Applications of Machine Learning Models to Predict and Prevent Obesity: A Mini-Review

Xiaobei Zhou [1,2], Lei Chen [1,2,3] and Hui-Xin Liu [1,2,3]*

[1] Health Sciences Institute, China Medical University, Shenyang, China, [2] Liaoning Key Laboratory of Obesity and Glucose/Lipid Associated Metabolic Diseases, China Medical University, Shenyang, China, [3] Institute of Life Sciences, China Medical University, Shenyang, China

Research on obesity and related diseases has received attention from government policymakers; interventions targeting nutrient intake, dietary patterns, and physical activity are deployed globally. An urgent issue now is how can we improve the efficiency of obesity research or obesity interventions. Currently, machine learning (ML) methods have been widely applied in obesity-related studies to detect obesity disease biomarkers or discover intervention strategies to optimize weight loss results. In addition, an open source of these algorithms is necessary to check the reproducibility of the research results. Furthermore, appropriate applications of these algorithms could greatly improve the efficiency of similar studies by other researchers. Here, we proposed a mini-review of several open-source ML algorithms, platforms, or related databases that are of particular interest or can be applied in the field of obesity research. We focus our topic on nutrition, environment and social factor, genetics or genomics, and microbiome-adopting ML algorithms.

Keywords: nutrition, obesity, machine learning, algorithm, genetics, environment

INTRODUCTION

Obesity is considered to be a chronic progressive disease caused by a combination of multiple determinants, including biological, genetic, social, environmental, and behavioral factors (1, 2). Research on obesity and related diseases has received attention from government policymakers; interventions targeting nutrient intake, dietary patterns, and physical activity are deployed globally. Considerable current effort has been made by the computer community and industry to apply artificial intelligence (AI) technology in the field of biology and biomedicine (3, 4); a possible solution is to use modified machine learning (ML) algorithms to detect obesity disease biomarkers or discover intervention strategies to optimize the weight loss results. ML is seen as part of AI, allowing software applications to predict outcomes without being interpretable (5). ML models can be summarized as two categories: (I) supervised learning models relying on labeled data to train a function that can finish the prediction tasks, and (II) unsupervised learning models focusing on summarizing the characteristics of data, such as dimensionality reduction analysis.

Machine learning models have been successfully used in many studies on obesity to predict obesity rates and identify the risk factors in samples of interest (6–9). In the field of ML research, it is important and necessary to share the code used in research to check the reproducibility of a related work. In addition, open-source tools could greatly improve the efficiency of similar studies by other researchers. However, these studies ignored one significant thing: They did not share

their ML methodologies and frameworks with the public for use by other researchers. At this stage, obesity-related research is considered a public health field and is mostly published in certain professional or medical journals, and these journals do not have strict requirements on the openness of the algorithms or frameworks used in the manuscripts, which are far less than the requirements for data sharing. After matching the keyword "github" (which is the world's most popular public code repository) in 491 abstract texts of obesity-ML-related research [using the PubMed search strategy "Obesity and Machine Learning and" ("1990/01/ 01" [PDAT]: "2022/04/01" [PDAT])], only one related paper was found. Open-source tools are tools for which the original source code is freely available and can be redistributed and modified (10). As a new and booming branch of computer science, ML inherits the tradition of open sharing in the computer community; relevant top journals or conferences have requirements for code sharing of published papers, and most researchers are also willing to upload their codes to the software source code repository.

Several existing reviews focused on the detailed illustration of the status (including the fundamentals, strength, limitation, and evaluation metric) of ML methods in obesity, while ignoring the discussion on the openness of algorithms (7, 11–13). To address this point, we proposed a mini-review for collecting open-source ML algorithms or platforms particularly focusing on the field of obesity research or that could be used in related problems in this article. Since the etiology and pathogenesis of obesity are extremely complex, we set the starting point of our study on the prior knowledge of the current authoritative study of obesity epidemiology (14). The traditional approach to epidemiological obesity research is to study numerous risk citations for obesity through specialized cohorts or epidemiological surveys (14). These customized ML tools for these specialized studies did not take our interest because of their low generality. We focused on ML algorithms related to obesity into several segments, including diet and nutrition, physical activity, geographic environment, genetics or genomics, and microbiome, in which the required data of ML models are available from the database (e.g., GEO database) or public platform (Google Map) without special input (**Figure 1**). In addition, we collected detailed information of 25 open-source ML algorithms or models applied in obesity or that could be used in related filed, including the project name, the relative website, applicable data types, and the simplified description of usage (**Table 1**).

METHODS

We designed a custom semi-automated data collection method incorporating data mining techniques for this mini-review. The PubMed and Google Scholar databases were searched from the inception until March 2022. The search terms used were: "obesity AND machine learning"; "nutrigenetics AND machine learning"; "food recommendation AND machine learning"; "obesity AND geographic information system"; "mhealth" AND ("smartphones" OR "mobile app" OR "mobile applications") AND ("platform" OR "framework")"; "(Genetics OR Genomics)

AND machine learning"; "microbiome AND machine learning." The inclusion criteria were as follows: (1) the abstract text contains the keyword "github" (text mining (TM) technology); (2) studies applying at least one ML algorithm focusing on predicting obesity and obesity-related diseases, preventing the obesity prevalence, or conducting the relationship between obesity and obesity risk factors [manually screening (MS)]; (3) the algorithms or frameworks in the studies must be suitable for a common data format, such as fasta or fastq format (MS). Articles excluded were the following: (1) the main text of studies was not available in the English language (TM technology); (2) review papers and papers not considered original studies (TM technology). The PRISMA flowchart of assessment of the literature is illustrated in **Figure 2**.
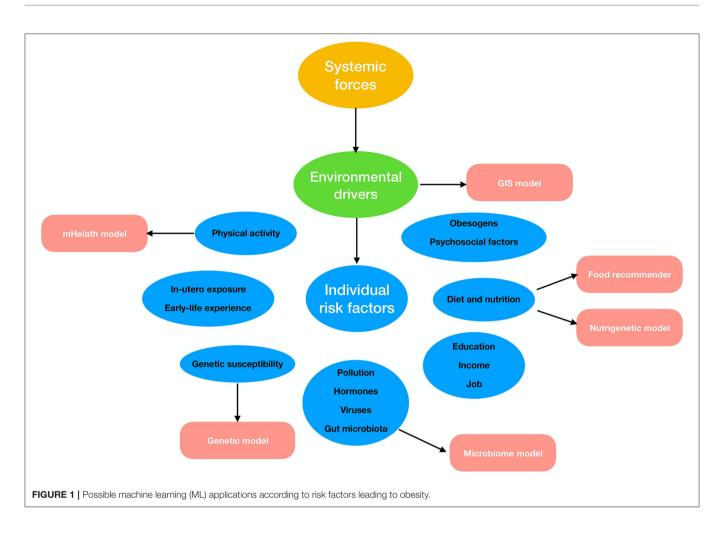
RESULTS

In summary, we included 25 open-source ML algorithms or models applied in obesity in this mini-review. In the following sections, we discuss each algorithm and describe its applications.

Personalized Food Recommender

A recommendation system can be considered a specified ML model predicting the "rating" or "preference" when given certain information about the user. Deep learning approaches based on the embedding models are now popularly used in modern recommendation systems (15). The core utility of the food recommendation system is to provide recipe retrieval (16). PFoodReq is a novel question-answering food recommendation system based on a large-scale food knowledge base/graph (17). In general, PFoodReq will follow the user's question, such as "What's good with bread for breakfast?", and then output all recipes from the model. The ingredients in these recipes are then rated for suitability and the top-rated recipes are recommended. FlavorGraph is a knowledge graph (KG) system by relations extracted from food recipes and information on flavor molecules from food databases (18). The two main usages of FlavorGraph are predicting compound food relationships and selecting or optimizing food pairings. Yum-me is a nutrition-based meal recommender that integrates state-of-the-art food image analysis models (19). Its input variables rely on two parts: (I) a survey of user dietary restrictions and nutritional expectations and (II) a visual interface that represents the user's food preferences. DeepFood, based on a deep learning approach, can recognize multi-item (food) images by detecting candidate regions or using a deep convolutional neural network (CNN) for food classification tasks (20). Market2Dish focused on the extension utility of user health profiling and health-aware food recommendation (21).

Nutrigenetic Model

Nutrigenetics is the science of studying the interaction between nutrition and gene information. LC-N2G is a novel statistical method inheriting genetic algorithm for ranking and identifying combinations of nutrients with gene expression (22). NutriGenomeDB is a nutrigenomics data platform based on the GSEA algorithm, which collects signature gene information of

FIGURE 1 | Possible machine learning (ML) applications according to risk factors leading to obesity.

nutrigenomics experimental expression data obtained from the gene expression omnibus (GEO) (23).

Geographic Information System (GIS) Model

Studying obesity-related issues through GIS data and methods is one of the hotspots in current public health research. A US team pioneered the use of deep learning techniques to assess obesity levels by identifying key patterns in the built environment from satellite imagery (24). The key idea is to extract thousands of hidden features from satellite images and identify the potential relationship between the hidden features and the body mass index (BMI). Unfortunately, the framework is not open to the public. MapMetadataEnrichment, a deep learning–based approach that can automatically generate labeled training map images using GIS data, is highly recommended as an alternative framework (25). Without a deep learning approach, spatial models could not be directly used to analyze GIS data to study obesity (26). In this article, we recommend the GWmodel, which is a widely used R-based package that implements geographically weighted models for exploring spatial heterogeneity (27). A classic use case of the application of the GWmodel can be found in a recent study in which a geographically weighted

regression (GWR) model (based on the GWmodel package) was adopted to analyze the relationship between socioeconomic factors, obesity, and air temperature and unhealthy behaviors in the USA (28).
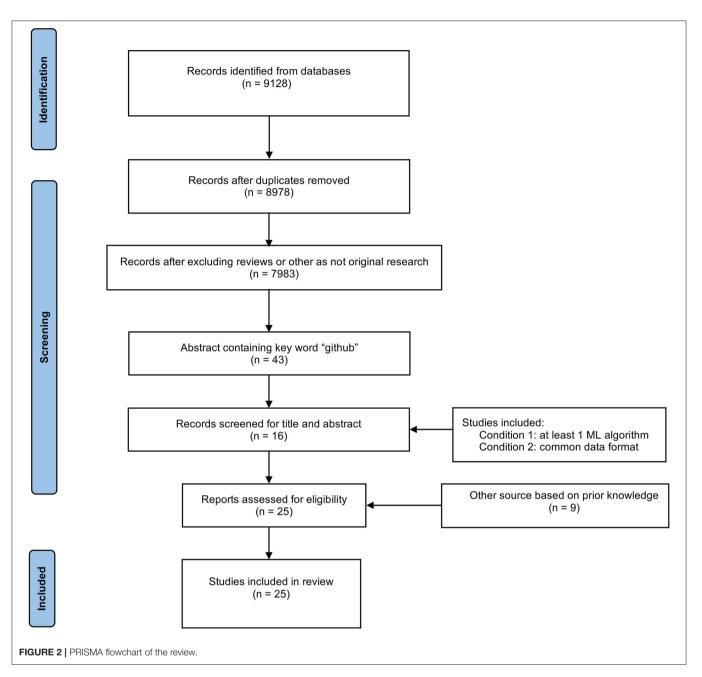
mHealth Platform

Recently, digital technologies that can monitor and manage our physical and mental health in our daily lives have rapidly developed and are being used to solve obesity-related problems (29, 30). Mobile health applications denoted mHealth apps have become increasingly popular with researchers and clinicians as effective tools for improving health behaviors. Several open-source mHealth platforms are available, including schema (31), mHealthDroid (32), and MobileCoach (33). These platforms are mentioned here since they can cooperate with ML models in the situation where the model uses the data provided by the platform. The mHealhDroid platform is designed to facilitate the fast and easy development of mHealth and biomedical applications; the schema platform is a lightweight cross-platform mobile application focused on mobile health monitoring and intervention research; the MobileCoach platform provides a one-stop solution for fully automated digital interventions.

TABLE 1 | Detailed information on these ML algorithms.

Name	Usage	Core algorithm	Input data	URL
PFoodReq	Food recommendation system	Knowledge graph (KG) & question answering (QA) & recipe retrieval	Text	https://github.com/hugochan/PFoodReq
FlavorGraph	Food recommendation system	KG & recipe retrieval & compound food relationship	Vector	https://github.com/lamypark/FlavorGraph
DeepFood	Food recommendation system	Deep learning (DL)& image recognition & recipe retrieval	Image	https://github.com/deercoder/DeepFood
Market2Dish	Food recommendation system	DL& image recognition & recipe retrieval	Image	https://github.com/WenjieWWJ/FoodRec
LC-N2G	Nutrigenetics analysis method	Statistical methods	Gene expression data (GSE85998)	https://sydneybiox.github.io/LCN2G/
NutriGenomeDB	Nutrigenetics platform	None	None	https://github.com/rmartin84/NutriGenomeDB (locally deployment)
MapMetadataEnrichment	GIS satellite image analysis tool	DL	Satellite image	https://github.com/geoai-lab/MapMetadataEnrichment
GWmodel	An R package for exploring spatial heterogeneity	Spatial regression models	POI data	https://cran.r-project.org/web/packages/GWmodel/index.html
Schema	Digital intervention	None	None	https://github.com/schema-app/schema
mHealthDroid	mHealth platform	None	None	https://github.com/mHealthTechnologies/mHealthDroid
MobileCoach	Digital intervention	None	None	https://github.com/schlegel11/MobileCoach
PGS Catalog	Polygenic score (PRS) database	None	None	https://www.pgscatalog.org/
Impute.me	Platform for direct-to-consumer genetic testing	none	23AndMe	https://github.com/lassefolkersen/impute-me
DeepVariant	Deep learning-based variant caller	DL	BAM or CRAM	https://github.com/google/deepvariant
NeuralCVD	Cardiovascular risk predictor	Survival machine algorithm	UK Biobank data	https://github.com/thbuerg/NeuralCVD
DeepCOMBI	AI tool for analysis of GWAS data	DL	GWAS data	https://github.com/AlexandreRozier/DeepCombi
DeepMicro	Taxonomic classifier	DL	Microbe data (csv)	https://github.com/minoh0201/DeepMicro
DeepMicrobes	Taxonomic classifier	DL	Microbe data (fasta)	https://github.com/MicrobeLab/DeepMicrobes
SortMeRNA	Taxonomic classifier	ML	Microbe data (fasta)	https://github.com/biocore/sortmerna
q2-feature-classifier	Taxonomic classifier	ML	Microbe data (fasta)	https://github.com/qiime2/q2-feature-classifier
Swarm	Taxonomic classifier	ML	Microbe data (fasta)	https://github.com/torognes/swarm
GEDFN	Microbial biomarker' identification	DL	OTU & IBD	https://github.com/MicroAVA/GEDFN
MDeep	Microbe-disease predictor	DL	OTU	https://github.com/lichen-lab/mdeep
TaxoNN	Microbe-disease predictor	DL	OTU	https://github.com/divya031090/taxonn_otu.
MetaPheno	Microbe-disease predictor	DL	Microbe data (fasta)	https://github.com/nlapier2/metapheno

Genetic Model

Genetics studies genes and the way certain traits or conditionsare passed from one generation to the next. The polygenic score (PGS) database collects published PGS information that provides the community with an open platform for PGS research. (34). PRS estimates an individual's genetic risk for complex diseases based on many genetic variants across the genome. In fact, a polygenic risk score (PRS) could be considered a specified regression model; there are many obesity PRS models available in the PGS database, such as PGP000017 and PGP000211. Impute.me is the first non-commercial platform for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores (35). DeepVariant was launched by Google, which uses deep neural networks to fast and accurately

FIGURE 2 | PRISMA flowchart of the review.

identify variation sites from DNA sequencing data (36). The NeuralCVD-based deep survival machine algorithm can estimate the cardiovascular risk for coronary heart disease prevention (37). DeepCOMBI uses AI for analysis and discovery in genome-wide association studies (38).

Microbiome Model

The gut microbiome is closely related to overall health. More and more studies suggest that obesity is associated with specific changes in the composition and function of the human gut microbiome (39, 40). Any study of the gut microbiome associated with obesity cannot skip the step of taxonomic classification to infer the relative abundance of different taxa. Typical ML taxonomic classifiers are DeepMicro (41), DeepMicrobes (42), SortMeRNA (43), q2-feature-classifier (44), swarm (45), etc.

We mentioned that q2 feature classifier has been implemented in QIIME 2 which is the most popular microbiome analysis platform (46). ML models [GEDFN (47), MDeep (48), TaxoNN (49), and MetaPheno (50)] can also be used to predict patient phenotype or obesity from their microbiome sequence data. MIPMLP provides a reproducible preprocessing ML pipeline for a microbiome analysis (51).

LIMITATIONS

Obesity, with its complex etiologies, is difficult to describe with a single theoretical model. The current trend in obesity research is to integrate health and medical big data and use high-throughput sequencing technology for multi-omics joint research. In general,

the current ML models and algorithms for obesity research are separated from a single field, and no model can span multiple fields or integrate multiple types of data across platforms for a joint analysis of homologous and heterogeneous data. For the nutrigenetic model, the development of open-source models is far behind the industry. Several startups have launched genetics-based food recipes and recommendation systems, while their models are "black box." A major limitation of genetic models of obesity is weak evidence from small sample sizes due to small variants of genetic mutations. For the mHealth platforms, we could not find open-source ML models in obesity, although we believe that ML models in obesity must be existing in commercial mobile APPs. For the open-source mHealth platform, we were unable to find an open-source ML model related to obesity, although we believe that such ML models already exist in commercial mobile apps.

PERSPECTIVE

Machine learning algorithms are a powerful analytic tool that enables us to conceptualize and study metabolic disorders within a fundamentally novel framework. In contrast to conventional statistical methods assessing single modality predictors, ML methods are capable of integrating multiple data types and sources to inform predictive models. Nonetheless, ML algorithms are limited by the type of data captured, the quality of available data, the conceptual frameworks algorithms are applied to, and the underlying assumptions. The use of ML algorithms in obesity will dramatically increase in the future, and a large number of ML algorithms will be implemented on a few specific platforms such as QIMME 2 to facilitate rapid application and collaborative work between multiple algorithms. Presently, most of the ML algorithms used in obesity-related research are isolated in a single field or the factor that causes obesity; in the future, the algorithms will work together across platforms or data types. Future research vistas are to optimize and prospectively test predictive models using external datasets.

AUTHOR CONTRIBUTIONS

FUNDING

REFERENCES

1. Apovian CM. Obesity: definition, comorbidities, causes, and burden. *Am J Manag Care*. (2016) 22(7 Suppl.):s176–85.

2. Kyle TK, Dhurandhar EJ, Allison DB. Regarding obesity as a disease: evolving policies and their implications. *Endocrinol Metab Clin North Am*. (2016) 45:511–20. doi: 10.1016/j.ecl.2016.04.004

3. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. (2018) 15:20170387. doi: 10.1098/rsif.2017.0387

4. Zhao Z, Woloszynek S, Agbavor F, Mell JC, Sokhansanj BA, Rosen GL. Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. *PLoS Comput Biol*. (2021) 17:e1009345. doi: 10.1371/journal.pcbi.1009345

5. Mitchell TM. *Machine Learning*. 1st ed. McGraw-Hill, Inc. (1997).

6. Chatterjee A, Gerdes MW, Martinez SG. Identification of risk factors associated with obesity and overweight-a machine learning overview. *Sensors*. (2020) 20:2734. doi: 10.3390/s20092734

7. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients*. (2020) 12:2466. doi: 10.3390/nu12082466

8. Duran I, Martakis K, Rehberg M, Semler O, Schoenau E. Diagnostic performance of an artificial neural network to predict excess body fat in children. *Pediatr Obes*. (2019) 14:e12494. doi: 10.1111/ijpo.12494

9. Safaei M, Sundararajan EA, Driss M, Boulila W, Shapi'i A. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Comput Biol Med*. (2021) 136:104754. doi: 10.1016/j.compbiomed.2021.104754

10. Corbly JE. The free software alternative: freeware, open-source software, and libraries. *Inform Technol Librar*. (2014) 33:65–75. doi: 10.6017/ital.v33i3.5105

11. Chimunhu P, Topal E, Ajak AD, Asad W. A review of machine learning applications for underground mine planning and scheduling. *Resour Policy*. (2022) 77:102693. doi: 10.1016/j.resourpol.2022.102693

12. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity. *Obes Rev*. (2018) 19:668–85. doi: 10.1111/obr.12667

13. Triantafyllidis A, Polychronidou E, Alexiadis A, Rocha CL, Oliveira DN, da Silva AS, et al., Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature. *Artif Intell Med*. (2020) 104:101844. doi: 10.1016/j.artmed.2020.101844

14. Pan XF, Wang L, Pan A. Epidemiology and determinants of obesity in China. *Lancet Diabetes Endocrinol*. (2021) 9:373–92. doi: 10.1016/S2213-8587(21)00045-0

15. Zhang S, Yao L, Sun A, Tay Y. Deep learning based recommender system: a survey and new perspectives. *ACM Comput Surv*. (2017) 52:1–35. doi: 10.1145/3285029

16. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health recommender systems: systematic review. *J Med Internet Res*. (2021) 23:e18035. doi: 10.2196/18035

17. Chen Y, Subburathinam A, Chen CH, Zaki MJ. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In: *WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining*. New York, NY: Association for Computing Machinery. (2021). doi: 10.1145/3437963.3441816

18. Park D, Kim K, Kim S, Spranger M, Kang J. FlavorGraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. *Sci Rep*. (2021) 11:931. doi: 10.1038/s41598-020-79422-8

19. Yang L, Hsieh C-K, Yang H, Pollak JP, Dell N, Belongie S, et al. Yum-Me: a personalized nutrient-based meal recommender system. *ACM Trans Inf Syst*. (2017) 36:7. doi: 10.1145/3072614

20. Jiang L, Qiu B, Liu X, Huang C, Lin K. DeepFood: food image analysis and dietary assessment via deep model. *IEEE Access*. (2020) 1–1. doi: 10.1109/ACCESS.2020.2973625

21. Wang W, Duan LY, Jiang H, Jing P, Song X, Nie L. Market2Dish: health-aware food recommendation. *ACM Trans Multimedia Comput Commun Applic*. (2021) 17:1–19. doi: 10.1145/3418211

22. Xu X, Solon-Biet SM, Senior A, Simpson SJ, Fontana L, Mueller S, et al. LC-N2G: a local consistency approach for nutrigenomics data analysis. *BMC Bioinformatics*. (2020) 21:530. doi: 10.1186/s12859-020-03861-3

23. Martín-Hernández R, Reglero G, Ordovás JM, Dávalos A. NutriGenomeDB: a nutrigenomics exploratory and analytical platform. *Database*. (2019) 2019:baz097. doi: 10.1093/database/baz097

24. Maharana A, Nsoesie EO. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw Open*. (2018) 1:e181535. doi: 10.1001/jamanetworkopen.2018.1535

25. Hu Y, Gui Z, Wang J, Li M. Enriching the metadata of map images: a deep learning approach with GIS-based data augmentation. *Int J Geogr Inform Sci*. (2021) 36:1–23. doi: 10.1080/13658816.2021.1968407

26. Kamel Boulos MN, Peng G, VoPham T. An overview of GeoAI applications in health and healthcare. *Int J Health Geogr*. (2019) 18:7. doi: 10.1186/s12942-019-0171-2

27. Cao S, Zheng H. A POI-based machine learning method for predicting residents' health status. In: *The International Conference on Computational Design and Robotic Fabrication*. Singapore: Springer. (2021). doi: 10.1007/978-981-16-5983-6_13

28. Lotfata A. Using geographically weighted models to explore obesity prevalence association with air temperature, socioeconomic factors, and unhealthy behavior in the USA. *J Geovisualiz Spatial Anal*. (2022) 6:1–12. doi: 10.1007/s41651-022-00108-y

29. Goldstein SP, Thomas JG, Foster GD, Turner-McGrievy G, Butryn ML, Herbert JD, et al. Refining an algorithm-powered just-in-time adaptive weight control intervention: a randomized controlled trial evaluating model performance and behavioral outcomes. *Health Informatics J*. (2020) 26:2315–31. doi: 10.1177/1460458220902330

30. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. *JMIR Diabetes*. (2017) 2:e28. doi: 10.2196/diabetes.8590

31. Shatte ABR, Teague SJ. Schema: an open-source, distributed mobile platform for deploying mhealth research tools and interventions. *BMC Med Res Methodol*. (2020) 20:91. doi: 10.1186/s12874-020-00973-5

32. Banos O, Garcia R, Holgado-Terriza JA, Damas M, Pomares H, Rojas I, et al. *mHealthDroid: A Novel Framework for Agile Development of Mobile Health Applications*. Cham: Springer International Publishing (2014). doi: 10.1007/978-3-319-13105-4_14

33. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. Design and evaluation of a mobile Chat App for the open source behavioral health intervention platform MobileCoach. In: *International Conference on Design Science Research in Information System and Technology*. Cham: Springer. (2017). doi: 10.1007/978-3-319-59144-5_36

34. Lambert SA, Gill L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*. (2021) 53:420–5. doi: 10.1038/s41588-021-00783-5

35. Folkersen L, Pain O, Ingason A, Werge T, Lewis CM, Austin J. Impute.me: an open-source, non-profit tool for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores. *Front Genet*. (2020) 11:578. doi: 10.3389/fgene.2020.00578

36. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*. (2021) 36:5582–9. doi: 10.1093/bioinformatics/btaa1081

37. Steinfeldt J, Buergel T, Loock L, Kittner P, Ruyoga G, Belzen JU, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. *Lancet Digit Heal*. (2022) 4:e84–94. doi: 10.1016/S2589-7500(21)00249-1

38. Mieth B, Rozier A, Rodriguez JA, Höhne MMC, Görnitz N, Müller KR. DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genom Bioinform*. (2021) 3:lqab065. doi: 10.1093/nargab/lqab065

39. Aoun A, Darwish F, Hamod N. The influence of the gut microbiome on obesity in adults and the role of probiotics, prebiotics, and synbiotics for weight loss. *Prev Nutr Food Sci*. (2020) 25:113–23. doi: 10.3746/pnf.2020.25.2.113

40. Davis CD. The gut microbiome and its role in obesity. *Nutr Today*. (2016) 51:167–74. doi: 10.1097/NT.0000000000000167

41. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci Rep*. (2020) 10:6026. doi: 10.1038/s41598-020-63159-5

42. Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom Bioinform*. (2020) 2:lqaa009. doi: 10.1093/nargab/lqaa009

43. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*. (2012) 28:3211–7. doi: 10.1093/bioinformatics/bts611

44. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*. (2018) 6:90. doi: 10.1186/s40168-018-0470-z

45. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*. (2014) 2:e593. doi: 10.7717/peerj.593

46. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. (2019) 37:852–57. doi: 10.1038/s41587-019-0209-9

47. Zhu Q, Jiang X, Zhu Q, Pan M, He T. Graph Embedding deep learning guides microbial biomarkers' identification. *Front Genet*. (2019) 10:1182. doi: 10.3389/fgene.2019.01182

48. Wang Y, Bhattacharya T, Jiang Y, Qin X, Wang Y, Liu Y, et al. A novel deep learning method for predictive modeling of microbiome data. *Briefings Bioinformatics*. (2020) 22:bbaa073. doi: 10.1093/bib/bbaa073

49. Sharma D, Paterson AD, Xu W. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics*. (2020) 36:4544–50. doi: 10.1093/bioinformatics/btaa542

50. LaPierre N, Ju CJT, Zhou G, Wang W. MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods*. (2019) 166:74–82. doi: 10.1016/j.ymeth.2019.03.003

51. Jasner Y, Belogolovski A, Ben-Itzhak M, Koren O, Louzoun Y. Microbiome preprocessing machine learning pipeline. *Front Immunol*. (2021) 12:677870. doi: 10.3389/fimmu.2021.677870