INPLASY PROTOCOL

To cite: Estrella et al. Machine learning for the analysis of healthy lifestyle data: a scoping review protocol. Inplasy protocol 202330065. doi: 10.37766/inplasy2023.3.0065

Received: 18 March 2023

Published: 18 March 2023

Corresponding author:
Tony Estrella

antonio.estrella@uab.cat

Author Affiliation:
Department of Basic Psychology and Sport Research Institute, Universitat Autonoma de Barcelona.

Support: **No financial support.**

Review Stage at time of this submission: **Piloting of the study selection process.**

Conflicts of interest:
None declared.

Machine learning for the analysis of healthy lifestyle data: a scoping review protocol

Estrella, T[1]; Alfonso, C[2]; Capdevila, L[3]; Losilla, JM[4].

Review question / Objective: **The objective of this scoping review is to identify and characterize machine learning algorithms used in data analysis of healthy lifestyle. The specific objectives are the study of a) terminology, b) healthy lifestyle variables analysed either input or output, c) programs and libraries used to analyse data, and d) sources, types, and quality of data analysed.**
Eligibility criteria: **In this scoping review the inclusion criteria from studies that provide empirical information are as follows: a) studies must use machine learning models either supervised or unsupervised learning to analyses lifestyle data (input or output), b) studies must use real data from individuals for analysis, and c) the language of the studies must be English or Spanish. Furthermore, theorical studies focusing on a) the mathematical approach to explain algorithm construction, and b) guidelines for implementing the use of machine learning in the field of health will be excluded. Additionally, studies with simulated data and those aiming to develop a robot or app based on machine learning will be excluded. Regarding lifestyle, studies whose main topic is substance use, such as alcohol intake, or those related to smoking cessation will be manually excluded.**

INPLASY registration number: This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 18 March 2023 and was last updated on 18 March 2023 (registration number INPLASY202330065).

INTRODUCTION

Review question / Objective: **The objective of this scoping review is to identify and characterize machine learning algorithms used in data analysis of healthy lifestyle. The specific objectives are the study of a) terminology, b) healthy lifestyle variables analysed either input or output, c) programs and libraries used to analyse data, and d) sources, types, and quality of data analysed.**

Background: Obesity, cardiovascular disease, poor sleep quality and stress are some of the biggest challenges that public systems must face worldwide (Lee & Yoon, 2018). Their incidence in the general population is increasing exponentially (Ng et al., 2022; Pickens et al., 2018). As a consequence, recommendations from public sectors are directed toward changes in modifiable lifestyle's behaviours to improve health and prevent diseases (Everest et al., 2022; Santos et al., 2023). Leading an unhealthy lifestyle has been linked to an increased risk of multimorbidity (Chudasama et al., 2020; Freisling et al., 2020). On the other hand, a healthy lifestyle (HL) in the general population has positive health consequences, reducing the risk of diseases (Aleksandrova et al., 2014; Vetter & Scheer, 2019; Zhong et al., 2022). A HL is considered a protective factor for chronic diseases including mental health (Ellis et al., 2022; Siew et al., 2022; Wang et al., 2021). Despite the fact that each lifestyle behaviour has been studied as an independent factor (Shan et al., 2020), the interest is growing in understanding the links between behaviours to figure out their influence on health (Aleksandrova et al., 2014; Chudasama et al., 2020; Furihata et al., 2018).

The current accessibility to mHealth and wearables (Free et al., 2010), make it possible to collect ecological and momentary data (EMA methodology), creating a detailed people's profile (Zapata-Lamana et al., 2020). Regarding lifestyle data, the main characteristic is the large volume coming from different sources. In addition, fluctuations in human behaviours make it difficult to extract patterns or predictions over time.

However, advances in computational power and artificial intelligence (AI) models, such as machine learning (ML), have been widely used for solving complex problems with large datasets (Beam & Kohane, 2018). ML models can learn through different type of data (inputs) and generate insight and knowledge (output) for better decision making and outcome prediction (Maleki et al., 2020; Secinaro et al., 2021).

In recent years, ML has moved from computer science to health domains, facilitating screening throughout a large number of variables. Although ML models have been applied for organic diseases (Sharma et al., 2021), they are also being used in the field of mental health (Habib et al., 2022; Le Glaz et al., 2021). Moreover, early identification of risk factors allows acting when treatments are more effective, increasing the likelihood of success (Kabalisa & Altmann, 2021).

Rationale: ML models can be classified into supervised learning (SL), unsupervised learning (USL) and reinforcement learning (RL). In SL the model is trained with labeled data, that is, for each observation there is an associated response measurement (James et al., 2021). The goal of SL is to fit a model that will be able to predict the response (or outcome) when applied to new data. The response could be a continuous variable, such as risk of having a certain disease; these problems are known as a regression algorithm. On the other hand, a model with a categorical variable as response is called a classification algorithm, for instance transforming the continuous variable risk exposed previously into two categories: low risk and high risk.

In contrast to SL, in USL models there is not an associated response to the input, and the model seeks relationships between the observations. Dimension reduction techniques such as principal components analysis or cluster analysis are some of the most common USL methods. RL is out of the scope of this paper, and we turn to Alharin et al. (2020) for more on it.

However, a challenge for ML models is the need of considerable domain expertise for transforming raw data into a suitable feature. Concerning this, deep learning (DL) is facing this challenge due to its better performance at discovering patterns in high-dimensional data (LeCun et al., 2015). DL is a specific subset of ML models based on neural networks with an input layer that receives the data, an output layer that returns the outcome, and typically more than one hidden layer that applies non-

linear transformations (Janiesch et al., 2021).

A recent scoping review aimed to provide an overview of ML methods used in health promotion and behavioral change (Goh et al., 2022), found that the main interventions are those related to physical activity. The results are highlighting an imbalance in lifestyle's studies. Hence, extending the search strategy to a global concept of healthy lifestyle may help with understanding relationships and understudied concepts.

Therefore, this scoping review is appropriate to identify and characterize the ML algorithms applied to HL data. In this way, the process to promote those behaviour that are considered relevant in a HL and prevent risk of disease could be performed from a data-driven approach that enables personalized decision-making (Manktelow et al., 2022).

METHODS

Strategy of data synthesis: In this scoping review we searched primary studies in the 3 principal health databases: PubMed by National Centre by Biotechnology Information (NCBI), PsychINFO by ProQuest and Web of Science by Clarivate.

The search strategy followed the Peer Review of Electronic Search Strategies (PRESS) (McGowan et al., 2016) and PRISMA for Searching (PRISMA-S) guidelines (Rethlefsen et al., 2021), and it consisted of two groups of search terms referring to a) healthy lifestyle, and b) machine learning. We also added a third group of terms preceded by the boolean operator NOT to improve the specificity of the search strategy.

The search strategy was adapted to the specific database syntax. In PubMed we used the following thesaurus subject terms: Machine Learning"[Mesh], "Deep Learning"[Mesh], "Artificial Intelligence"[Mesh], "Neural Networks, Computer"[Mesh], "Healthy Lifestyle"[Mesh], "Health Behavior"[Mesh], "Diet, Healthy"[Mesh], "Exercise"[Mesh], "Sedentary Behavior"[Mesh], "Sleep Hygiene"[Mesh], and "Sleep Quality" [Mesh].

For instance, the search strategy in PubMed was as follows: ((("Healthy Lifestyle"[Mesh] OR "health lifestyle*"[Title/Abstract] OR "healthy lifestyle*"[Title/Abstract] OR "Health Behavior"[Mesh] OR "health behavio*"[Title/Abstract] OR "healthy behavio*"[Title/Abstract] OR (("Sleep Hygiene"[Mesh] OR "sleep hygiene"[Title/Abstract] OR "hygiene, sleep"[Title/Abstract] OR "good sleep habit*"[Title/Abstract] OR "Sleep Quality"[Mesh] OR "sleep qualit*"[Title/Abstract] OR "sleep quant*"[Title/Abstract]) AND ("Sedentary Behavior"[Mesh] OR "sedentary behavio*"[Title/Abstract] OR "behavior, sedentary"[Title/Abstract] OR "sedentary lifestyle"[Title/Abstract] OR "physical inactivity"[Title/Abstract] OR "inactivity, physical"[Title/Abstract] OR "lack of physical activity"[Title/Abstract] OR "Exercise"[Mesh] OR "exercise*"[Title/Abstract] OR "physical activit*"[Title/Abstract]) AND ("Diet, Healthy"[Mesh] OR "diet, health*"[Title/Abstract] OR "health diet*"[Title/Abstract] OR "healthy diet*"[Title/Abstract] OR "health nutrition*"[Title/Abstract] OR "healthy nutrition*"[Title/Abstract] OR "health eat*"[Title/Abstract] OR "healthy eat*"[Title/Abstract])))) AND ("Machine Learning"[Mesh] OR "machine learning"[Title/Abstract] OR "learning, machine"[Title/Abstract] OR "supervised learning"[Title/Abstract] OR "unsupervised learning"[Title/Abstract] OR "transfer learning"[Title/Abstract] OR "Deep Learning"[Mesh] OR "deep learning"[Title/Abstract] OR "learning, deep"[Title/Abstract] OR "Artificial Intelligence"[Mesh] OR "Artificial Intelligence"[Title/Abstract] OR "Computational Intelligence"[Title/Abstract] OR "Machine Intelligence"[Title/Abstract] OR "Computer Vision System*"[Title/Abstract] OR "Neural Networks, Computer"[Mesh] OR "Neural Network, Computer"[Title/Abstract] OR "Neural Networks, Computer"[Title/Abstract] OR "Computer Neural Network*"[Title/Abstract] OR "Neural Network Model*"[Title/Abstract] OR "perceptron*"[Title/Abstract])) NOT ("Robot*"[Title/Abstract] OR "Reinforcement Learning"[Title/Abstract]

OR "Deep Reinforcement Learning"[Title/Abstract])) NOT ("systematic review"[Title/Abstract] OR "review"[Title/Abstract] OR "meta-analys*"[Title/Abstract])
The research was applied in January 2023, being restricted only by language (English and Spanish), with no restriction placed by years of publication.

Eligibility criteria: In this scoping review the inclusion criteria from studies that provide empirical information are as follows: a) studies must use machine learning models either supervised or unsupervised learning to analyses lifestyle data (input or output), b) studies must use real data from individuals for analysis, and c) the language of the studies must be English or Spanish.
Furthermore, theorical studies focusing on a) the mathematical approach to explain algorithm construction, and b) guidelines for implementing the use of machine learning in the field of health will be excluded. Additionally, studies with simulated data and those aiming to develop a robot or app based on machine learning will be excluded. Regarding lifestyle, studies whose main topic is substance use, such as alcohol intake, or those related to smoking cessation will be manually excluded.

Source of evidence screening and selection: In the first place, titles and abstract will be screened by two reviewers, in case of doubt a third reviewer will evaluate whether the paper is going to be included in the next step. In this first round the two reviewers will report their degrade of agreement using the Cohen's kappa coefficient. Second, the full text of the resulting papers will be checked by two reviewers. Each reviewer will mark the papers as include or exclude to the final stage. Discrepancies in this process will be solved by a third reviewer. As well as the first stage, the level of agreement will be reported. Finally, the information will be extracted from the papers included in the review using a checklist. The agreement during the information extraction process will also be reported.

Data management: Mendeley will be used as a reference manager software, results of the search strategy will be entered, and duplicates will be merged or removed.
An ad hoc checklist will be used to extract the information from the papers included. The checklist will be divided into 5 sections:
- General information: authors, title, year, and country of affiliation.
- Methodological data: type of study, aim, year of data collection, form of data acquisition, sample, countries represented in the data, and bias detected and reported.
- Study variables: health issue, lifestyle features, and model's input/output variables.
- Software: Statistical programming language, libraries, and packages.
- Model aspects: Type of problem, stages of ML analysis, ML methods, optimization technique, cross validation, and model performance.

Presentation of the results: The review will be presented as a narrative synthesis and the information will be summarized in tables and figures. The present review will follow the PRISMA-ScR Guidelines (Tricco et al., 2018).
Two tables will be presented, one will include the general information and the methodological data exposed previously, and the second one will include study variables and model aspects. Regarding the figures, two choropleth maps will be created, one for the countries where the papers were published and another for the countries where the data were collected. Also, a figure with the principal software and libraries used in the data analysis process will be described. Finally, a bibliometric analysis will be reported through a thematic map connecting papers.

Language restriction: Included papers must be published in English or Spanish.

Country(ies) involved: Spain.

Keywords: Machine Learning; Artificial Intelligence; Lifestyle; Scoping Review.

Contributions of each author:

Author 1 - Tony Estrella.
Email: antonio.estrella@uab.cat
Author 2 - Carla Alfonso.
Email: carla.alfonso@uab.cat
Author 3 - Lluis Capdevila.
Email: lluis.capdevila@uab.cat
Author 4 - Josep-Maria Losilla.
Email: josepmaria.losilla@uab.cat