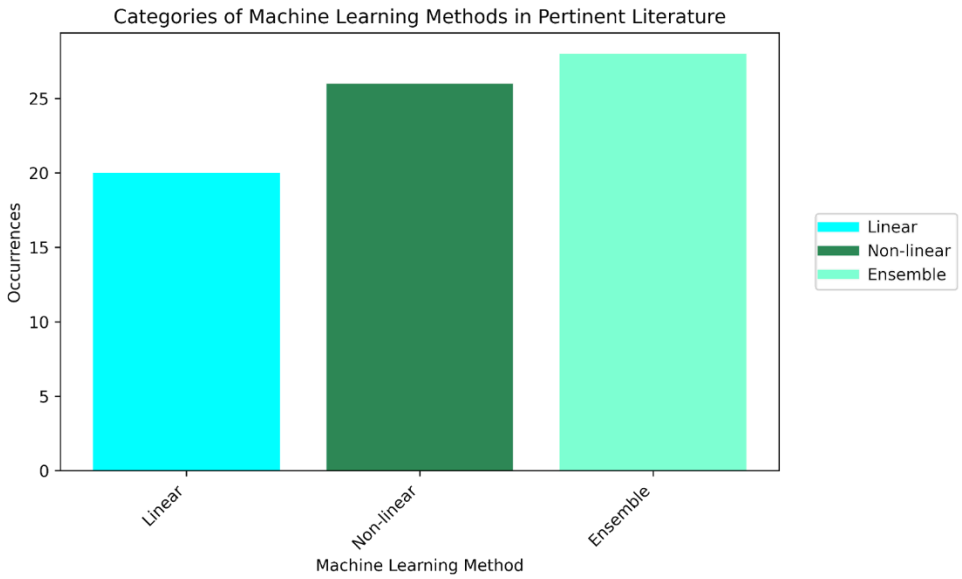

Figure 4. Popularity of categories of ML algorithms as reflected by the number of corresponding occurrences in identified literature.

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 [40]	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 [39]	Regression	Lasso, RF, XGBoost	10-fold cross-validation	R2	SHAP
Nissa et. al. 2024 [41]	Classification	DT, RF, GB, CatBoost, XGBoost, AdaBoost, LGBM	Data splitting with 60:40 ration + 5 and 10-fold cross-validation	ACC, sensitivity PEC, Precision, NPV, FPR, FDR, FNR, F1, MCC	Correlation-based
Nghiem et et al. 2023 [67]	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 [43]	Regression	ANN	cross-validation + external	MAPE	connection weight method
Li et. al. 2023 [68]	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 [63]	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 [51]	Classification	XGBoost	Internal + External	Sensitivity, Specificity, Precision, AUC-ROC	SHAP
Dong *et al*. 2023 [42]	Classification	CART	Internal + External		RF

Martin-Morales 2023 [44]	Classification	LR, SVM, RF, XGBoost, LGBM	5-fold cross-validation	Accuracy, Recall, Precision, F1, AUC-ROC	SHAP
Bhakta et al. 2023 [47] [12]	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 [73]	Regression	LightGBM	10-fold cross-validation	RMSE, MAE	Boruta SHAP
Sun et al. 2023 [58]	Regression	XGBoost	-	-	Gini index
Wang et. al. 2023 [70]	Classification	LR, RF, SVC, MLP, Cox Survival Regression, Random survival Forest, Fast Survival SVM	5-fold cross-validation	AUC, F1 score, Precision, Recall	SHAP
Huang et. al. 2022 [71]	Classification, Regression	Ensemble Classifier based on Naive Bayes, RF and SVM. Ensemble regressor based on General Linear Regression, Support Vector Regressor and Stochastic Gradient Descent)	5-fold cross-validation	AUC-ROC	Support Vector Classifier
Lee *et al*. 2022 [11]	Classification	LR	10-fold cross-validation+ test set	FDR	Lasso, knockoff boosted trees (KOBT), SHAP
Lee et al. 2022 [45]	Regression	XGBoost, RF, Lasso, Ridge, and EN , LSTM, stacked LSTM	10-fold cross-validation + external	MAE, RMSE	-
Marien *et al*. 2022 [46]	Regression	DT, RF, GBR, Ridge, MLP	10-fold cross-validation + external	R2, adjusted R2, ME, RMSE, BIC	DT, RF and GBR, PCA
Yao *et al*. 2022 [50]	Classification and regression	RF	-	Kappa coefficient, Precision, Recall, F1 score and AUC-ROC	SHAP
Li et al. 2022 [55]	Classification	RF, Random Ferns, and XGBoost	10-fold cross-validation + test set	AUC-ROC, AUC-PR	Boruta
Liu et al. 2022 [48]	Classification	LR, ANN, H2O Driverless AI, Isolation Forest (IF).	cross-validation	AUC-ROC, AUC-PR, Categorical Net Reclassification	BoostARoota, SHAP

				Improvement (NRI), Integrated Discrimination Improvement (IDI)	
Testa et al. 2022 [72]	Clustering and inferential analysis	Hierarchical Clustering, cluster selection based on dendrograms	n/a	n/a	n/a
Usmani *et al*. 2022 [69]	Regression	LSTM, ELSTM, DL	train + test	RMSE, MAE	-
Lin *et al*. 2021 [59]	Regression	LSTM		RMSE, MAPE	-
Monaco *et al.* 2021 [49]	Classification	RF, ANN, GLM	*k*=10,5,3 and Leave One Out cross-validation (LOO-CV)	Sensitivity, Precision, F1 score	Boruta
Jalili et al. 2021 [65]	Classification	ANN, LR	cross-validation +testing (Data splitting 70:15:15)	MSE, ER, Pearson's r	
Sajid et. al. 2021 [60]	Classification	LR, ANN, SVM, RF	Data split (70:30) and 10-fold cross validation	Sensitivity, Specificity, Precision, ACC, AUC	Gini
Usmani *et al*. 2021 [64]	Regression	LSTM, ELSTM, DL	Data split (70:30)	MAE, RMSE	
Qiu *et al*. 2020 [62]	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
Chen *et al*. 2019 [61]	Classification	LR, XGBLinear, KNN, RF, XGBTree, and SVMLinear	10-fold cross-validation	AUC-ROC, sensitivity, specificity	Lasso
Alaa et al. 2019 [52]	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,	No selection, only rankings using Gini

				Precision, Recall	Coefficient and Relative Influence
Park 2017 [54]	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 [56]	Classification	Autoencoder and a Softmax classifier	-	-	-
Dominici et al. 2006 [57]	Clustering, Regression	K-means	n/a	n/a	n/a

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

Reference	**Best performer**
Chen et al. 2024 [35]	ET
Hossain et al. 2024 [40]	RF
Guimbaud et al. 2024 [39]	XGBoost
Nissa et al. 2024 [41]	AdaBoost
Bhakta et al. 2023 [47]	XGBoost
Li et al. 2023 [68]	RF
Lu and Qiu, 2023 [63]	Stacking Ensemble Model with four base learners (Ridge, RF, GB, and ANN)
Martin-Morales et al. 2023 [44]	RF
Nghiem et al. 2023 [67]	RF
Wang et al. 2023 [70]	Cox survival regression (L1 penalty)
Lee et al. 2022 [45]	XGBoost, Ridge, and EN
Li et al. 2022 [55]	RF
Marien et al. 2022 [46]	Ridge, MLP
Usmani et al. 2022 [69]	ELSTM
Jalili et al. 2021 [65]	ANN
Monaco et al. 2021 [49]	RF
Sajid et al. 2021 [60]	RF
Usmani et al. 2021 [64]	ELSTM

Hu et al. 2020 [74]	BART
Qiu et al. 2020 [62]	LGBM
Chen et al. 2019 [61]	LR
Ren et al. 2018 [53]	RF
Park et al. 2017 [54]	AENET-I

Table 2. Algorithms reported as top-performers in corresponding comparative studies

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 pertinent literature, whether individually or in combination, are environmental exposure factors, lifestyle και socio-economic status-related factors. The majority of identified articles exploit a combination of the aforementioned categories while the remaining works exclusively consider environmental factors. For a detailed reporting of the categories of exposomics features exploited in literature, see Table 3.

Environmental exposure is the most widely adopted category of exposome with the bulk of relevant literature exploring a broad palette of factors ranging from pollutants concentrations, average sound levels and meteorological parameters to biomarkers reflecting the extent of human exposure to heavy metals. Following closely are socio-economic factors, including factors such as current employment status, education level, income, lack of health insurance and food insecurity. Lastly, lifestyle-related factors, mostly reflecting dietary/sleep patterns along with various health habits, are also addressed in a significant portion of pertinent literature. Notably, environmental exposure is the only category often investigated individually while socioeconomic and lifestyle factors are typically studied in combination with other categories.

The most widely used environmental parameters include air pollutant concentrations and meteorological parameters [43], [45], [46], [53], [64], [69], [72]. Additionally but less often, parameters reflecting noise pollution [39], [51], green spaces [39], [46] and traffic proximity/volume [41], [66] are also included in the exploited feature space. In very few cases, variables reflecting daily or work exposures such as number of smokers at home [11], [39], [52], biohazardous materials [11] and heavy metals concentration based on human scalp hair analysis [49], blood or urine samples are exploited [68]. 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.

Regarding socio-economic factors, a wide range of variables are used, primarily reflecting income, education levels, economic status and/or the subject's occupation [39], [41], [47], [48], [51], [55], [67], [71]. Less often, extra variables such as lack of private health insurance, home ownership, housing type or severe housing cost burden are also considered [42], [44], [58], [70], [74].

In terms of lifestyle factors the focus is predominantly on health-related habits and adopted dietary, physical activity and sleep patterns/quality [11], [39], [41], [42], [44], [48], [51], [52], [55], [58], [67], [70], [71]. The most commonly

considered health habit-related parameters is smoking status and/or alcohol consumption [11], [41], [44], [52], [55], [58], [67], [70], [71].

A considerable part of literature explores combinations of all three categories i.e. environmental, lifestyle, and socio-economic factors [11], [39], [41], [47], [51], [52], [74]. Additionally, socio-economic and lifestyle-related parameters are frequently combined to investigate associations of exposomics data with CVD-related variables [42], [44], [48], [55], [58], [67], [70], [71]. One last commonly encountered combination of categories consists in simultaneous investigation of socio-economic and environmental categories [35], [54], [60], [68].

The remaining part of identified articles exclusively considers environmental parameters, with a part of literature exploiting air quality-related features (mostly air pollutants concentrations) to forecast CVD-related variables [57], [64], [65], [69], and another part further expanding the feature space to include meteorological parameters as well [43], [45], [59], [63], [72]. Additionally, Marien et al. [46] further expanded the feature space with vegetation index. Ohashi et al. [73] is the only work exclusively based on meteorological parameters. Lastly, Monaco et al. [49] focused on heavy metals concentrations obtained by human scalp hair analysis tests to perform ML-enabled estimation of the severity of CVD.

Reference	Exposomic categories included in feature space	Target outcome/variable
Chen et. al. 2024 [35]	socio-economic and, built environment	Coronary heart disease prevalence
Hossain et al. 2024 [40]	socio-economic, lifestyle	CVD diagnosis
Guimbaud et al. 2024 [39]	Air quality (pollutant concentrations), meteorological factors, traffic noise, traffic indicators, natural space, built environment, lifestyle, pollutant biomarkers	Cardiometabolic risk
Nissa et. al. 2024 [41]	environmental, socio-economic and lifestyle factors (including diet, physical activity and sleep hours)	Risk of heart attack
Bhakta et al. 2023 [47]	occupation, environment, lifestyle habits	Heart disease detection
Leiriao *et al*. 2023 [43]	Air quality, meteorological parameters	Cardiorespiratory mortality
Li et al. 2023 [68]	Exposure to heavy metals, socioeconomic	Coronary Heart Disease Risk
Lu and Qiu, 2023 [63]	Air quality (pollutants concentrations)	Daily counts of hospital admissions due to cardiovascular diseases and stroke

Atehortua *et al*. 2023 [51]	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 [42]	Air quality (pollutants concentration), lifestyle, socioeconomic status	Premature cardiovascular mortality
Martin-Mor ales 2023 [44]	Dietary and non-diet-related health data	CVD mortality
Ohashi et. al. 2023 [73]	Daily weather data	CVD Mortality risk
Sun et al. 2023 [58]	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 [70]	Socioeconomic status, lifestyle (including dietary patterns)	CVD Mortality, IHD hospitalization
Huang et. al. 2022 [71]	Social-demographics, lifestyle factors (Dietary, Physical activity, sleep pattern)	Cardiovascular risk
Lee *et al*. 2022 [45]	Meteorological variables, air pollutants	Non-accidental, cardiovascular, and respiratory mortality
Lee et al. 2022 [11]	Chemical and biological exposures, socioeconomic status, food and diet, prescription medication and comorbidities, emotional and mental health, sleep, and genetics.	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 [46]	Meteorological, air quality (pollutants concentration), vegetation index	Daily and annual incidents of myocardial infarctions
Yao *et al*. 2022 [50]	Urban multi-source spatio-temporal big data, road network, demographic,	MI disease severity, MI mortality

	economic, meteorologic data and air pollutants	
Li et al. 2022 [55]	Socio-economic and lifestyle factors	CVD diagnoses
Liu et al 2022 [48]	Demographic Information,Dietary Intake, Health Behaviors,	Stroke incidents
Testa et al. 2022 [72]	Air Pollutants and weather variables	Acute cardiac or cerebrovascular events[60]
Sajid et. al. 2021 [60]	Non-clinical factors such as education, area of living, occupation, exposure to economic problems and the role in the family	Cardiovascular risk score
Usmani *et al*. 2022 [69]	Air quality (pollutants concentration)	Cardiorespiratory mortality
Lin *et al*. 2021 [59]	Air quality, meteorological	Emergency department visits with CVD-related etiology
Monaco *et al.* 2021 [49]	Heavy metal concentrations are extracted by means of TMA hair tests.	Level of CVD clinical risk
Jalili et al. 2021 [65]	Air quality (pollutants concentration), meteorological	CVD rate (as reflected by hospital admissions)
Usmani *et al*. 2021 [64]	Air quality (pollutants concentration)	Monthly cardiorespiratory hospitalization
Qiu *et al*. 2020 [62]	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)
Chen *et al*. 2019 [61]	Air quality (pollutants concentration), meteorological	Demand for hemorrhagic stroke admissions to medical institutions
Alaa et al 2019 [52]	Health and medical history, lifestyle and environment, physical activity, psychosocial factors, dietary and nutritional information, and sociodemographics	CVD risk

Ren et. al. 2018 [53]	Air quality (pollutants concentration), meteorological (humidity, temperature)	Risk of Congenital Heart Defects
Park 2017 [54]	Metal biomarkers in blood and urine reflecting environmental exposure in heavy metals	blood pressure, hypertension and total and CVD mortality
Hsiao *et al*. 2016 [56]	Air quality (pollutants concentration), meteorological	Risk of four categories of cardiovascular diseases (hypertensive, ischemic heart, cerebrovascular disease, other forms of heart disease)[57]
Dominici et al. 2006 [57]	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.

Discussion

This scoping review aimed at presenting the state-of-the-art of ML applications on exposomic 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. study objectives, ML techniques employed and the categories of exposomic data exploited. Based on the insights from the previous sections, we will take a step further and identify key challenges and opportunities in ML-enabled exploitation of exposomic data for CVD-related variables.

As reflected by the timeline of identified publications the presented 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 [75].

Two main categories of studies have been identified based on the selected target variable and the application context of the system under development. The first category focuses on forecasting crucial cardiovascular outcomes or associated risk levels while the second targets resource-related variables typically reflecting healthcare demand such as the number of CVD-induced hospital admissions or the identification of peak demand days of hospitalizations. The former studies

often aim to identify key determinants of the target outcomes and rank them in terms of feature importance and the latter usually aspires to build and deploy an early-warning system for medical resource allocation and management. Overall, the direction of research efforts falls within the scope of AI-driven health interventions aimed at effective risk stratification, public health surveillance, and health policy planning, all of which are described to contribute to health-related sustainable development goals [76].

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 machine learning 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. Ensemble algorithms seem to be the most widely adopted approach and especially Random Forest 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 also enhance predictive modeling. An addition to the toolbox of the well-established linear clustering methods would 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 resides [77].

With respect to investigations related to the employed feature space, environmental exposure seems to be the most widely adopted category within the exposome. The bulk of relevant literature explores diverse factors ranging from pollutant concentrations, average sound levels and meteorological parameters to biomarkers reflecting the extent of human exposure to heavy metals. Following closely are socio-economic factors, encompassing aspects such as current employment status, education level, income, lack of health insurance and food insecurity. Lastly, lifestyle-related factors, mostly reflecting dietary/sleep patterns and quality along with various health habits such as drinking alcohol or smoking, are also addressed in a significant portion of the identified literature. Notably, environmental exposure is the only category often investigated individually while socioeconomic and lifestyle factors are typically studied in combination with other categories.

Important limitations have been reported across pertinent literature. Firstly, validity of exploited ground truth cannot be ascertained e.g. upon the use of diagnostic codes corresponding to hospitalizations or death certificates. Second, in case of self-reported data categories such as lifestyle data, subjectivity and recall bias cannot be avoided. Third, the bulk of conducted research only involved incidents that occurred in a single region. 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. Most of the time, the data sources are not comprehensive and/ or standardized [36]. Fifth, several works report limited availability of data sources and finally, it is difficult to directly compare results from different studies since they are heterogeneous in terms of target, problem framing, study design and sample size and outcome assessment as reflected by the occasional lack of an external validation process but also by the diverse metric scores reported.

Extending the feature space to include non-traditional CVD risk factors spanning environmental, social and life-style domains has shown promising results in improving the performance of conventional models targeting CVD-related variables. In this sense, exploitation of exposome-related variables brings the researcher community one step closer to identifying major environmental, social and lifestyle determinants of CVDs. This additional knowledge at a personal level could form the basis for developing tools for personalized healthcare management. At a broader level (neighborhood, city, etc.), it could help prioritize and allocate healthcare resources. The latter would enhance healthcare workflow optimization and implementation of prevention and intervention measures to reduce CVD-induced healthcare costs.

At this point it should be stressed that a lack of consensus is observed within the research community regarding the distinct categories that constitute the exposome and the specific variables included in each of these categories. This leads to inconsistencies in terminology and classification, hindering comparability and standardization across studies. For instance, part of the research community might refer to 'smoking status' while another part may be making use of alternative terms such as 'tobacco use' etc. This holds for other terms as well, that may be used by researchers interchangeably, despite potentially referring to slightly different concepts. To enhance the reliability and comparability of exposome research, there is a need for standardization of terms and definitions used to describe predictor variables. Overall, standardized protocols for data collection and sharing should be developed.

Finally, there is a need to explore exposomics from a multi-disciplinary perspective. Since its scope is becoming clearer and clearer every day and more and more studies include exposomics in etiological research, actions are needed from multiple stakeholders to join forces for the unification of frameworks and the establishment of guidelines regarding an Exposome Study Design and even a comprehensive exposome database [78].

Conclusions

Even though ML techniques application on exposomics data with a focus on cardiovascular diseases is in its early stages compared to similar use cases that are based on other kinds of -omic data such as genomic data, there is a pronounced increase of pertinent publications during the last years. However, the vast majority of relevant studies has been based on data outside EU territory and specifically on data originating from the US and China. Regarding the ML framing of CVD-related problems, it is worth highlighting the nearly exclusive adoption of a supervised context across identified literature, irrespective of the cardiovascular outcome addressed, with just two works addressing a clustering task. As machine learning applications on exposomics data expand and reach a higher maturity level, it seems to hold promise for uncovering new insights into the environmental determinants of health but also for identifying valuable strategies for CVD prevention and healthcare resource allocation. Towards this aim, further research could focus on the more manageable and easily adjustable modifiable factors in contrast to those that are stiffer such as the socio-economic status and try to use the first as inhibitors to avert a “poor” exposome. Understanding the modifiability of different factors is crucial for public health strategies. Focusing on more modifiable factors can empower individuals to make positive changes in their lives, while also pursuing broader societal and policy changes to address the other less modifiable factors. Finally, there is need of standardization in terms of language so as to enable comparability between different studies as this is often hard even to categorize similar studies using the most prevalent keywords in the literature.

References

[1] “Fact sheets for Press.” Accessed: Jul. 01, 2024. [Online]. Available: https://www.escardio.org/The-ESC/Press-Office/Fact-sheets

[2] “Beating cardiovascular disease — the role of Europe’s environment,” European Environment Agency. Accessed: Jul. 01, 2024. [Online]. Available: https://www.eea.europa.eu/publications/beating-cardiovascular-disease/beating-cardiovascular-disease-the

[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,” *Eur. Heart J.*, 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,” *Eur. Heart J. Suppl.*, 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. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, 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. Off. J. Soc. Toxicol.*, 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 Primer*, 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," *Environ. Res.*, 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," *Circ. Res.*, 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," *Lancet Planet. 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," *Environ. Chall.*, 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," *Environ. Pollut.*, 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 J. Cardiol.*, 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," *Am. J. Lifestyle Med.*, 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," *Environ. Int.*, 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," *Comput. Struct. Biotechnol. J.*, 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," *Eur. Heart J.*, 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 Machine 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] 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.

[40] 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 Cardiovasc. Disord.*, vol. 24, no. 1, p. 214, Apr. 2024, doi: 10.1186/s12872-024-03883-2.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] L. Marien *et al.*, "Machine learning models to predict myocardial infarctions from past climatic and environmental conditions," *Nat. Hazards Earth Syst. Sci.*, vol. 22, no. 9, pp. 3015–3039, Sep. 2022, doi: 10.5194/nhess-22-3015-2022.

[47] 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.

[48] 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.

[49] A. Monaco *et al.*, "Random Forests Highlight the Combined Effect of Environmental Heavy Metals Exposure and Genetic Damages for Cardiovascular Diseases," *Appl. Sci.*, vol. 11, no. 18, Art. no. 18, Jan. 2021, doi: 10.3390/app11188405.

[50] Y. Yao *et al.*, "Assessing myocardial infarction severity from the urban environment perspective in Wuhan, China," *J. Environ. Manage.*, vol. 317, p. 115438, Sep. 2022, doi: 10.1016/j.jenvman.2022.115438.

[51] A. Atehortúa *et al.*, "Cardiometabolic risk estimation using exposome data and machine learning," *Int. J. Med. Inf.*, vol. 179, p. 105209, Nov. 2023, doi: 10.1016/j.ijmedinf.2023.105209.

[52] 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.

[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] 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," *Environ. Health*, vol. 16, no. 1, p. 102, Sep. 2017, doi: 10.1186/s12940-017-0310-9.

[55] 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.

[56] 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.

[57] 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.

[58] F. Sun *et al.*, "Social Determinants, Cardiovascular Disease, and Health Care Cost: A Nationwide Study in the United States Using Machine Learning," *J. Am. Heart Assoc.*, vol. 12, no. 5, p. e027919, Mar. 2023, doi: 10.1161/JAHA.122.027919.

[59] 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.

[60] 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.

[61] J. Chen *et al.*, "Machine Learning-Based Forecast of Hemorrhagic Stroke Healthcare Service Demand considering Air Pollution," *J. Healthc. Eng.*, vol. 2019, no. 1, p. 7463242, 2019, doi: 10.1155/2019/7463242.

[62] 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 Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 83, May 2020, doi: 10.1186/s12911-020-1101-8.

[63] X. Lu and H. Qiu, "Explainable prediction of daily hospitalizations for cerebrovascular disease using stacked ensemble learning," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 59, Apr. 2023, doi: 10.1186/s12911-023-02159-7.

[64] 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.

[65] 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.

[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] 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 Econ. Rev.*, vol. 13, no. 1, p. 9, Feb. 2023, doi: 10.1186/s13561-023-00422-1.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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. Cardiovasc. Cerebrovasc. Dis.*, vol. 9, no. 22, p. e016745, Nov. 2020, doi: 10.1161/JAHA.120.016745.

[75] 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," *Soc. Sci. Med.*, p. 117056, Jun. 2024, doi: 10.1016/j.socscimed.2024.117056.

[76] 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.

[77] M. Meilă and H. Zhang, "Manifold Learning: What, How, and Why," *Annu. Rev. Stat. Its Appl.*, vol. 11, no. Volume 11, 2024, pp. 393–417, Apr. 2024, doi: 10.1146/annurev-statistics-040522-115238.

[78] "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

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

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

CART, Classification And Regression Tree

CatBoost, Categorical Boosting

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

FDR, False Discovery Rate

FNR, False Negative Rate

FPR, False Positive Rate

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

PRESS

RF, Random Forest

RMSE, Root Mean Squared Error

SHAP, SHapley Additive exPlanations

SVC, Support Vector Classification

SVM, Support Vector Machines

XGBoost, Extreme Gradient Boosting