Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
2(1): 32–48
DOI: 10.3934/aci.2022002
Received: 24 January 2022
Revised: 30 March 2022
Accepted: 31 March 2022
Published: 31 March 2022

Review

A review of the application of machine learning in adult obesity studies

Mohammad Alkhalaf[1,2], **Ping Yu**[1,3,*], **Jun Shen**[1] **and Chao Deng**[3,4]

[1] School of Computing and Information Technology, University of Wollongong, Wollongong, Australia
[2] School of Computer Science, Qassim University, Qassim, Saudi Arabia
[3] Illawarra Health and Medical Research Institute, Wollongong, Australia
[4] School of Medicine, University of Wollongong, Wollongong, Australia
* **Correspondence**: ping@uow.edu.au

Academic Editor: Chih-Cheng Hung

Abstract: In obesity studies, several researchers have been applying machine learning tools to identify factors affecting human body weight. However, a proper review of strength, limitations and evaluation metrics of machine learning algorithms in obesity is lacking. This study reviews the status of application of machine learning algorithms in obesity studies and to identify strength and weaknesses of these methods. A scoping review of paper focusing on obesity was conducted. PubMed and Scopus databases were searched for the application of machine learning in obesity using different keywords. Only English papers in adult obesity between 2014 and 2019 were included. Also, only papers that focused on controllable factors (e.g., nutrition intake, dietary pattern and/or physical activity) were reviewed in depth. Papers on genetic or childhood obesity were excluded. Twenty reviewed papers used machine learning algorithms to identify the relationship between the contributing factors and obesity. Regression algorithms were widely applied. Other algorithms such as neural network, random forest and deep learning were less exploited. Limitations regarding data priori assumptions, overfitting and hyperparameter optimization were discussed. Performance metrics and validation techniques were identified. Machine learning applications are positively impacting obesity research. The nature and objective of a study and available data are key factors to consider in selecting the appropriate algorithms. The future research direction is to further explore and take advantage of the modern methods, i.e., neural network and deep learning, in obesity studies.

Keywords: machine learning; prediction; regression; neural network; deep learning; obesity

1. Introduction

Obesity is increasing at an alarming rate. The prevalence of obesity has tripled in the last four decades [1]. The body mass index (BMI) is the most commonly used indicator to assess obesity that, if BMI is 30 kg/m^2 or larger, it falls into the obese range [1]. Obesity is a major cause for several chronic diseases [2] such as cardiovascular disease and diabetes, which are the leading causes of death around the world [3]. Obesity causes around three million death per year [4]. It has also led to substantial financial burden in obesity treatment. In the U.S., treatment of obesity and its related illnesses costs more than 200 billion dollars per year [5]. Despite the severe impact of obesity on human life, there is limited effective prevention mechanisms to control obesity at population levels, although some therapeutic methods such as bariatric surgery have been successfully conducted on some obese individuals [6]. On the other hand, interventions in nutrition intake, dietary pattern and/or physical activity have been promoted worldwide with certain level of success [7]. Now the key issue is how we can improve the effectiveness of these interventions. A promising approach is to apply artificial intelligence (AI) and machine learning (ML) technology to discover the effective intervention strategies for optimal outcomes.

Machine learning is a data analysis method that applies computer systems to execute tasks without being explicitly programmed [8]. It generally takes dataset as input, learns from data and then outputs predictions with minimal human intervention. Machine learning methods can be classified as supervised learning and unsupervised learning. Supervised learning uses labeled data and tries to predict the outcomes from the input variables (i.e., predicting body weight from nutrition intake data). Unsupervised learning uses unlabeled data and aims to find hidden relationships between variables (i.e., clustering different eating behaviors from dataset). Both methods are fast and efficient, thus have been widely applied in many fields such as healthcare [9], finance [10] and autonomous cars [11]. The purpose of using machine learning is diverse, including but not limited to, extract useful information from data, recognize hidden patterns, acquire knowledge, predict the future and make recommendations. Applications of machine learning have enriched traditional data analytic methods in various fields including health [12].

Rapid development in computer technology and computational capabilities in the last decade has led to massive improvement in accuracy and running time of ML. Recent studies have suggested that the performance of ML techniques in data analytics can outperform that of the traditional statistical methods [13]. It is predicted that ML will increasingly play vital role in health research and the application of ML in analyzing medical data will become increasingly crucial [14]. For example, biomedical researchers have successfully applied ML to describe and/or predict factors that cause a certain diseases, improve quality of clinical decision and reduce medical errors in treating diseases such as cancer [15], diabetes [16], cardiovascular diseases [17] and finding relationship between comorbideties [18]. These studies have demonstrated the effectiveness of ML methods in discovering and predicting the relationships between independent variables and dependent variables. Other studies even suggested that ML will soon exceed human specialists in diagnostic accuracy [19].

The amount of healthcare data generated through smart devices is enormous, in addition to the huge amount of patient data captured in hospital electronic medical records [12]. Increasing availability of digital data due to increased computerized data capture has also accelerated the advancement of ML technology and generated needs for researchers to build more robust prediction models to support data-driven health care [20].

Many ML algorithms have been applied in healthcare such as regression, classification, clustering, and neural network algorithms.

Logistic regression: is used to test a predefined hypothesis and find a relationship between input and output variables when the output variables are categorical in nature (i.e., weight gain or loss) [21]. Linear regression is similar to logistic regression in terms of examining the association between input and output variables. Its output is continuous, not binary variable (i.e., weight changes in kilograms). It also assumes a linear relationship between input and output variables [22].

Classification: is widely applied data-based technique [23]. It classifies data into predefined classes. Algorithms such as decision trees, random forest, naïve Bayes and support vector machine (SVM) are all examples of classification algorithms.

Decision tree: is an algorithm that uses functions to classify data in a shape that is similar to a tree structure. It classifies data by sorting them from root down to leaf node. Each node in the tree represents a variable and each branch from a node represent a possible value of the attribute. It is applied to classify samples to specific classes based on their values. These classes are divided based on specific calculated thresholds. Decision tree is simple, easy to apply, uses both categorical and numerical data and produce promising results [22,24].

Random forest: is basically a large number of decision trees, could be a couple of hundreds that would function in aggregate to improve the effectiveness of a prediction model [25].

Naïve Bayes: is also a simple classifier to calculate event probability. It needs a small number of data points to find relationship between the probabilities and conditional probabilities of two events. It assumes that existence of features of a class are independent of each other, and the input data is normally distributed. It is fast, scalable and effective in handling missing data [22].

Support vector machine: is another well-established and robust classification technique. Essentially it works as a hyperplane and divides the positive and negative classes in supervised learning to separate the cases of the target variables. It divides cases of two categories on two sides and tries to reach maximum margin between the two sides [26]. If dataset contains categorical values then the use of methods such as one-hot-encoding is needed to transform them into binary values [27].

Extreme gradient boosting (XGBoost): is an improvement to already existing ed gradient tree boosting algorithm with increased speed and scalability than many other ML algorithms. It is widely applied these days in ML competitions for its effective performance [28].

Neural network: works like the human brain where there is a web of connected neurons. It basically comprises of number of layers; input layer, hidden layer, and output layer; with number of neurons in each layer. Each neuron in neural network algorithm takes input values and computes its weight (strength between neurons). It then applies activation function to produce a single output value. Neural network is used to predict both continuous and categorical data. They can be effective in models where the relationship between input and output variables is non-linear [29].

Deep learning: is becoming more popular nowadays in text analyses and image recognition [30]. It is essentially a neural network with many hidden layers where each layer uses the output of previous layer as its input. In general, deep learning algorithm takes inputs data with their labels and passes them through multiple hidden layers to extract features and generate classification. These hidden layers are often called black boxes since the user does not see the work of each layer.

Clustering: is used for analyzing unsupervised data in order to group similar objects [14]. It does not require predetermined hypothesis; instead, it produces one by uncovering the causal relationships between the data assembled together with certain features [31], thus can serve the purpose of medical research.

Despite the potential importance of ML application in obesity studies, to the best of our knowledge, very few studies have paid attention to what ML algorithms have been applied in this field and their effectiveness. Therefore, this study aims to understand the current state of application of ML in obesity studies. It will discuss how ML technique is applied in obesity studies, what are the contributions and limitations, how are algorithms performance measured and evaluated, what is the type of data used, and provide recommendations for the further application of ML in obesity research.

2. Method of conducting the review

A Scoping review based on [32] was conducted to address this research topic.

Keywords: used in literature search include *obesity*, *physical activity, nutrition, machine learning, regression*, *support vector machine, decision tree, neural network*, *deep learning, cluster analysis*. These keywords are used in different combinations for literature search. Information about the use of keywords and their various combinations is presented in Appendix (Table 4). Two online databases, PubMed and Scopus, were selected to search for articles published between 2014 and 2019. The databases search was undertaken up to 1st of January 2020.

Inclusion criteria: we included papers that have applied at least one ML algorithm to predict obesity, to study/prevent obesity prevalence or to describe the relationships between obesity and factors that are controllable by human actions (i.e., nutrition intake, dietary pattern and/or physical activity). The study subjects are adult people (i.e., aged above 18 years). 10 algorithms have been selected for this review: logistic and linear regression, decision trees, SVM, naïve Bayes, random forest, neural network, deep learning, XGBoost, and clustering.

Exclusion criteria: we excluded papers that focused on bariatric surgeries, laboratory data and papers that only reported the relationships between obesity and uncontrollable factors such as genetic or childhood obesity. We also excluded papers published in language other than English. We chose 2 studies for each algorithm and review them extensively. Other studies that used the same machine learning were excluded because of repetition in algorithm.

Assessment of literature: keyword search returned 6087 papers. We first scanned the title to assess whether the article met the inclusion and exclusion criteria. 440 papers were duplicates and 837 did not meet the criteria. Next, we scanned the abstract for most relevant titles. We applied the ‘stop search’ logic in literature selection. When we find two papers that meet the inclusion criteria, and apply the same machine learning method, we included them. Any further papers reporting the same machine learning method were excluded.

In summary, we included 20 papers in this review. Constant comparison was conducted among the included studies to extract the content of the studies and synthesise the key information into the report.

3. Results

Twenty papers were identified to apply different ML algorithm to study obesity. In terms of topics, four papers focused on individual’s nutrition intake to understand obesity [33–36]. Three focused on factors for weight loss and fat prediction [37–39]. Three focused on obesity and obese individual status [40–42]. Two focused on demographic variables to understand obesity [43,44]. Two focused only on physical activities [45,46]. Two focused on predicting obesity based on eating behaviors [47,48]. Another two tried to help fight obesity by focusing on calorie calculation from food pictures [49,50]. One paper focused on calculating energy expenditure [51] and another focused on obesity prevalence [52]. Data type of papers include clinical data [40,43–46,48,51] and cross-sectional questionnaire survey data [33–39,41,42,47,52] (Table 1).

Data of six papers were from United States [38,39,42,47,48,52]. Five are from Europe [37,41,43,45,51], two from Australia [44,46] and one from each of Canada [33], Malaysia [34], Mexico [35] and Chile [40]. Mixed data from over 70 countries was used in one paper [36]. Two papers used collection of food images from images dataset [49,50] (Table 1).

In general, many ML techniques have been applied in obesity prediction. In this review, 11 papers applied only one algorithm [33–35,39,41,43,48-52] while 9 papers applied more than one algorithm [36–38,40,42,44–47]. Eight of the nine papers applied and compared performances between multiple algorithms [36–38,40,42,44–46].

Accuracy, AUC, sensitivity, specificity and precision were the main performance metrics for models. K-Fold cross-validation, bootstrapping, data splitting were the most used model validation techniques (Table 2).

In the following sections, we will discuss each algorithm, describe its applications and provide examples of papers applied it.

Table 1. Overview of machine learning application in obesity studies.

Authors	Paper Focus	Important features	No. Data Records	Study type
[52]	Obesity prevalence	Tweets about food and physical activities + Google search trends about food	More than 4 million tweets	Cross-sectional study
[49]	Calorie count (food images)	Image color, texture, size and shape	3500 images	-
[50]	Calorie count (food images)	Food images from ImageNet dataset	1316 images	-
[38]	Numerous factor/body fat prediction	BMI, WC, Socioeconomic (education levels), demographic (race, gender and age)	25367	Cross-sectional study
[35]	Nutrition Intake	BMI, gender, food groups (cereals and grains, vegetables, fresh fruits, dairy, meat, fish and eggs, sugars and fats, and fast food)	18385	Cross-sectional study
[47]	Eating behavior	BMI, age, gender, race, frequency of eating (healthy food, unhealthy food, breakfast, snacking), overall diet quality and problem eating behaviors	9977	Cross-sectional study
[33]	Nutrition intake	BMI, food group, serving size, age, sex, marital status, race, employment status, student status, education, personal income, province of residence, living in urban/rural area, physical activity	6202	Cross-sectional study
[42]	Obesity status	historical hospitalization records	4787	Cross-sectional study

Continued on next page

[37]	Numerous factors	BMI, age, sex, physical activity, food frequency, education, smoking status	4757	Cross-sectional study
[41]	Obesity status	BMI, age, gender, race, health status, smoking, alcohol, physical activity	4144	Cross-sectional survey
[39]	Numerous factors	BMI, WC, age, race, education, employment, sex, food intake, types of physical and sedentary activity, television and video viewing, and computer use	4100	Cross-sectional study
[40]	Obesity Status	Electronic medical record	3015	Cohort study
[43]	Demographic variables/ body fat prediction	BMI, age, gender, body fat percentage	2755	Clinical study
[51]	Energy expenditure	BMI, weight, height, sex, age,	565	Clinical study
[46]	Physical activity	BMI, WC, daily steps, body fat, age, gender, cholesterol levels.	295	Clinical study
[34]	Nutrition intake (grocery sale)	BMI, physical activity, age, gender, calories intake, carbohydrate, fat, protein, raw and processed food	170	Cross-sectional study
[48]	Eating behavior	BMI, age, gender, food intake	120	Clinical study
[36]	Nutrition intake (grocery sale)	Nationwide food sale data (79 countries)	79	Cross-national study
[44]	Demographic variables and weight loss	BMI, sex and age	76	Clinical study
[45]	Physical activity	BMI, age, multi-sensor system for physical activities (sedentary, household, moderate, vigorous)	17	Clinical study

BMI: Body Mass Index; **WC**: waist circumference

Table 2. Algorithms applied, metrics evaluation and validation techniques.

Authors	Algorithm applied	Main performance metrics of models/variables	Model validation
[52]	LR	R^2	K-fold Cross-validation
[49]	SVM	75% to 99% Accuracy	Groups of train and test images
[50]	Deep CNN	0.95% Accuracy	Divided dataset for training and testing

Continued on next page

[38]	DL ANN DT LRs	Results of men body fat DL: AUC of 0.95 ANN: AUC of 0.95 DT: AUC of 0.92 LRs: AUC of 0.89	Leave-one-out Cross-validation
[35]	NB	Sensitivity	Bootstrapping
[47]	K-means LR	10000 repetitions of k-means	Split data into 2 samples
[33]	LR	CI	Bootstrap
[42]	XGB RF	XGB: AUC of 0.68 RF: AUC of 0.69	K-fold Cross-validation
[37]	RF LR	RF: OOB error estimate 41% for men data, 37% for women data LRs: OOB error estimate 39% for men data, 35% for women data	OOB
[41]	agglomerative hierarchical clustering	Replication analysis	Randomly divide samples into half
[39]	LRs	OR CI	-
[40]	NB SVM	NB: 91.44% Average Accuracy SVM: 96.99 Average Accuracy	K-fold Cross-validation
[43]	ANN	80.43% Predictive Accuracy	Randomly divided into training and testing
[51]	ANN	73% Predictive Precision	K fold Cross-validation
[46]	DT RF LRs SVM ANN	DT: AUC of 0.70 RF: AUC of 0.75 SVM: AUC of 0.69 LRs: AUC of 0.67 ANN: AUC of 0.66	70%,15%,15% training, validation and testing
[34]	DT	89% Accuracy	K-fold Cross-validation
[48]	SVM	82% Accuracy	K-fold Cross-validation
[36]	XGB RF SVM	XGB: RMSE: 0.05 RF: RMSE: 0.057 SVM: RMSE: 0.06	Leave-one-out Cross-validation
[44]	LRs DT	DT: AUC of 0.72 Better than LRs	K-fold Cross-validation

Continued on next page

[45]	RF SVM DT	RF: 94% Accuracy SVM: 84% Accuracy DT: 93% Accuracy	10 times randomly selecting of 90% training 10% testing

Table 3. Abbreviations.

LRs	Logistic Regression	**LR**	Linear Regression
DT	Decision Trees	**RF**	Random Forest
ANN	Artificial Neural Network	**XGB**	Extreme Gradient Boosting
SVM	Support Vector Machine	**NB**	Na ïve Bayes
CNN	Convolutional Neural Network	**DL**	Deep Learning
AUC	Area Under the Curve	**OR**	Odds ratio
CI	Confidence interval	**RMSE**	Root mean square error
OOB	Out-of-bag		

3.1. Logistic and linear regression

Logistic regression is used in obesity to predict future trends and prevalence or to classify risk associated with obesity. It is used to validate the relationships between the independent variables, i.e., nutrition intake, night-time eating and the binary dependent variable obesity (yes/no). Batterham et al. [44] used logistic regression, among other algorithms, to predict weight changes at the first month and at the end of a one-year dietary intervention. The algorithm had a moderate AUC in comparison to decision trees. Authors stated that the algorithm assumed linearity which was not the case in the data sample they used. Kim et al. [39] applied logistic regression to investigate the effect of food intake and physical activity on obesity among U.S. adults. Odds ratio was used to quantify the relationship between variables and outcome. Although the model was successful, using cross-sectional data prevented a causal inference of the findings.

Linear regression is similar to logistic regression in terms of its applications in obesity studies. So et al. [33] used linear regression to examine the relationship between obesity and consumption of four different food groups based on Canadian food guide. The model calculated each variable coefficient and tested the statistical significance of each variable. They used methods to validate reported BMI and energy intake measures which add strength to their model. However, it is still self-reported data which might include inaccurate measurements. Another example of the application of linear regression is to predict obesity prevalence based on data from Twitter and Google search results [52]. Authors built a model to study obesity prevalence in United States based on millions of tweets that include keywords of food and physical activities. Their model had R^2 between 0.83 and 0.79. The authors believed that this result will encourage governments to utilize ML on social-media data to have real-time understanding of obesity prevalence.

3.2. Decision tree

Decision tree is used in obesity research for dietary pattern prediction, diagnosis and risk analysis [38]. Batterham et al. [46] applied a decision tree algorithm, among other algorithms, to detect factors help in prediction whether an individual would adhere to daily physical activity goal of 10,000 steps. The algorithm was able to detect number of predictors with an AUC of 0.70. The authors stated that overfitting was clear limitation in decision tree analysis and lead to some inexplicable rules. Daud et al. [34] used decision trees to predict obesity using grocery data. For every household, grocery shopping data over five months was converted into nutrition intake. Although the model was able to predict obesity with 89 % accuracy, it could not be applied on individual level.

3.3. Random forest

Applications of random forest in obesity include obesity prediction [36], physical activity recognition [45] and nutrition intervention [46]. Feng et al. [45] used random forest for physical activity recognition. The model was able to recognize 19 different types of physical activities with 93.4% accuracy but the few number of subjects might limit the generalizability of the model. Kanerva et al. [37] used random forest to examine factors that affect bodyweight such as lifestyle and sociodemographic factors. The model had an estimated error rate of 40%. Authors stated that algorithm was able to handle highly correlated variables. However, the low number of these correlated variables used in the paper affected the model accuracy. Another issue stated is the difficulty to interpret the algorithm classification process.

3.4. Naïve Bayes

Naïve Bayes is proven to be effective in many practical applications such as predicting dietary pattern [35] and obesity risk factors. Easton et al. [35] built a predictive model based on naïve Bayes to predict health status based on dietary pattern. It measured the differences in eating behaviors between Mexican adults with and without obesity. The model had higher sensitivity than the average and was successful in subcategorizing participants based on health status, but limitation of data could have affected the ability to uniquely interpret the model. Figueroa and Flores [40] applied naïve Bayes and SVM to identify obesity status (i.e., morbid obesity, severe obesity) using electronic medical records. They stated that applying feature selection technique produced a good naïve Bayes model with an average accuracy of 91%.

3.5. Support vector machine

Utilization of SVM in obesity studies includes physical activity recognition, obesity status [40] and food image recognition. Sarasfis et al. [48] used SVM to assess in-meal eating behavior [48]. Accuracy of algorithm was ranged from 60% to 82% based on population groups but the lack of similar populations data prevented testing the robustness of the model. Pouladzadeh et al. [49] used SVM to calculate calorie intake. The algorithm used food images provided by the user. Then, based on the features of image color, texture, food portion, size and shape, it provides output of the calorie estimate of that food. The algorithm was able to recognize food images with an accuracy between 75% to 99%. Although the model had promising results, authors stated that it could not achieve same results with other food pictures due to reasons such as different plates colors and textures.

3.6. Extreme gradient boosting

Despite recent success and popularity of this algorithm, it is rarely applied in obesity studies. Dunstan et al. [36] is one of a few studies that applied this algorithm to build a model to predict obesity based on nationwide food sales data. The algorithm was able to predict obesity with 80% accuracy. However, authors questioned robustness of XGBoost process when obtaining the variable important list. Another paper [42] applied XGBoost to predict 30-day readmission of obese patients based on historic hospitalization records. Their model had an AUC of 0.68. Although it had the best result, authors stated that interpreting the model was challenging.

3.7. Neural network

Neural networks are used in a few obesity studies. Their use includes calorie measurement, resting energy prediction [51], and body fat percentage. Disse et al. [51] used a neural network algorithm to calculate resting energy expenditure. Their algorithm had an accuracy of 73% and outperformed current statistical methods. However, they stated that neural network model is mathematically challenging in comparison to statistical methods which affected its acceptance among clinicians. Kupusinac et al. [43] also successfully applied neural network to calculate percentage of body fat with an accuracy of 80%. Nonetheless, the model was limited to certain population and might not be as accurate when applied to another population.

3.8. Deep learning

Application of deep learning in obesity prediction is rare. It mainly involves image analysis to estimate food calorie [50,53] or to understand obesity prevalence from built environment images. Heravi et al. [50] used a deep neural network to calculate food calories from food pictures taken by smartphone. The model was more accurate than previous models and had an accuracy ranged from 0.62 to 0.96 for different food classes. The authors believed that the lack of training images has affected the model accuracy in recognizing some food classes. In [38] deep learning was applied to classify health risk by using body fat levels and blood pressure. Their model was able to classify the risk with an AUC of 0.94. Although the model had good results, authors mentioned some difficulties applying deep learning such as the need for large amount of data and for hyperparameter optimization.

3.9. Clustering

Clustering techniques answer hypothesis about the causal effects in obesity studies [38]. It is mainly applied to discover dietary pattern and dietary behaviors. K-means cluster analysis was used in [47] to determine eating styles from eating behaviors and examine the relations between these styles and weight status. It was able to find 4 clusters of eating habits (healthy, unhealthy, healthy with problem eating behaviors, unhealthy with problem eating behaviors). Clustering analysis was also used in [41] to group obese individuals based on features such as age, health and demographic information. Both papers mentioned that data used was self-reported which might be based on inaccurate measures and subject to biases.

4. Discussion

This study reviewed the previous research that applied different ML algorithms to predict obesity. The overall impact of ML in obesity studies is promising because many studies have reported moderate to high accuracy of models ranging between 0.70 to 0.96, which gives researchers the confidence to use ML in studying obesity. Regression models were the most frequently applied in obesity prediction. Other ML models had promising performances but were less exploited. Despite being generally successful, ML has some limitations.

4.1. Limitations of algorithms

Different ML algorithms have different limitations. Regression algorithms suffer from linearity assumptions and data priori assumptions such as normal distribution [54], which is not always the case in obesity dataset. Their accuracy also decreases with increased number of outliers. In [44], for example, the relationship between the success of weight loss and first month weight loss was nonlinear. Applying regression in a situation like that can lead to losing valuable information and negatively affects model performance [55]. Despite these limitations, regression algorithms are still widely applied because they are easy to learn, apply, interpret [56] and most researchers have received training on them [38]. Additionally, they are less mathematically challenging than other algorithms such as deep neural networks [57].

Other non-regression algorithms (e.g., random forest, deep neural network) are also not without flaws. They suffer from the limitations such as overfitting where a model fits certain data perfectly but fails when applied to other datasets. Overfitting models are ungeneralizable. The mediation techniques include cross-validation and bootstrapping [54]. Another limitation is hyper-parameter optimization. Every ML algorithm needs predefined inputs to find the optimal unbiased models. This is uneasy task and could positively or negatively affect the model performance [38,55]. Model interpretation is also a concern when applying some new, advanced ML such as XGBoost, random forest [36,42] and deep neural network [57]. Another issue with deep neural network is that they can be computationally expensive, however this has changed lately due to the recent advancement in computer processors and graphical processing units.

4.2. Limitations in terms of data

Machine learning (deep neural network in particular) needs a massive amount of data to produce good results [19]; however, clinical studies are costly and data could be difficult to obtain [14]. Thus, the majority of papers reviewed here relied on large population surveys, social media or grocery sales to collect data. However, these surveys could be inaccurate and limit the generalizability of the study outcomes. Data in the self-reported studies could also be biased. Subjects could under-report actual weight or less dietary intake [58]. Other studies compared clinical data to survey data and argued that the difference has minimal effect on the overall accuracy when adjusted for socio demographic differences [59]. To reach a middle-ground, we found that applying ML-based equations such as in [60] to cross-sectional data could help minimize the effect of false measurement reporting.

Cross-sectional study design also limits the inference of causality where it is unknown, for example, if nutrition and lifestyle causes obesity or vice versa [33,39].

Another limitation with healthcare datasets is that they suffer from a huge amount of noise (i.e., missing data, data entry errors, unbalanced dataset). This affects data quality, and can lead to weaker predictive models [61]. Ferenci and Kovács [62] suggested that data quality could be improved by removing every subject with missing values or by applying different data processing techniques such as dimensionality reduction, feature selection or feature extraction. These techniques will help in building more robust models [15].

4.3. Recommendations for future machine learning use in obesity prediction

There are a wide range of ML algorithms, each with its own advantages and limitations. Accuracy of algorithms differs between studies suggesting that quality and properties of available data along with the nature of study play important role in applying the right algorithm (Table 2). Therefore, it is recommended to use more than one model especially if hypothesis is unclear [23,46].

New advancement in ML algorithms such as deep neural network and XGBoost have opened new opportunities for innovative research in obesity. Easton et al. [35] point out that new ML techniques are widely used in studying various health issues such as heart disease and diabetes. However, their utilization in obesity and nutritional studies is still limited despite their high promise. Several other researchers also suggest to apply new ML tools to improve the accuracy in obesity prediction [13,57,63]. Jothi et al. [56] recommend to apply these algorithms to analyze the big volume of healthcare data produced, which is beneficial for generating insight from large, complex data.

4.4. Limitation of this review study

As the purpose of this study was to assess the application of ML methods in obesity studies, not the number of obesity studies applied ML methods, we only included two studies applied the same ML methods in the review. This may cause unintentional exclusion of some ML techniques that have not been included, e.g., reinforcement learning, association rules and principal component analysis. In addition, studies used ML or data mining prior to 2014 were not included in this study.

5. Conclusions

Understanding the nature of available data and study methods is a key factor for selecting the suitable machine learning technique and model for health research. This review of the application of ML in obesity studies suggest that ML provides the essential, useful analytical tools in predicting obesity. However, the modern ML techniques have not been sufficiently applied in obesity studies, despite the promising performance. Further use of the recent development in ML technology should be promoted in the obesity research.

Acknowledgments

The author, Mohammad Alkhalaf, is supported by a full PhD scholarship from Qassim University, Saudi Arabia.

Conflict of interest

All authors declare that there is no conflict of interests in this paper.

References

1. WHO, Obesity and Overweight, World Health Organization, 2020. Available from: https://wwwwhoint/news-room/fact-sheets/detail/obesity-and-overweight.
2. A. Hruby, J. E. Manson, L. Qi, V. S. Malik, E. B. Rimm, Q. Sun, W. C. Willett, F. B. Hu, Determinants and consequences of obesity, *Am. J. Public Health*, **106** (2016), 1656–1662. https://doi.org/https://doi.org/10.2105/AJPH.2016.303326

3. WHO, The top 10 causes of death, World Health Organization, 2018. Available from: https://wwwwhoint/news-room/fact-sheets/detail/the-top-10-causes-of-death.
4. WHO, 10 facts on obesity, World Health Organization, 2017. Available from: https://wwwwhoint/features/factfiles/obesity/en/..
5. J. Cawley, C. Meyerhoefer, The medical care costs of obesity: An instrumental variables approach, *J. Health Econ.,* **31** (2012), 219–230. https://doi.org/10.1016/j.jhealeco.2011.10.003
6. L. Angrisani, A. Santonicola, P. Iovino, G. Formisani, H. Buchwald, N. Scopinaro, Bariatric Surgery Worldwide 2013, *Obes. Surg.,* **25** (2015), 1822–1832. https://doi.org/10.1007/s11695-015-1657-z
7. T. Bhurosy, R. Jeewon, Overweight and obesity epidemic in developing countries: A problem with diet, physical activity, or socioeconomic status? *Scientific World Journal,* **2014** (2014). https://doi.org/10.1155/2014/964236
8. E. Alpaydin, *Introduction to Machine Learning,* Cambridge: MIT press, 2014.
9. N. S. Rajliwall, R. Davey, G. Chetty, Machine learning based models for cardiovascular risk prediction, *International Conference on Machine Learning and Data Engineering 2018, (iCMLDE),* (2018), 142–148. https://doi.org/10.1109/iCMLDE.2018.00034
10. J. B. Heaton, N. G. Polson, J. H. Witte, Deep learning for finance: deep portfolios, *Appl. Stoch. Model. Bus.,* **33** (2017), 3–12. https://doi.org/10.1002/asmb.2209
11. J. Kim, J. Canny, Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention, *Proceedings of the IEEE International Conference on Computer Vision,* (2017), 2942–2950. https://doi.org/10.1109/ICCV.2017.320
12. D. Gruson, T. Helleputte, P. Rousseau, D. Gruson, Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation, *Clin. Biochem.,* **69** (2019), 1–7. https://doi.org/10.1016/j.clinbiochem.2019.04.013
13. D. Panaretos, E. Koloverou, A. C. Dimopoulos, G. M. Kouli, M. Vamvakari, G. Tzavelas, C. Pitsavos, D. B. Panagiotakos, A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): The ATTICA study, *Brit. J. Nutr.,* **120** (2018), 326–334. https://doi.org/10.1017/S0007114518001150
14. H. C. Koh, G. Tan, Data Mining Applications in Healthcare, *Journal of Healthcare Information Management,* **19** (2011), 64–72.
15. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, D. I. Fotiadis, Machine learning applications in cancer prognosis and prediction, *Comput. Struct. Biotec.,* **13** (2015), 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
16. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *JAMA - Journal of the American Medical Association,* **316** (2016), 2402–2410. https://doi.org/10.1001/jama.2016.17216
17. Y. Xing, J. Wang, Z. Zhao, Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease, *International Conference on Convergence Information Technology, (ICCIT) 2007,* (2007), 868–872. https://doi.org/10.1109/ICCIT.2007.4420369
18. P. Fränti, S. Sieranoja, K. Wikström, T. Laatikainen, Clustering diagnoses from 58M patient visits in Finland during 2015-2018, *JMIR Medical Informatics,* (2022). https://doi.org/10.2196/35422

19. Z. Obermeyer, E. J. Emanuel, Predicting the Future: Big Data, Machine Learning, and Clinical Medicine, *The New England journal of medicine,* **375** (2016), 1216–1219. https://doi.org/doi:10.1056/NEJMp1606181
20. M. A. Morris, E. Wilkins, K. A. Timmins, M. Bryant, M. Birkin, C. Griffiths, Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map, *Int. J. Obesity,* 42 (2018), 1963–1976. https://doi.org/10.1038/s41366-018-0184-0
21. C. Y. J. Peng, K. L. Lee, G. M. Ingersoll, An introduction to logistic regression analysis and reporting, *J. Educ. Res.,* **96** (2002), 3–14. https://doi.org/10.1080/00220670209598786
22. D. Dietrich, B. Heller, Y. Beibei, *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data,* Indianapolis: Wiley, 2015.
23. H. O. Alanazi, A. H. Abdullah, K. N. Qureshi, A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care, *J. Med. Syst.,* **41** (2017), 1–10. https://doi.org/10.1007/s10916-017-0715-6
24. Y. Y. Song, L. U. Ying, Decision tree methods: applications for classification and prediction, *Shanghai Archives of Psychiatry,* **27** (2015), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
25. M. Pal, Random forest classifier for remote sensing classification, *Int. J. Remote Sens.,* **26** (2005), 217–222. https://doi.org/10.1080/01431160412331269698
26. S. V. Vishwanathan, M. N. Murty, SSVM: A simple SVM algorithm, *International Joint Conference on Neural Networks (IJCNN) 2002,* **3** (2002), 2393–2398. https://doi.org/10.1109/IJCNN.2002.1007516
27. Y. Qu, B. Fang, W. Zhang, R. Tang, M. Niu, H. Guo, Y. Yu, X. He, Product-Based Neural Networks for User Response Prediction over Multi-Field Categorical Data, *ACM T. Inform. Syst.,* **37** (2019), 1–35. https://doi.org/10.1145/3233770
28. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (2016), 785–794. https://doi.org/10.1145/2939672.2939785
29. A. T. C. Goh, Back-propagation neural networks for modeling complex systems, *Artificial Intelligence in Engineering,* **9** (1995), 143–151. https://doi.org/10.1016/0954-1810(94)00011-S
30. Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature,* **521** (2015), 436–444. https://doi.org/10.1038/nature14539
31. A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, *ACM Comput. Surv.,* **31** (1999), 264–323. https://doi.org/10.1145/331499.331504
32. H. Arksey, L. O'Malley, Scoping studies: towards a methodological framework, *Int. J. Soc. Res. Method.,* **8** (2005), 19–32. https://doi.org/10.1080/1364557032000119616
33. H. So, L. McLaren, G. C. Currie, The relationship between health eating and overweight/obesity in Canada: cross-sectional study using the CCHS, *Obesity Science and Practice,* **3** (2017), 399–406. https://doi.org/10.1002/osp4.123
34. N. Daud, N. L. Mohd Noor, S. A. Aljunid, N. Noordin, N. I. M. F. Teng, Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity, *2018 IEEE Conference on Big Data and Analytics, ICBDA,* (2018), 1–6. https://doi.org/10.1109/ICBDAA.2018.8629623
35. J. F. Easton, H. Román Sicilia, C. R. Stephens, Classification of diagnostic subcategories for obesity and diabetes based on eating patterns, *Nutr. Diet.,* **76** (2019), 104–109. https://doi.org/10.1111/1747-0080.12495

36. J. Dunstan, M. Aguirre, M. Bast ás, C. Nau, T. A. Glass, F. Tobar, Predicting nationwide obesity from food sales using machine learning, *Health Inform. J.,* **26** (2019), 652–663. https://doi.org/10.1177/1460458219845959
37. N. Kanerva, J. Kontto, M. Erkkola, J. Nevalainen, S. Mannisto, Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design, *Scand. J. Public Health,* **46** (2018), 557–564. https://doi.org/10.1177/1403494817736944
38. K. W. DeGregory, P. Kuiper, T. DeSilvio, J. D. Pleuss, R. Miller, J. W. Roginski, C. B. Fisher, D. Harness, et al., A review of machine learning in obesity, *Obes. Rev.,* **19** (2018), 668–685. https://doi.org/10.1111/obr.12667
39. D. Kim, W. Hou, F. Wang, C. Arcan, Factors Affecting Obesity and Waist Circumference Among US Adults, *Prev. Chronic Dis.,* **16** (2019). https://doi.org/10.5888/pcd16.180220
40. R. L. Figueroa, C. A. Flores, Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures, *J. Med. Syst.,* **40** (2016). https://doi.org/10.1007/s10916-016-0548-8
41. M. A. Green, M. Strong, F. Razak, S. V. Subramanian, C. Relton, P. Bissell, Who are the obese? A cluster analysis exploring subgroups of the obese, *J. Public Health (UK),* **38** (2016), 258–264. https://doi.org/10.1093/pubmed/fdv040
42. P. P. Brzan, Z. Obradovic, G. Stiglic, Contribution of temporal data to predictive performance in 30-day readmission of morbidly obese patients, *PeerJ,* **5** (2017), e3230. https://doi.org/10.7717/peerj.3230
43. A. Kupusinac, E. Stokić, R. Doroslovački, Predicting body fat percentage based on gender, age and BMI by using artificial neural networks, *Comput. Meth. Prog. Bio.,* **113** (2014), 610–619. https://doi.org/10.1016/j.cmpb.2013.10.013
44. M. Batterham, L. Tapsell, K. Charlton, J. O'shea, R. Thorne, Using data mining to predict success in a weight loss trial, *J. Hum. Nutr. Diet.,* **30** (2017), 471–478. https://doi.org/10.1111/jhn.12448
45. Z. Feng, L. Mo, M. Li, A Random Forest-based ensemble method for activity recognition, *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2015 EMBS,* (2015), 5074–5077. https://doi.org/10.1109/EMBC.2015.7319532
46. M. Batterham, E. Neale, A. Martin, L. Tapsell, Data mining: Potential applications in research on nutrition and health, *Nutr. Diet.,* **74** (2017), 3–10. https://doi.org/10.1111/1747-0080.12337
47. W. J. Heerman, N. Jackson, M. Hargreaves, S. A. Mulvaney, D. Schlundt, K. A. Wallston, R. L. Rothman, Clusters of Healthy and Unhealthy Eating Behaviors Are Associated With Body Mass Index Among Adults, *J. Nutr. Educ. Behav.,* **49** (2017), 415–421. https://doi.org/10.1016/j.jneb.2017.02.001
48. I. Sarasfis, C. Diou, I. Ioakimidis, A. Delopoulos, Assessment of In-Meal Eating Behaviour using Fuzzy SVM, *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* (2019), 6939–6942. https://doi.org/10.1109/EMBC.2019.8857606
49. P. Pouladzadeh, S. Shirmohammadi, A. Bakirov, A. Bulut, A. Yassine, Cloud-based SVM for food categorization, *Multimed. Tools Appl.,* **74** (2015), 5243–5260. https://doi.org/10.1007/s11042-014-2116-x

50. E. J. Heravi, H. Habibi Aghdam, D. Puig, A deep convolutional neural network for recognizing foods, *Eighth International Conference on Machine Vision (ICMV),* **9875** (2015), 98751D. https://doi.org/10.1117/12.2228875

51. E. Disse, S. Ledoux, C. Bétry, C. Caussy, C. Maitrepierre, M. Coupaye, M. Laville, C. Simon, An artificial neural network to predict resting energy expenditure in obesity, *Clin. Nutr.,* **37** (2018), 1661–1669. https://doi.org/10.1016/j.clnu.2017.07.017

52. N. Cesare, P. Dwivedi, Q. C. Nguyen, E. O. Nsoesie, Use of social media, search queries, and demographic data to assess obesity prevalence in the United States, *Palgrave Communications,* **5** (2019), 1–9. https://doi.org/10.1057/s41599-019-0314-x

53. P. Kuhad, A. Yassine, S. Shimohammadi, Using distance estimation and deep learning to simplify calibration in food calorie measurement, *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA,* (2015),1–6. https://doi.org/10.1109/CIVEMSA.2015.7158594

54. K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, P. P. Sengupta, Machine learning in cardiovascular medicine: Are we there yet? *Heart,* **104** (2018), 1156–1164. https://doi.org/10.1136/heartjnl-2017-311198

55. B. A. Goldstein, A. M. Navar, R. E. Carter, Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges, *Eur. Heart J.,* **38** (2017), 1805–1814. https://doi.org/10.1093/eurheartj/ehw302

56. N. Jothi, N. A. A. Rashid, W. Husain, Data Mining in Healthcare - A Review, *Procedia Computer Science,* **72** (2015), 306–313. https://doi.org/10.1016/j.procs.2015.12.145

57. A. L. Beam, I. S. Kohane, Big data and machine learning in health care, *JAMA - Journal of the American Medical Association,* **319** (2018), 1317–1318. https://doi.org/10.1001/jama.2017.18391

58. A. Mozumdar, G. Liguori, Corrective Equations to Self-Reported Height and Weight for Obesity Estimates among U.S. Adults: NHANES 1999-2008, *Res. Q. Exercise Sport,* **87** (2016), 47–58. https://doi.org/10.1080/02701367.2015.1124971

59. M. Stommel, C. A. Schoenborn, Accuracy and usefulness of BMI measures based on self-reported weight and height: Findings from the NHANES & NHIS 2001-2006, *BMC Public Health,* **9** (2009), 1–10. https://doi.org/10.1186/1471-2458-9-421

60. D. Rativa, B. J. T. Fernandes, A. Roque, Height and Weight Estimation from Anthropometric Measurements Using Machine Learning Regressions, *IEEE J. Transl. Eng. He.,* **6** (2018), 1–9. https://doi.org/10.1109/JTEHM.2018.2797983

61. J. A. Sáez, J. Luengo, F. Herrera, Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification, *Pattern Recogn.,* **46** (2013), 355–364. https://doi.org/10.1016/j.patcog.2012.07.009

62. T. Ferenci, L. Kovács, Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks, *Applied Soft Computing Journal,* **67** (2018), 834–839. https://doi.org/10.1016/j.asoc.2017.05.063

63. S. P. Goldstein, F. Zhang, J. G. Thomas, M. L. Butryn, J. D. Herbert, E. M. Forman, Application of Machine Learning to Predict Dietary Lapses During Weight Loss, *Journal of Diabetes Science and Technology,* **12** (2018), 1045–1052. https://doi.org/10.1177/1932296818775757

Appendix

Table 4. Sample of search terms used in this literature review.

For all search keywords obesity, physical activity, nutrition and diet used Ex: Obesity AND "machine learning" : physical activity AND “machine learning”
Obesity AND "Logistic regression"
Obesity AND "Linear regression"
Obesity AND "decision trees"
Obesity AND "Na ïve Bayes"
Obesity AND "neural network"
Obesity AND "deep learning"
Obesity AND "Random Forest"
Obesity AND “Extreme gradient boosting”
Obesity AND "cluster analysis"
Obesity AND "support vector machine"