Lo Sasso et al. *EPJ Data Science* (2025) 14:76
https://doi.org/10.1140/epjds/s13688-025-00590-6

EPJ Data Science
a SpringerOpen Journal

RESEARCH

Open Access

Exploring the relationship between cancer incidence and the sustainable development goals through complex networks and machine learning

Andrea Lo Sasso[1,2], Loredana Bellantuono[2,3*] and Elisa Omodei[4]

Handling Editor: Mattia Mazzoli

*Correspondence:
loredana.bellantuono@uniba.it
[2]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70125, Bari, Italy
[3]Dipartimento di Biomedicina Traslazionale e Neuroscienze (DiBraiN), Università degli Studi di Bari Aldo Moro, 70124, Bari, Italy
Full list of author information is available at the end of the article

Abstract

The awareness that socioeconomic factors play a significant role in the potential onset of cancer is increasingly widespread; however, a clear understanding of the most influential factors is still lacking. In this study, we explore the relationship between cancer incidence, as recorded by the International Agency for Research on Cancer, and the environmental and socioeconomic well-being of countries, measured by the Sustainable Development Goals (SDGs) indicators. To identify relevant predictors of cancer incidence, we construct a weighted complex network where nodes represent SDG indicators, and links correspond to statistically significant correlations between them. We implement community detection to identify a subset of indicators that incorporates the non-redundant dataset's information, and use the selected features for a machine learning prediction of cancer incidence rates. Furthermore, we highlight the most influential SDG indicators by means of an eXplainable Artificial Intelligence analysis. We find that not only health-related indicators play a key role in explaining cancer incidence, but also factors related to agriculture, resource availability, and water cleanliness. These findings provide insights into the complex interplay between socioeconomic, environmental, and health factors. This study aims to expand knowledge on non-intuitive associations related to cancer onset and may contribute to the development of effective public prevention policies.

Keywords: Network science; Machine learning; SDG; Cancer incidence; XAI

1 Introduction

In recent years, the scientific community has shown a growing awareness of how environmental, economic, and social factors, such as lifestyle and human-nature interactions, affect citizens' well-being and the prevention of chronic diseases [1–3]. In addition to various innovative techniques developed to predict the onset of diseases such as cancer [4–6], research works have investigated correlations between socioeconomic indicators and disease incidence in several countries [7, 8]. For example, studies conducted across different regions, including United States, Finland, and Germany, have demonstrated that social vulnerability and lower socioeconomic status are associated with higher cancer incidence and mortality rates for various cancer types [9–11]. On a global scale, countries with higher levels of education, higher income and lower inequalities, active cancer control policies and programs, and high-performing health systems exhibit better cancer outcomes, as reflected in lower mortality-to-incidence ratios [12].

To date, a considerable number of studies have been conducted in an attempt to identify correlations between cancer incidence and various lifestyle patterns and trends [13]. Some associations between cancer incidence and socioeconomic status (SES) at the local level emerged a few years ago through the analysis of SES patterns in European countries. A population-based review examined 91 English-language studies that focused on cancer registry data, and reported relative risks of European adults between 2000 and 2019 [14]. The review highlighted that subjects with low SES have an increased risk of head and neck, oesophagogastric, liver and gallbladder, pancreatic, lung, kidney, bladder, penile, and cervical cancers. Conversely, high SES was associated with an increased risk of thyroid, breast, prostate, and skin cancers. This body of evidence underscores the relationship between SES and cancer incidence across European countries.

While these investigations primarily focus on identifying correlations between individual indicators and incidence rates, such single-factor analyses are insufficient for capturing the true complexity of today's world. A more holistic perspective is needed, one that acknowledges how the severity and prevalence of a disease are shaped by the combined action of multiple socioeconomic, demographic, and environmental factors. This multifactorial approach is a core tenet of the One Health paradigm, which has already proven invaluable in studies on other disorders where dietary habits, lifestyle, healthcare accessibility, and environmental conditions play a key role [15, 16]. For example, this is the case of cholera infections, which are often a result of fewer water connections and high slum dweller density [17], as well as conditions like antimicrobial resistance, which is related to water pollution and accessibility [18].

Another significant drawback of current research is the absence of a comprehensive global overview that identifies the determinants of cancer incidence across all United Nations Member States (UNMS). Such an analysis would be invaluable, allowing for a comparative assessment of the role of various factors in vastly different socioeconomic and environmental contexts. A prerequisite for this kind of study is the availability of harmonized data acquired using consistent standards across all UNMS. The Sustainable Development Goals (SDGs) framework perfectly addresses this need. Established by the 2030 Agenda for Sustainable Development [19], the SDGs provide a set of globally standardized indicators that offer a common ground for analysis. Specifically, SDG 3 (Good Health and Well-being) focuses on ensuring healthy lives and promoting well-being for all, with Target 3.4 aiming to reduce premature mortality from non-communicable diseases (including cancer) by one-third by 2030 [20].

Recent advancements in Social Physics and complexity science provide new tools to investigate the multi-dimensional socioeconomic determinants of phenomena like cancer incidence. As highlighted by Helbing et al. [21], the quantitative study of many large-scale societal challenges, from systemic inequalities to feedback mechanisms, require a holistic, systems-level approach. In this context, methods from statistical physics can be employed to analyse collective human behaviour and network dynamics, offering tools to understand how lifestyles and social structures may propagate health-related risks [22]. The power of this approach lies in its ability to model the complex interdependencies among socioeconomic factors, which are often overlooked by traditional analyses. Complex networks, in particular, have emerged as a powerful tool for this purpose [23]. This framework allows to capture the multifaceted influence of socioeconomic factors among countries to analyse development patterns, identify communities, and enable more meaningful comparisons across development levels. For instance, research works have used complex network approaches in several ways to investigate the global SDG framework, including: (i) network models where the nodes represent the SDGs themselves, built upon the statistical correlations among their indicators across the UNMS [24]; (ii) networks of UNMS, constructed from the correlations of their socioeconomic indicators [25]; (iii) bipartite networks, comprising both SDGs and UNMS as nodes, to compare countries' progress towards achieving SDG targets [26]. Together, these studies demonstrate how complex networks can reveal the underlying structure of global development and provide a solid foundation for investigating multi-factor phenomena like global cancer incidence.

Building on this, an integrated methodology combining network science with Machine Learning (ML) has emerged as a powerful tool for uncovering hidden structures in complex data and develop robust predictive models. This approach has been further reinforced by the advent of eXplainable Artificial Intelligence (XAI) [27], which addresses the *black box* issue by improving model interpretability, enhancing the reliability and applicability of ML-driven analyses across diverse contexts [28]. Implementing ML and XAI has also generated new insights in the study of SDGs. ML demonstrates significant potential in addressing urban sustainability challenges, particularly about SDG 11, detecting poverty and socioeconomic inequalities [29], monitoring and improving access to quality food and water [30], and enhancing traffic flow management [31]. With regard to factors influencing diseases, a recent review on ML applications in the context of female cancer care underscores the value of computational approaches for identifying and analysing determinants of health outcomes [32]. In this study, Bukke et al. emphasize the links between the SDGs and cancer incidence, demonstrating how disparities in medical treatment and prevention can contribute to cancer incidence. In fact, data analysis reveals that women in regions with limited access to quality healthcare experience higher rates of late cancer diagnosis and mortality. This outcome highlights the need for targeted interventions to improve care access and reduce disparities thereby informing health policies and advancing the achievement of SDG 3 (Good Health and Well-being) and SDG 10 (Reduced Inequalities).

Developing ML models to predict global cancer incidence from SDG indicators represents a compelling research question due to its potential to inform tailored policy design for each country based on its specific socioeconomic and environmental context. To the best of our knowledge, a comprehensive, data-driven approach using the SDG framework for this purpose has not yet been undertaken. Our study addresses this significant gap by developing a predictive ML model to forecast cancer incidence of UNMS from SDG indicators, and implementing XAI to measure the impact of individual features on each prediction.

Although the SDG framework offers a unique, standardized, and multidimensional lens to investigate cancer incidence on a global scale, the considered research question involves significant methodological challenges. Firstly, the stark differences among UNMS lead to high heterogeneity in data availability, which can compromise model performance. Furthermore, the disproportion between the numbers of SDG indicators and UNMS instances poses a substantial risk of overfitting. A key challenge also arises from the fact that XAI methods perform optimally on datasets consisting of non-redundant features. Actually, using XAI on highly correlated indicators can lead to unreliable explanations [33, 34].

To overcome these issues, a robust preprocessing pipeline is essential to obtain a final feature set that is informative, widely available across UNMS, and non-redundant, with a number of features comparable to the number of training instances. In this study, we propose a novel network-based feature selection approach, and apply it to the problem of predicting and explaining global cancer incidence from SDG indicators. We construct a weighted complex network, where nodes represent the SDG indicators, and links correspond to their statistically significant correlations across UNMS. Thus, after filtering on network edges and performing community detection, we identify the most representative indicator in each community. The feature set selected through this network-based procedure is then used to train a ML regression model for predicting cancer incidence rates in UNMS. Finally, we implement XAI techniques to identify the indicators with the largest impact on ML prediction. It is worth highlighting that the pipeline proposed in this study can be transferred to other predictive modelling problems using features from any SDG dataset.

2 Results

2.1 Complex network clustering of SDG indicators

We collect SDG indicator data from the United Nations Sustainable Development Goals Data Portal [35] in 2022 and cancer incidence data from the Global Cancer Observatory, managed by the International Agency for Research on Cancer (IARC) [36], covering the same year. Globally, we have the incidence cancer rate and 124 SDG indicators of 145 UNMS, obtained after applying a selection criterion of at least 70% data completeness on SDG indicators to ensure reliability. This choice excludes all indicators from the SDG 13 goal (Climate Action), as none met the 70% availability threshold in the 2022 dataset. All dataset information is detailed in the *Data collection* section in the *Methods*. We build a network of 124 nodes, each representing an SDG indicator, where links correspond to statistically significant correlations between nodes across UNMS. In the *Network construction* section, we also describe the procedure to identify the optimal network backbone through the disparity filter [37], which removes irrelevant links from the graph. We observe that the disparity filter makes 12 nodes disconnected from the rest of the network, as they have irrelevant connections. Then, we find communities (i.e., groups of nodes that have a stronger relationship with each other than with the rest of the nodes in the network) using the Leiden algorithm [38]. We adopt this algorithm to have a descriptive subdivision of network nodes into communities. Then, we apply community detection to the network giant component, which identifies 15 communities. Moreover, each of the 12 isolated nodes previously cut out by the disparity filter, is treated as a single community.

Figure 1 shows the graph representation of the SDG network after the backbone filtering and community detection. The colour of nodes represents their SDG membership, while the circles represent the communities in which the nodes cluster. Each community is annotated with keywords that provide a comprehensive summary of all indicators included therein. In the Supplementary Data file (SD), Table SD1 reports the community membership of each node.

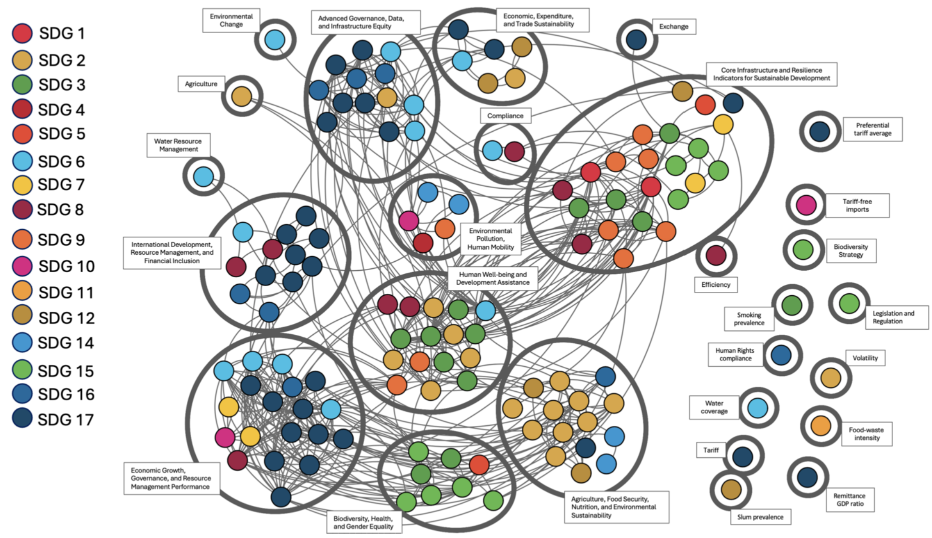

Figure 1 SDG network after the backbone filtering and community detection. The colour of nodes represents their SDG membership, while the circles represent the communities in which the nodes cluster. We adopt the Gephi ForceAtlas2 layout [39] to arrange the nodes of the network. Each community is labelled by keywords summarizing the content of its indicators. This network representation emphasizes how the indicators referred to the same SDG (nodes of the same colour) are not clustered within a single community; rather, network communities are composed of indicators from different SDGs, and no specific goal prevails

Figure 1 shows that network communities aggregate indicators from different SDGs. We observe that SDG 2 (Zero Hunger), SDG 3 (Good Health and Well-Being), SDG 6 (Clean water and sanitation), and SDG 17 (Partnerships for the Goals) are not only the most abundant in terms of indicators but also the ones present in multiple network communities. SDG 6 appears in nine different communities, two of which (*Environmental Change* and *Water Resource Management*) contain only a single indicator. SDG 17 is distributed in seven distinct communities.

Communities identified with this approach consist of indicators originating from different SDGs. From a mathematical standpoint, this behaviour suggests that indicators from the same SDG, on average, are not more significantly correlated with each other than with indicators of different goals. This finding suggests that each SDG exhibits significant variability among its indicators. Moreover, network communities help to clarify which indicators researchers should analyse together in thematic studies rather than concentrating on a single SDG in isolation.

Community detection is employed to identify clusters of highly correlated, and thus redundant, indicators. This process allows us to perform a completely data-driven reduction of the SDG indicators, yielding a dataset that is both informative and non-redundant. As we shall see, this optimized dataset is crucial for the subsequent ML pipeline used to predict global cancer incidence rates. To select a single, representative indicator for each community, we use three criteria based on maximizing different node centrality metrics: betweenness, closeness, and strength, measured at the community level. These metrics are chosen because they quantify different aspects of an indicator's position and influence within the community. Briefly, betweenness identifies indicators that act as bridges between groups of nodes otherwise disconnected or connected through much longer paths, closeness measures an indicator's proximity to all the others, and strength reflects the cu-

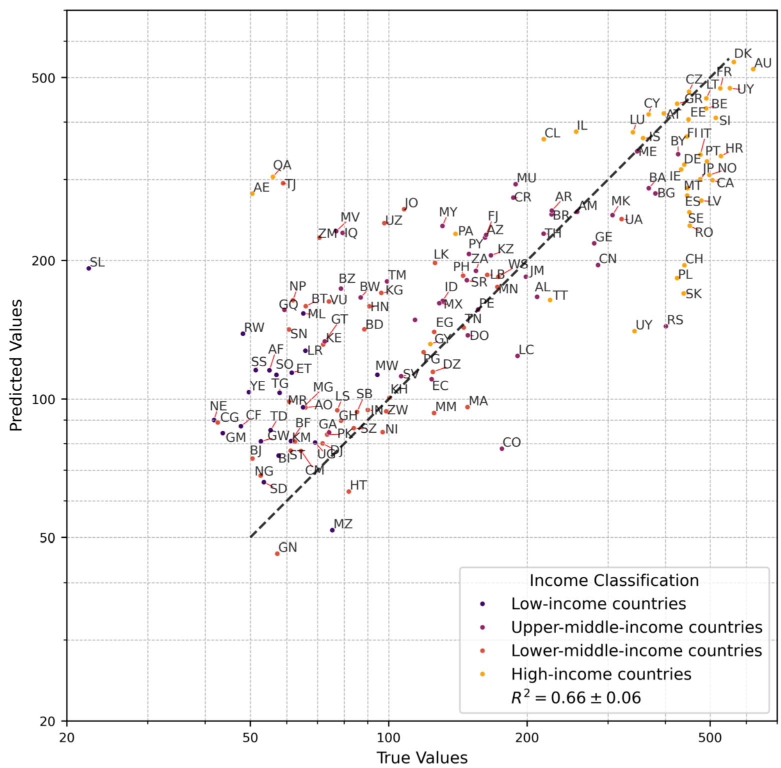

Figure 2 Log scale scatter plot of cancer incidence ML prediction over the UNMS against true IARC values. This scatter plot specifically refers to prediction based on the betweenness dataset. Data points are coloured according to the UNMS income labels assigned by the World Bank [41]

mulative weight of its connections. Detailed definitions of these node centrality metrics can be found in the *Methods* section. In 19 out of 27 communities, the same indicator is selected across all three metrics. On the other hand, even when the indicators selected through the metrics are different, we find significant correlations between them.

Betweenness, closeness and strength provide a robust basis for pinpointing the most representative indicator within each community, as they can be systematically computed for all nodes. Thus, we construct three reduced feature sets, each comprising the most central nodes of network communities, as determined by the three centrality metrics.

2.2 Machine learning predictions

We employ a Machine Learning (ML) pipeline that applies XGBoost [40] to perform regression-based cancer incidence predictions and evaluate how SDG indicators identified through the network are associated with cancer incidence. The pipeline used is described in the *Machine Learning workflow* section, in *Methods*.

Figure 2 shows that our model can predict 66% of the variance in the test set using the betweenness dataset. To evaluate the robustness of our methodology, we also consider the cancer incidence predictions obtained using the closeness and strength datasets. The scatter plots of true cancer incidence rates and predictions from the closeness and strength datasets are shown in Figs. S1a and S2a in the Supplementary Information (SI). The ML models trained on closeness and strength datasets can predict 65% and 61% of the vari-

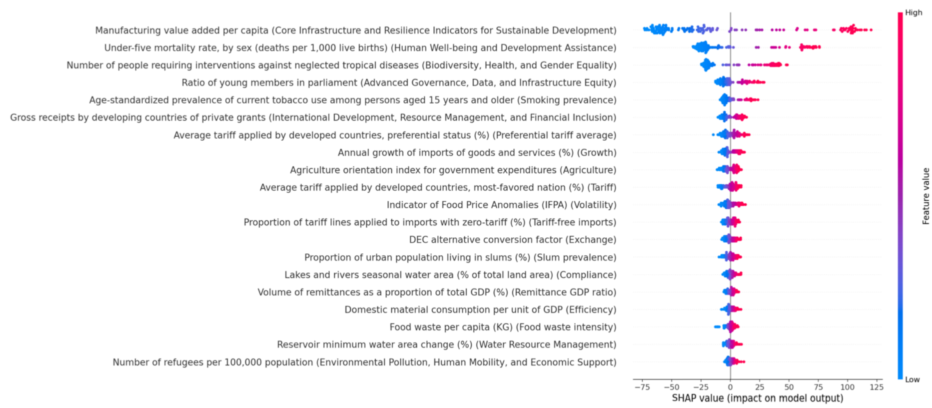

Figure 3 Shap plot highlighting the 20 most important SDG indicators in the ML prediction; each point represents an individual observation, the colour indicates the feature value, and the horizontal position represents the feature's impact on the prediction. Additionally, the SHAP plot ranks features by their importance, positioning in the top the ones with the highest impact on the model's predictions. The SHAP plot highlights that the most important features in the prediction are not solely health-related, but also include socioeconomic, governance, and natural resource supply dimensions

ance in the test set, respectively, showing that our performances are robust and do not depend on the metric used to identify the communities' representative indicators.

In Fig. 2, Fig. S1a, and Fig. S2a, the data points for UNMS are coloured according to the labels of income groups assigned by the World Bank [41]. In line with previous studies on the correlation between GDP and cancer incidence [42–45], our findings indicate that countries with the highest predicted cancer incidence rates are also those classified as high-income by the World Bank.

The predictive performances of ML models trained on feature sets selected through a network-based approach are also compared with those obtained from recursive feature elimination. In particular, we implement the Boruta algorithm [46], a wrapper-based feature selection method that leverages Random Forests [47] to identify all relevant variables in a prediction task. The network-based approach proposed in the present study achieves a performance (R^2 = 0.66 ± 0.06 for the betweenness dataset) comparable to the one obtained by employing the Boruta feature selection method (R^2 = 0.63 ± 0.02). The indicators identified through the Boruta recursive feature selection algorithm, listed in Supplementary Table S1, partially overlap with those obtained through a network-based approach.

2.3 Exploring the most important indicators by means of XAI

To understand the importance of each feature in the ML prediction, we implement a XAI analysis using SHAP values [48], which quantify the impact of SDG indicators on predictions. In a SHAP plot, the points represent individual observations, the colour indicates the feature value, and the horizontal position represents the feature's impact on the prediction. Additionally, features are ranked by their contribution, highlighting at the top which features impact the prediction across the dataset the most.

Figure 3 shows the top 20 SDG indicators selected based on betweenness. The most impactful indicator is *Manufacturing value added per capita* (SDG 9), concerning the country's industrialization level, and belonging to community *Core Infrastructure and Resilience Indicators for Sustainable Development*, which consists of indicators measuring the development of UNMS. Higher values of this indicator positively impact predictions. Descending in the SHAP ranking, we find indicators such as the *Under-five mortality rate* and the *Number of people requiring interventions against neglected tropical diseases*. These indicators are both included in SDG 3 (Good Health and Well-being) but belong to two different communities: *Under-five mortality rate* is part of community *Human Well-being and Development Assistance Indicators*, while *Number of people requiring interventions against neglected tropical diseases* belongs to community *Biodiversity, Health, and Gender Equality Indicators*.

Further down the ranking, we observe the presence of indicators related to food economy and/or waste, such as the *Agriculture orientation index for government expenditures* (SDG 2), the *Indicator of Food Price Anomalies* (SDG 2), and *Food waste per capita* (SDG 12). It is well established that investments in agriculture, along with the need to control food price volatility and food waste are related to an increasing use of pesticides. This usage is further intensified in more developed high-income countries, where pesticide application helps to prolong the shelf life of agricultural products that would otherwise be discarded [49]. Although pesticides enhance the longevity and durability of plant-based products, prolonged exposure may adversely affect consumer health [50], potentially contributing to an increased risk of cancer. The link between agricultural chemical usage and increased cancer mortality rates in rural areas was unveiled in different studies related to the United States [51] and Argentina [52]. The prominent results obtained through the SHAP analysis provide additional evidence supporting the association between higher levels of food waste, increased government agricultural spending, and elevated cancer incidence rates.

A direct link with cancer also emerges through the contribution of indicators such as the *Age-standardized prevalence of current tobacco use*, which corroborates research highlighting a strong correlation between early-onset smoking and an increased risk of cancer, particularly lung cancer. The high correlation between smoking and lung cancer incidence is already well known in the literature, as many studies state [53–58]. Another interesting yet non-intuitive result is the presence of indicators related to SDG 6 (Clean water and sanitation) on water, its management and the natural resources UNMS have in their land. The SHAP analysis returns SDG 6 indicators as *Lakes and rivers seasonal water area (% of total land area)* and *Reservoir minimum water area change (%)*. This finding further reinforces existing research that has established associations between various water pollutants and specific types of cancer in different regions. Such associations have been identified in the United States [59], the United Kingdom [60], and China [61]. Another water-related indicator emerges in Fig. S1b and Fig. S2b in SI: the *Proportion of population using basic drinking water services, by location (%)*, which pertains to SDG 1. This indicator is absent in Fig. 3, replaced by the indicator on manufacturing (see Table 1 in SD for more information).

The other indicators reported in Fig. 3 have a more substantial economic connotation and reflect cooperation among UNMS. This is the case for *Gross receipts by developing countries of private grants*, *Annual growth of imports of goods and services (%)*, *Decimal Currency units (DEC) alternative conversion factor*, *Average tariff applied by developed countries (most-favoured nation status)*, *Average tariff applied by developed countries (and preferential status)*, and *Volume of remittances (in United States dollars) as a proportion of total GDP (%)*, all of which pertain to SDG 17 (Partnerships for the Goals). The high values of these indicators in the SHAP plots confirm that countries with higher socioeconomic status tend to exhibit higher cancer incidence. The relationship between socioeconomic status and cancer incidence is well-documented in literature. In wealthier countries, a higher incidence of cancer is primarily observed due to more advanced diagnostic and registration systems [62], which often lead to overdiagnosis—that is, the identification of multiple conditions in the same patient—thereby increasing overall incidence rates [63]. Additionally, in more developed high-income countries, sedentary lifestyles and high-calorie diets further contribute to this increase [64]. Moreover, higher levels of urbanization, characteristic of these developed countries, are associated with greater incidence of certain diseases, largely due to exposure to environmental pollutants in urban areas [65]. Finally, considering the higher socioeconomic status and longer life expectancy in these countries, chronic diseases typical of advanced age, such as cancer, become more prevalent, resulting in an overall higher cancer incidence rate [66].

3 Discussion and conclusions

Predicting cancer incidence rates using Sustainable Development Goal (SDG) indicators through complex network modelling provides valuable insights. Notably, SDG indicators often show clustering with those associated with other SDGs. As illustrated in Fig. 1, community detection highlights groups of indicators that span multiple SDGs, indicating low internal correlation within individual SDGs and stronger connections between them. This suggests a high degree of variability among indicators of each SDG, with community detection offering a more diverse and non-redundant representation. It also emphasizes which indicators should be analyzed together in thematic studies, transcending SDG boundaries.

In order to construct an efficient and interpretable model, the datasets used for the ML predictions are generated through a network-based approach, which involves selecting the most representative indicators based on a reproducible pipeline while avoiding biases in the feature extraction criteria. Furthermore, employing different network metrics for node selection provides robust results, as the prediction performances obtained using betweenness, closeness, and strength datasets are comparable. The proposed construction of a network and subsequent community detection constitutes a methodological advancement designed to reduce the number of indicators within a feature dataset. This approach is generalizable to diverse feature selection tasks, extending well beyond the specific ML application to SDG analysis. We also implement a ML prediction pipeline using recursive feature elimination with the Boruta algorithm. The performance achieved by this approach is comparable with the network-based one. However, it is worth noting that recursive feature elimination methods select indicators by relying on access to the prediction target within a the training subset. By contrast, our network-based approach does not utilize the target variable at any point during the feature selection process. Additionally, it allows for the reduction of redundant features within the dataset used for prediction.

Then, the XAI analysis we implement underscores the key role of indicators that impact predictions. This study focuses on predicting cancer incidence, highlighting the significant role of indicators related to Sustainable Development Goal 3 (SDG 3), which rank among the top positions in the SHAP plot. As shown in Fig. 3, socioeconomic indicators related to the agricultural economy and the management and cleanliness of water resources (linked to SDG 6) emerge as influential factors. Our findings, discussed in the Results section regarding XAI outcomes, align with previous studies that identify a direct correlation between economic factors and cancer incidence, referring to issues such as food management [51, 52], water cleanliness [59–61], and smoking exposure [53–58]. The socioeconomic indicators presented in the top-20 SHAP plot (Fig. 3) underscore their role as data-driven insights in cancer incidence prediction, enabling the formulation of effective policies aimed at enhancing the overall well-being of UNMS. The betweenness dataset exhibits slightly higher predictive performance (0.66 ± 0.06) compared to the other datasets, indicating that distinct indicators are crucial for prediction.

Our work achieves satisfactory performance in predicting cancer incidence, by relying solely on the widest set of SDG indicators available for the UNMS. In a recent contribution, Umar et al. [67] developed a ML model that reached higher predictive accuracy for cancer incidence in the United States ($R^2 = 0.81$) by integrating both social indicators, related to education and employment, and cancer-specific variables, such as the mean number of reported cancer deaths and the mean per capita cancer diagnoses. To the best of our knowledge, however, a ML model capable of effectively predicting cancer incidence on a global scale using exclusively SDG indicators across a broad set of UNMS is still missing.

To conclude, we should also highlight that the comprehensive implementation of ML-based tools, like the one presented in this study, for sustainable development, faces several limitations, as highlighted in other studies, mainly due to low model transferability and poor generalizability [68, 69]. One limitation we encounter in this study concerns the availability of SDG indicator data. We analyze all indicators from each SDG, selecting only those available for at least 70% of UN member states. This choice is crucial for implementing a reliable and generalizable ML model. However, this selection excludes all SDG 13 indicators (Climate Action), as none meet the 70% availability threshold in the 2022 dataset. Additionally, cancer incidence data provided by the International Agency for Research on Cancer (IARC) is not fully available in all countries. Data completeness is especially important for lower and lower-middle-income countries, where disease detection rates are often low due to inadequate tools for accurate and comprehensive case registration.

Finally, the interesting, albeit not immediately intuitive, results presented in this work enhance our understanding of the complex interactions at play in cancer incidence. Furthermore, the study contributes to the development of effective public prevention policies by advocating for an approach that does not focus solely on a single mainly correlated factors, but rather on a varied set of social, economic and environmental aspects.

4 Methods

4.1 Data collection

The SDG dataset is obtained from the United Nations Sustainable Development Goals Data Portal [35], which gathers all sub-indicators linked to each SDG (we further refer to them as SDG indicators). For our analysis, we selected SDG indicators from 2022. The 2022 dataset consists of 509 indicators, and, for each, we apply a selection criterion of at least 70% data completeness (i.e., the indicator needs to be available for at least 70% of the UNMS to be included in our analysis) to ensure variability of data and generalizability [70]. The selection outputs 141 indicators from all SDGs, except for SDG 13 (Climate Action), which records data only for a limited number of UNMS. The cancer incidence data, also for 2022, is downloaded by the *Cancer Today* dataset available on the Global Cancer

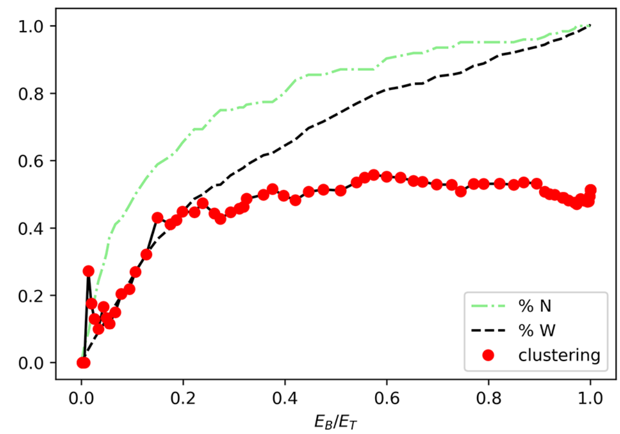

Figure 4 Behaviour of the network as the parameter α varies within the interval [0,1]. For each increasing value of α, we report the fraction of network nodes (%N, green dotted line), weights (%W, black dotted line) and clustering coefficient (red dots) as a function of the fraction of edges between the thresholded network and the whole network (E_B/E_T). Following the literature, the threshold α is chosen in order to maximize the clustering coefficient. In this study, α is equal to 0.56

Observatory, managed by the International Agency for Research on Cancer (IARC) [36]. It tracks the incidence and mortality rates in UNMS, offering a standardized approach to analysing cancer trends across regions for research, prevention, and policy-making. Globally, the number of UNMS initially considered in the construction of the network and community detection is 185. This number decreases to 145 for the ML predictions, as 145 corresponds to the UNMS for which cancer incidence values are available from IARC.

4.2 Network construction

For this study, we construct a weighted complex network. In this network, each node is an SDG indicator, and links correspond to statistically significant correlations between nodes calculated by SDG data recorded across UNMS. If an indicator appears in the lists of different SDGs, we count it once with the first SDG affiliation we find (see Doubled column in Table SD1 in the Supplementary Data for more details). This process returns 124 indicators we use for network modeling. We calculate the Pearson correlation between indicators and the related p-value using the Python Sklearn package [71]. Links between nodes are accepted and weighted using the value of their Pearson correlation, provided that the significance condition (i.e., p-value $\leq 10^{-3}$) is satisfied.

To further reduce network complexity by preserving significant connections while eliminating statistically weak links, we apply the Disparity Filter [37], i.e., a filter based on the local identification of the statistically relevant weight heterogeneities. This method filters the backbone of dominant connections in complex weighted networks, preserving structural properties and hierarchies at all scales. For each link, the disparity filter imposes a significance level α, selecting those weighted links that can be considered incompatible with a random distribution.

Figure 4 illustrates the behaviour of the network as the parameter α varies within the interval [0, 1]. For each increasing value of α, we construct a new version of the network that includes only the edges whose associated α value is below the given threshold. For each of these thresholded networks, we track the fraction of remaining nodes (N%) and edges (E_B/E_T) with respect to the original network, as well as the total retained weight of the surviving connections (W%), and the average clustering coefficient. In Fig. 4, we plot the fraction of N% and W% as a function of the fraction of edges in the thresholded networks, within the corresponding clustering coefficient value.

Following the same procedure adopted in Ref. [37], we use the clustering coefficient as a metric to decide the optimal value of significance level α. We choose $\alpha = 0.56$, i.e. the significance value that maximizes the clustering coefficient in our network.

4.3 Community detection

To identify a descriptive pattern of how nodes aggregate into communities, we perform community detection using the Leiden algorithm [38]. We set the resolution parameter, denoted by γ, to 1, while keeping the remaining algorithm parameters at their default values ($\beta = 0.05$, objective function = modularity). For each resolution value, we conduct 100 algorithm runs (K = 100), each with a different pseudo-random number generator seed. We then apply majority voting to select the most consistent partitions from the resulting outputs. To enhance the robustness of this procedure, we utilize a stability criterion that assesses the similarity between different partitions p_j (where j = 1, ..., K) based on the average Normalized Mutual Information

$$< NMI >= \frac{2}{K(K-1)} \sum_{a=1}^{K-1} \sum_{b=a+1}^{K} NMI(p_a, p_b), \tag{1}$$

where $NMI(p_a, p_b)$ is the Normalized Mutual Information between a given pair of partitions, and $\frac{2}{K(K-1)}$ is the number of distinct pairs. The majority partition over K = 100 runs can be approved only if $\langle \text{NMI} \rangle \geq 0.95$, and if

- it is non-trivial (i.e., not consisting of a single community)
- it is not too fragmented, namely it does not contain communities whose cardinality is less than 5% of the cardinality of the partitioned network.

If the majority voting results obtained for 100 runs, at different values of the resolution γ, satisfy the above conditions, we choose the output with larger $\langle \text{NMI} \rangle$, and the majority partition corresponding to this choice is identified as the result of community detection. We record $\langle \text{NMI} \rangle = 0.98$, which attests the stability and robustness of the partitions we obtained during iterations.

4.4 Constructing ML feature datasets by means of network metrics

In order to build a parsimonious ML model, we reduce the feature set dimensionality. We compute for each node the following centrality metrics: betweenness, closeness, and strength [72], which allow to characterize the centrality of a node within a network. In particular, betweenness quantifies the extent to which a node v lies on the shortest paths between pairs of other nodes in the network. It is mathematically defined as:

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{2}$$

where σ_{st} is the total number of shortest paths between nodes s and t, $\sigma_{st}(v)$ is the number of these shortest paths that pass through node v. The sum runs over all node pairs

(s,t) such that $s \neq v \neq t$. In weighted graphs like the one in our analysis, shortest paths are computed using edge weights as distances. Closeness centrality, then, measures how efficiently information can spread from a node to all other nodes in the network. It is formally defined as:

$$C(v) = \frac{N-1}{\sum_{u \neq v} d(v,u)}, \tag{3}$$

where N is the total number of nodes in the network, $d(v,u)$ is the shortest path distance between nodes v and u, and the sum runs over all nodes $u \neq v$. Also here for weighted graphs, shortest paths are calculated based on the sum of edge weights. The *strength* of a node v in a weighted graph represents the sum of the weights of all edges incident to v. It is defined as:

$$S(v) = \sum_{u \in N(v)} w_{vu}, \tag{4}$$

where $N(v)$ denotes the set of neighbours of node v, w_{vu} is the weight of the edge connecting v and u. These metrics are calculated for the network nodes by considering, for each community, only the sub-graph composed of the subset of nodes within community, isolating them from the rest of the network. For each of the 15 communities, we select the most representative indicator for each metric as the node with the highest centrality value. Specifically, we construct: (i) the *betweenness dataset* of features by choosing, for each community, the node with the highest betweenness centrality; (ii) the *closeness dataset* of features by selecting the node with the highest closeness centrality; (iii) the *strength dataset* by identifying the nodes with the highest strength. If a community contains a single node (i.e., the isolated nodes), we select it as the representative indicator. All information on datasets is displayed in Table SD1 in the Supplementary Data.

We also include into ML dataset those indicators that are detached from the rest of the network. Since they exhibit very weak correlations with other indicators, they provide non-redundant information compared to the indicators within the communities. For this reason, these indicators can be retained in the dataset for ML prediction. We report all information regarding nodes included in the dataset for ML analysis in Table SD1.

4.5 Machine learning workflow

Machine learning regression is conducted following the workflow shown in Fig. 5. We apply the same procedure across all datasets exposed in the previous section. The aim of this approach is evaluating the predictive power of those indicators in forecast cancer incidence data, and identifying the most predictive SDG indicators using XAI.

To achieve our aim, we separate initial data into two parts: the *training set*, consisting of 90% of samples, and the *test set*, consisting of the remaining 10%. We implement an eXtreme Gradient Boosting (XGB) regressor [40] using a Grid Search Cross-Validation with stratification on the training set. This technique allows to find the best hyper-parameter set for a machine learning model by systematically searching through a predefined parameter grid and evaluating each parameter combination using cross-validation to select the optimal one. In particular, we adopt a stratified Grid Search CV, ensuring a balanced class distribution of UNMS on the basis of their income group (as defined by the World Bank)

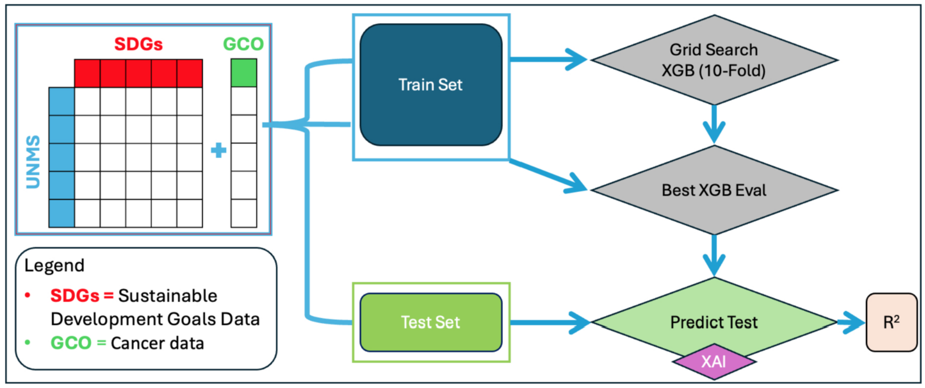

Figure 5 Overview of the workflow for training model, validation, and XAI analysis adopting the XGB regression algorithm. During the training process, the GridSearchCV algorithm is applied to find the best configuration of the XGB regressor. Then, SHAP values are recorded during prediction of the optimal configuration of the regression algorithm, to identify the subset of features which impact the most on the model's outcomes. For further details on XAI, refer to sections *eXplainable Artificial Intelligence*. We repeat this entire workflow 10 times, changing the random seed parameter at each repetition

Table 1 Parameters implemented for Grid Search Cross-Validation on the training set by means of XGB regressor. Grid Search Cross-Validation explores a predefined hyper-parameter grid, maintaining balanced income-group stratification across folds to identify the optimal parameter combination

Parameter	Value
Max depth	3, 5, 7
Learning rate	0.01, 0.1, 0.3
Number of estimators	400, 450, 500, 550, 600
Min child weight	1, 3, 5
Gamma	5, 10
Subsample	0.3, 0.7, 0.8
Colsample by tree	0.1, 0.2, 0.3
Eta	0.1, 0.3, 0.5, 0.7, 1

across folds while tuning hyper-parameters. In Table 1, we report the XGB parameters we varied during Grid Search CV.

This analysis is conducted using the Python Sklearn package [71]. Once the best parameter configuration is obtained, we use it to train the XGB regression model and then we move to evaluate its predictive power on test set using the coefficient of determination. Every time the algorithm is trained over a repetition, we also record SHAP values (see section *eXplainable Artificial Intelligence* for more details). SHAP is implemented to identify the subset of features (SDGs indicator) that impact the most on the regression outcomes.

4.6 eXplainable Artificial Intelligence (XAI)

The eXplainable Artificial Intelligence (XAI) framework includes various techniques emphasizing informativeness, uncertainty estimation, generalization, and transparency. Among those techniques, we use SHapley Additive exPlanations (SHAP), a model-agnostic posthoc local explanation method based on Shapley values which originates from cooperative game theory [27, 48]. We employ the SHAP algorithm to determine the importance of SDG indicators. The SHAP method implements interpretable linear models for each instance, highlighting the contribution of each feature to its prediction. The SHAP value for a particular (feature, instance) pair is obtained by assessing the difference in the model's prediction with and without that feature. In the context of XAI, this procedure involves evaluating the model's output across all possible subsets $F \subseteq S$ of the complete feature set S, simulating the effect of including or excluding specific features in order to determine their individual contributions to the prediction. Specifically, if $f_x(F)$ is the model's prediction for instance x using a subset F that excludes, for example, the jth feature, and $f_x(F \cup \{j\})$ is the prediction when the jth feature is included, the marginal contribution of the jth feature is calculated as the difference $f_x(F \cup \{j\}) - f_x(F)$. The SHAP value for the jth feature in instance x is then obtained by summing this difference over all possible subsets:

$$SHAP_j(x) = \sum_{F \subseteq S-\{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!}[f_x(F \cup \{j\}) - f_x(F)], \tag{5}$$

where $|F|!$ and $(|S| - |F| - 1)!$ represent the number of permutations of features in the subsets F and $S - \left(F \cup \{j\}\right)$, respectively, and $|S|!$ is the total number of feature permutations [73]. In the present study, we apply the XAI analysis to the outcomes provided by the best-performing XGB regressor for each of the considered datasets.

SHAP summary plots, such as the one displayed in Fig. 3, provide a very effective way to visualize the XAI outcomes. In such representations, each data point corresponds to a specific instance and indicates the SHAP value (horizontal coordinate) for a given feature (row name). SHAP values quantify the impact of each feature on the prediction for a given sample. In summary plots, features are ranked in descending order according to their mean absolute SHAP value across all observations. The colour of each point represents the value of the corresponding feature. Therefore, a pronounced chromatic separation in the distribution of SHAP values for a given feature corresponds to a situation of high interpretability, in which the considered feature consistently impacts the regressor's predictions across all instances. Hence, the colour and position of points in the SHAP summary plot are critical elements for interpreting the contributions of different features to the predictions made by ML models. In this study, the SHAP value computation is implemented in the Python (version 3.9) package Shap [74] (version 0.41.0).

Abbreviations

MIRs, Mortality-to-Incidence Ratios; UN, United Nations; SDGs, Sustainable Development Goals; UNMS, United Nations Member States; IARC, International Agency for Research on Cancer; XAI, eXplainable Artificial Intelligence; ML, Machine Learning; NMI, Normalized Mutual Information; XGB, eXtreme Gradient Boosting; SHAP, SHapley Additive exPlanations.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00590-6.

Additional file 1. (XLSX 19 kB)
Additional file 2. (PDF 2.0 MB)

Acknowledgements

Code development/testing and results were obtained on the IT resources hosted at ReCaS data center financed by the Italian MIUR (PONa3_00052, Avviso 254/Ric.).

Author contributions

A.L.S., L.B. and E.O. conceived the model; A.L.S. wrote the code used for computations and performed the analysis, A.L.S., L.B. and E.O. wrote the paper; L.B. and E.O. supervised the design of the rating scheme of performances; A.L.S., L.B. and E.O. interpreted the results, revised the text and approved the final version of the paper.

Funding information
L.B. is supported by Università degli Studi di Bari Aldo Moro through Fondo per la Qualità e l'Internazionalizzazione della Ricerca.

Data availability
The data that support the findings of this study are either publicly available on databases cited in the bibliography, or reported in the Supplementary Data file.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Dipartimento Interateneo di Fisica, Università degli Studi di Bari Aldo Moro, 70125, Bari, Italy. [2]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70125, Bari, Italy. [3]Dipartimento di Biomedicina Traslazionale e Neuroscienze (DiBraiN), Università degli Studi di Bari Aldo Moro, 70124, Bari, Italy. [4]Department of Network and Data Science, Central European University, A-1100, Vienna, Austria.

Received: 11 May 2025 Accepted: 28 September 2025 Published online: 28 October 2025

References

1. Braveman P, Egerter S, Williams DR (2011) The social determinants of health: coming of age. Annu Rev Public Health 32(1):381–398
2. Livingston V, Jackson-Nevels B, Reddy VV (2022) Social, cultural, and economic determinants of well-being. Encyclopedia 2(3):1183–1199
3. Metzler M (2007) Social determinants of health: what, how, why, and now. Prev Chronic Dis 4(4):85
4. Weinstein IB (1991) Cancer prevention: recent progress and future opportunities. Cancer Res 51(18_Supplement):5080–5085
5. Umar A, Dunn BK, Greenwald P (2012) Future directions in cancer prevention. Nat Rev Cancer 12(12):835–848
6. Gu KJ, Li G (2020) An overview of cancer prevention: chemoprevention and immunoprevention. J Cancer Prev 25(3):127
7. Grant WB (2013) A multicountry ecological study of cancer incidence rates in 2008 with respect to various risk-modifying factors. Nutrients 6(1):163–189
8. McCormack V, Boffetta P (2011) Today's lifestyles, tomorrow's cancers: trends in lifestyle risk factors for cancer in low-and middle-income countries. Ann Oncol 22(11):2349–2357
9. Mehta A, Jeon WJ, Nagaraj G (2024) Association of us county-level social vulnerability index with breast, colorectal, and lung cancer screening, incidence, and mortality rates across us counties. Front Oncol 14:1422475
10. Benigni R, Giuliani A (2002) Cancer incidence and socioeconomic geography of Finland: a correlation study. J Environ Sci Health, Part C 20(1):29–43
11. Eberle A, Luttmann S, Foraita R, Pohlabeln H (2010) Socioeconomic inequalities in cancer incidence and mortality—a spatial analysis in Bremen, Germany. J Public Health 18:227–235
12. McDaniel JT, Nuhu K, Ruiz J, Alorbi G (2019) Social determinants of cancer incidence and mortality around the world: an ecological study. Glob Health Promot 26(1):41–49
13. Bahrami H (2024) Interpreting cancer incidence rates and trends: a review of control factors and worldwide statistics. J Cancer Res Pract 11(1):7–17
14. Mihor A, Tomsic S, Zagar T, Lokar K, Zadnik V (2020) Socioeconomic inequalities in cancer incidence in Europe: a comprehensive review of population-based epidemiological studies. Radiol Oncol 54(1):1
15. Marais B, Hesseling A, Cotton M (2009) Poverty and tuberculosis: is it truly a simple inverse linear correlation? Eur Respir J 33(4):943–944
16. Bellantuono L, Monaco A, Amoroso N, Lacalamita A, Pantaleo E, Tangaro S, Bellotti R (2022) Worldwide impact of lifestyle predictors of dementia prevalence: an eXplainable Artificial Intelligence analysis. Front Big Data 5:1027783
17. Velimirovic B, Subramanian M, Sadek F (1975) Socio-economic and environmental factors and human health example of cholera el tor in manila. Zentralbl. Bakteriol. Orig. Reihe B 160(1):1–27
18. Monaco A, Caruso M, Bellantuono L, Cazzolla Gatti R, Fania A, Lacalamita A, La Rocca M, Maggipinto T, Pantaleo E, Tangaro S, Amoroso N, Bellotti R (2025) Measuring water pollution effects on antimicrobial resistance through explainable artificial intelligence. Environ Pollut 367:125620
19. Colglazier W (2015) Sustainable development agenda: 2030. Science 349(6252):1048–1050
20. Bennett JE, Kontis V, Mathers CD, Guillot M, Rehm J, Chalkidou K, Kengne AP, Carrillo-Larco RM, Bawah AA, Dain K, et al (2020) Ncd countdown 2030: pathways to achieving sustainable development goal target 3.4. Lancet 396(10255):918–934
21. Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U, Woolley-Meza O, Moussaid M, Johansson A, Krause J, Schutte S, et al (2015) Saving human lives: what complexity science and information systems can contribute. J Stat Phys 158(3):735–781

22. Jusup M, Holme P, Kanazawa K, Takayasu M, Romić I, Wang Z, Geček S, Lipić T, Podobnik B, Wang L, et al (2022) Social physics. Phys Rep 948:1–148
23. Amaral LA, Ottino JM (2004) Complex networks: augmenting the framework for the study of complex systems. Eur Phys J B 38:147–162
24. Bellantuono L, Monaco A, Amoroso N, Aquaro V, Lombardi A, Tangaro S, Bellotti R (2022) Sustainable development goals: conceptualization, communication and achievement synergies in a complex network framework. Appl Netw Sci 7(1):14
25. Bellantuono L, Monaco A, Tangaro S, Amoroso N, Aquaro V, Bellotti R (2020) An equity-oriented rethink of global rankings with complex networks mapping development. Sci Rep 10(1):18046
26. Sciarra C, Chiarotti G, Ridolfi L, Laio F (2021) A network approach to rank countries chasing sustainable development. Sci Rep 11:15441
27. Merrick L, Taly A (2020) The explanation game: explaining machine learning models using Shapley values. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E (eds) Machine learning and knowledge extraction. Springer, Cham, pp 17–38
28. Bellantuono L, Tommasi R, Pantaleo E, Verri M, Amoroso N, Crucitti P, Di Gioacchino M, Longo F, Monaco A, Naciu AM, et al (2023) An explainable artificial intelligence analysis of Raman spectra for thyroid cancer diagnosis. Sci Rep 13(1):16590
29. Omodei E, Garcia-Herranz M, Paolotti D, Tizzoni M (2022) Complex systems for the most vulnerable. J Phys Complex 3(2):021001
30. Omodei E (2023) Using computational tools to monitor and improve access to quality food and water. Nat Comput Sci 3(9):726–728
31. Leal Filho W, Mbah MF, Dinis MAP, Trevisan LV, Lange D, Mishra A, Rebelatto B, Hassen TB, Aina YA (2024) The role of artificial intelligence in the implementation of the un sustainable development goal 11: fostering sustainable cities and communities. Cities 150:105021
32. Bukke SPN, Komarla Kumarachari R, Komarla Rajasekhar ES, Dudekula JB, Kamati M (2024) Computational intelligence techniques for achieving sustainable development goals in female cancer care. Discov Sustain 5(1):390
33. Salih AM, Galazzo IB, Raisi-Estabragh Z, Petersen SE, Menegaz G, Radeva P (2024) Characterizing the contribution of dependent features in xai methods. IEEE J Biomed Health Inform 28(11):6466–6473
34. Krell E, Mamalakis A, King SA, Tissot P, Ebert-Uphoff I (2025) The influence of correlated features on neural network attribution methods in geoscience. Environ Data Sci 4:29
35. United Nations - Department of Economic and social affairs. SDG Indicators Database. https://unstats.un.org/sdgs/dataportal/database. Accessed: 2025-02-20
36. Global Cancer Observatory - International Agency for Research on Cancer. https://gco.iarc.fr/today/en/dataviz/maps-most-common-sites?mode=cancer&key=total&types=1&cancers=15. Accessed: 2025-02-20
37. Serrano MÁ, Boguná M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106(16):6483–6488
38. Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9(1):1–12
39. Jacomy M, Venturini T, Heymann S, Bastian M (2014) Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLoS ONE 9(6):98679
40. Chen T (2015) Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4)
41. Fantom NJ, Serajuddin U (2016) The world bank's classification of countries by income. World Bank Policy Research Working Paper (7528)
42. Chahoud J, Semaan A, Rieber A (2016) Wealth, health expenditure, and cancer: a national perspective. J Natl Compr Cancer Netw 14(8):972–978
43. Global B, et al (2017) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol
44. Yang Z, Zheng R, Zhang S, Zeng H, Xia C, Li H, Wang L, Wang Y, Chen W (2017) Comparison of cancer incidence and mortality in three gdp per capita levels in China, 2013. Chin J Cancer Res 29(5):385
45. Torre LA, Siegel RL, Ward EM, Jemal A (2016) Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiol Biomark Prev 25(1):16–27
46. Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Stat Softw 36:1–13
47. Biau G, Scornet E (2016) A random forest guided tour. Test 25(2):197–227
48. Louhichi M, Nesmaoui R, Mbarek M, Lazaar M (2023) Shapley values for explaining the black box nature of machine learning model clustering. Proc Comput Sci 220:806–811
49. Conrad Z, Niles MT, Neher DA, Roy ED, Tichenor NE, Jahns L (2018) Relationship between food waste, diet quality, and environmental sustainability. PLoS ONE 13(4):0195405
50. Botnaru AA, Lupu A, Morariu PC, Pop OL, Nedelcu AH, Morariu BA, Cioancă O, Di Gioia ML, Lupu VV, Avasilcai L, et al (2025) Balancing health and sustainability: assessing the benefits of plant-based diets and the risk of pesticide residues. Nutrients 17(4):727
51. Stokes CS, Brace KD (1988) Agricultural chemical use and cancer mortality in selected rural counties in the usa. J Rural Stud 4(3):239–247
52. Verzeñassi D, Vallini A, Fernández F, Ferrazini L, Lasagna M, Sosa AJ, Hough GE (2023) Cancer incidence and death rates in Argentine rural towns surrounded by pesticide-treated agricultural land. Clin Epidemiol Global Health 20:101239
53. Peace L (1985) A time correlation between cigarette smoking and lung cancer. J R Stat Soc, Ser D, Stat 34(4):371–381
54. Strand T-E, Malayeri C, Eskonsipo P, Grimsrud TK, Norstein J, Grotmol T (2005) Teenage smoking and lung cancer incidence in early adult age, 1954-1998. Tidsskr Nor Laegeforen 125(9):1174–1176
55. Lee PN, Forey BA, Coombs KJ (2012) Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. BMC Cancer 12:1–90
56. Loeb LA, Emster VL, Warner KE, Abbotts J, Laszlo J (1984) Smoking and lung cancer: an overview. Cancer Res 44(12_Part_1):5940–5958

57. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 22(1):173–203
58. He H, He M-M, Wang H, Qiu W, Liu L, Long L, Shen Q, Zhang S, Qin S, Lu Z, et al (2023) In utero and childhood/adolescence exposure to tobacco smoke, genetic risk, and lung cancer incidence and mortality in adulthood. Am J Respir Crit Care Med 207(2):173–182
59. Garcia D, Matthews T (2024) Contaminated drinking water and its effect on cancer. ACS ES&T Water 4(8):3340–3347
60. Beresford SA (1983) Cancer incidence and reuse of drinking water. Am J Epidemiol 117(3):258–268
61. Wang Z, Gu W, Guo X, Xue F, Zhao J, Han W, Li H, Chen W, Hu Y, Yang C, et al (2023) Spatial association of surface water quality and human cancer in China. NPJ Clean Water 6(1):53
62. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6):394–424
63. Welch HG, Black WC (2010) Overdiagnosis in cancer. J Natl Cancer Inst 102(9):605–613
64. Research W (2018) Diet, nutrition, physical activity and cancer: a global perspective. World Cancer Research Fund/American Institute for Cancer Research
65. Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, et al (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet 389(10082):1907–1918
66. Murray CJ, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, Abd-Allah F, Abdelalim A, Abdollahi M, Abdollahpour I, et al (2020) Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet 396(10258):1223–1249
67. Umar MA, Nawaz A, Qayyum T (2023) Lrcmp: a sequential statistical framework for predicting cancer mortality rate. In: 2023 7th international multi-topic ICT conference (IMTIC). IEEE, pp 1–8
68. Omodei E, Kim D, Garcia-Herranz M, Sekara V (2023) Are machine learning, AI, and big data tools ready to be used for sustainable development? Challenges, and limitations of current approaches. Frontiers Media SA
69. Berti-Equille L, Dao D, Ermon S, Goswami B (2021) Challenges in kdd and ml for sustainable development. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 4031–4032
70. United Nations DESA (2007) Indicators of sustainable development: guidelines and methodologies. United Nations, New York
71. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
72. Newman ME (2008) The mathematics of networks. New Palgrave Encycl Econ 2(2008):1–12
73. Ergün S (2023) Explaining xgboost predictions with shap value: a comprehensive guide to interpreting decision tree-based models. New Trends Comput Sci 1(1):19–31
74. Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41:647–665

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.