IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 5, 2021, accepted November 21, 2021, date of publication November 25, 2021,
date of current version December 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131128

A Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity

HERA SIDDIQUI[1], AJITA RATTANI[1], NIKKI K. WOODS[2], LAILA CURE[3], RHONDA K. LEWIS[4], JANET TWOMEY[3], BETTY SMITH-CAMPBELL[5], AND TWYLA J. HILL[6]

[1]School of Computing, Wichita State University, Wichita, KS 67260, USA
[2]Department of Public Health Sciences, Wichita State University, Wichita, KS 67260, USA
[3]Department of Industrial and Manufacturing Engineering, Wichita State University, Wichita, KS 67260, USA
[4]Department of Psychology, Wichita State University, Wichita, KS 67260, USA
[5]School of Nursing, Wichita State University, Wichita, KS 67260, USA *(Retired)*
[6]Department of Sociology, Wichita State University, Wichita, KS 67260, USA

Corresponding author: Hera Siddiqui (hxsiddiqui@shockers.wichita.edu)

This work was supported in part by the President's Convergence Science Initiative Grant from Wichita State University under Grant U29001.

ABSTRACT Childhood and adolescent obesity is a serious health problem that is on the rise at the global level. Earlier, certain diseases such as Type 2 diabetes, high blood pressure, and heart disease affected only adults, but now they are being detected in young children as well. Several studies based on machine learning have been proposed to develop obesity prediction models or to determine key determinants of obesity for designing intervention tools. Despite having a rich and diverse set of literature on obesity prediction models, obesity rates are at an all-time high for both children and adolescents. There is a need of proper understanding and critical analysis of existing machine learning models in order to design effective strategies for curbing obesity at childhood and adolescent level. This paper surveys the growing body of recent literature on machine and deep learning models for obesity prediction by providing a coherent view (critical analysis) of the limitations of the existing systems. The taxonomy of the existing literature on obesity prediction into methods used, predicted outcome, factors used, type of datasets, and the associated purpose, is discussed for analysis of the state-of-the-art. This analysis revealed that a) prediction-focused models do not use variables from as many domains as predictor-focused models do, b) very few studies proposed gender-specific and race-specific obesity prediction models, c) lack of large-scale multimodal datasets and d) existing predictor-focused models obtain an accuracy range of [53.7%, 96%] with an optimum set of predictors. Further, computer vision-based methods for obesity prediction and interpretable techniques for understanding the outcome of the models are discussed as well. In addition, we have also identified novel research directions. The overall aim is to advance the state-of-the-art and improve the quality of discourse in this field.

INDEX TERMS Adolescent obesity, childhood obesity, deep learning, machine learning, obesity prediction, key determinants.

I. INTRODUCTION

World Health Organization (WHO) defines overweight and obesity as ''abnormal or excessive fat accumulation that presents a risk to health''.[1] One of the most challenging healthcare problems that the world is facing today is that of childhood and adolescent obesity. It is not limited to a single country but has become a global public health crisis. For the year 2017-2018 about 14.4 million children and adolescents (aged 2-19 years) were affected by obesity in the United States. Among 2-5 years old, its prevalence was 13.4%, for 6-11 years old, it was 20.3%, and 21.2% for 12-19 years old. Data also shows that the distribution is not uniform, as certain populations (Hispanic and non-Hispanic Black) are more vulnerable to obesity. Among Hispanic children, the prevalence of obesity was 25.6%, 24.2% among non-Hispanic Black children, for non-Hispanic White children it was 16.1%, and 8.7% among non-Hispanic Asian children.[2] In not just the United States, the prevalence of obesity among children and adolescents is increasing all over the world. Globally about 41 million children under the age of 5 were overweight. Children under the age of 5 living in Asia contributed half to this number and about one quarter belonged to Africa.[3]

The associate editor coordinating the review of this manuscript and approving it for publication was Aasia Khanum.

[1]https://www.who.int/health-topics/obesity

[2]https://www.cdc.gov/obesity/data/childhood.html

[3]https://www.who.int/news-room/q-a-detail/noncommunicable-diseases-childhood-overweight-and-obesity

Diseases such as Type 2 diabetes, heart disease, blood vessel disease, and obesity-related depression, and social isolation, usually associated with adults are now being detected in children as well. Once obesity sets in, it is very difficult to treat it as the causes of childhood and adolescent obesity are complex and multifaceted which makes it a very challenging task. Therefore, immediate steps must be taken to prevent obesity. If obesity is not curbed at childhood itself then there is a greater chance of its persistence into adulthood. Adult obese people have a higher risk of developing diabetes, high blood pressure, and heart disease.

Adults are classified as overweight or obese based on Body Mass Index (BMI) which is defined as the weight in kilograms divided by the square of the height in meters (kg/m^2). For children and adolescents, this metric of classification is not that simple because as they grow, their bodies undergo several physiological changes. Factors such as age, puberty, and growth rate influence the rate of fat deposition and removal. As a result, defining a standard for overweight and obesity that incorporates all age groups is very difficult [1]. To assess obesity, three classification systems are used at the international level: International Obesity Task Force (IOTF), the United States Centers for Disease Control and Prevention (CDC) growth charts issued in 2000, and the World Health Organization (WHO) criteria. International Obesity Task Force (IOTF) [2], constructed in 2000 and updated in 2002 with the help of datasets from 6 countries (Singapore, Netherlands, Brazil, Hong Kong, the UK, the USA), uses sex-specific BMI curves that match adult BMI values of $25kg/m^2$ (Overweight) and $30kg/m^2$ (Obesity) at 18 years. The World Health Organization (WHO) criteria [3] supplemented with data from the WHO Child Growth Standards for children aged 5 years and younger, was developed in 2007 using the 1977 National Center for Health Statistics (NCHS) growth reference from 5 to 19 years. It defines overweight as a BMI $>$ 1 standard deviation (SD) and obesity as a BMI $>$ 2 SD from the mean of the WHO reference population. CDC growth charts [4] were a revision of the NCHS 1977 growth reference that incorporated data from five national surveys conducted between 1963 and 1994 in the United States of America. This system defines overweight as a BMI $>$ 85th percentile of the reference population and obesity as a BMI $>$ 95th percentile.

Obesity in children was already on the rise however amid the COVID-19 crisis, it has increased further due to stricter lockdowns, homeschooling, and other steps taken to stop the transmission of the virus. Children are attending virtual schools, there is not much physical activity, and there is more reliance on high-calorie food. The increase in prevalence was more in young children who depend on the family for dietary choices.[4] This increase is especially substantial for Black, Hispanic, and low-income children which in turn has widened the pre-existing disparity. Many countries have initiated stricter measures to deal with the epidemic of childhood obesity. For instance, UK is planning to ban junk food advertisements online on social media, and before 9 pm on Television from 2023 to combat child obesity.[5] Oaxaca, in Mexico, which has the highest childhood obesity rate, in early August 2020, passed a ban that prohibited retailers from selling or promoting processed snacks such as candy, chips, and soda to children under the age of 18.[6] Germany has also tightened its rules on marketing junk food to children.[7] However, such strict policies have not been implemented in the US, rather new research shows that Black and Hispanic youth are unfairly targeted by advertisements.[8] Both these ethnicities have higher rates of childhood obesity compared to others in the United States.

A number of studies have been proposed that use machine/deep learning models on different formats of data (a) for **obesity prediction** [5]–[8], and (b) for determining the **key determinants** of obesity [5], [9]–[13] for developing intervention techniques.

The aim of this paper is *to survey the recent body of growing literature* on Machine Learning (ML) and Deep Learning (DL) models for childhood and adolescent obesity prediction. To this aim, a taxonomy of the existing literature on machine learning and deep learning models for childhood and adolescent obesity prediction was developed. This paper provides a detailed analysis of predictor-focused (to study factors associated) and prediction-focused (accurate prediction) models which helps in identifying the existing gap between the two approaches. Future studies can build upon the findings of this survey to develop better prediction models that include datasets from various domains. This paper provides analysis on both association and prediction of childhood and adolescent obesity which can help in developing effective obesity intervention programs.

Google Scholar, PubMed, Scopus, and IEEE Xplore were used for finding papers using different combinations of keywords such as child, childhood, adolescent, obesity, machine learning, deep learning, etc. Papers were included in this survey using the following rule: 1.) those proposing machine learning models that extract patterns, and risk profiles using BMI or BMI categories, 2.) ML models that predict overweight/obesity or BMI in the future, 3.) those based on childhood and adolescent obesity with subjects in the age range 0-18, and 4.) papers from 2010 onwards up until March 2021 were included. Table 1 provides the list of abbreviations used in the manuscript.

This paper is organized as follows: We discuss the existing survey papers on childhood and adolescent obesity in

[4]https://www.inquirer.com/health/coronavirus/pandemic-diabetes-obesity-kids-increase-20210630.html

[5]https://www.theguardian.com/media/2021/jun/23/uk-to-ban-junk-food-advertising-online-and-before-9pm-on-tv-from-2023

[6]https://www.nycfoodpolicy.org/food-policy-snapshot-oaxaca-junk-food-ban-for-minors/

[7]https://www.foodnavigator.com/News/Policy/Germany-tightens-rules-on-marketing-food-to-children-Advertising-must-not-induce-children-to-eat-unhealthily

[8]https://www.medicalnewstoday.com/articles/fast-foods-equity-problem-black-and-hispanic-youth-unfairly-targeted-by-ads

TABLE 1. List of abbreviations used in this manuscript and their full forms.

Abbreviation	Full Form
ABCD	Adolescent Brain Cognitive Development
ANN	Artificial Neural Network
AUC	Area Under the ROC
BCH	Boston's Children Hospital
BMI	Body Mass Index
BN	Bayes Net
CART	Classification and Regression Trees
CCHMC	Cincinnati Children's Hospital and Medical Center
CDC	United States Centers for Disease Control and Prevention
CDS	Computerized Decision Support
CHICA	Child Health Improvement through Automation
CNN	Convolutional Neural Network
CRF	Conditional Random Forest
CYKIDS	Cyprus Kids Study
DC	Degree Centrality
DL	Deep Learning
DMRN	Deep Multi-Cue Regression Network
DT	Decision Trees
ECLS-B	Early Childhood Longitudinal Study Birth Cohort
EHR	Electronic Health Record
FGFQ	Food Group Frequency Questionnaire
GA	Genetic Algorithm
GBM	Gradient Boosting Methods
GBN	General Bayesian Network
GIS	Geographic Information System
HIMSS	Health Information and Management Systems Society
ID3	Iterative Dichotomiser 3
IDT	Improved Decision Tree
IOTF	International Obesity Task Force
KCDC	Korean Centers for Disease Control and Prevention
KNHI	Korean National Health Insurance
KNN	K-Nearest Neighbor
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCS	Millennium Cohort Study
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NB	Naive Bayes
NCHS	National Center for Health Statistics
NHGS	NHLBI Growth and Health Study
NKI-RS	Enhanced Nathan Kline Institute Rockland Sample
NLP	Natural Language Processing
PBD	Pediatric Bigdata Repository
PCA	Principal Component Analysis
PE	Physical Education
PFT	Physical Fitness Test
PPV	Positive Predicted Value
QUALITY	Quebec Adiposity and Lifestyle Investigation in Youth
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
ROI	Region of Interest
rs-fMRI	Resting-state Functional Magnetic Resonance Imaging
SES	Socioeconomic Status
SHAP	Shapley Additive exPlanation
SVM	Support Vector Machine
SVR	Support Vector Regressor
WC	Waist Circumference
WHO	World Health Organization
YRBS	National Youth Risk Behavior Survey

Section II along with our contributions. In Section III, we present the taxonomy of the existing studies on obesity. Methods, Datasets, Outcome Predicted, and Factors used in Machine Learning Models are discussed in Section III along with the purpose that they serve. Discussion and future research directions are present in Section IV.

II. PREVIOUS SURVEYS ON OBESITY

There are only four survey papers on Machine Learning models for childhood and adolescent obesity to date to the best of our knowledge. Table 2 summarizes the existing surveys (reviews) published until 2021.

The four reviews are very different in their approach. Adnan *et al.* [14] summarized papers based on only three models - Artificial Neural Network (ANN), Naive Bayes (NB), and Decision Trees (DT). This review was mostly on the effectiveness, efficacy, strengths, and weaknesses of these three data mining techniques, and only two papers were covered which is justified because at the time of their review there were not many papers published ([15]–[17]) that utilized Machine Learning methods for childhood obesity prediction. ANN was primarily used for finding out the risk factors that were associated with childhood obesity, Naive Bayes for classification, and Decision Tree for weight management counseling system. Their survey paper included studies that used Machine Learning for three different aspects of childhood obesity i.e., risk factor association, classification, and weight management/intervention. The authors summarized that for an ANN classifier, the attribute environment is considered as an important factor, Naive Bayes focused more on a child's personal attributes and Decision Trees are easier to build than the other two methods on the Wirral dataset.

Recently, in 2020 survey papers on obesity prediction were published by Triantafyllidis *et al.* [18] and Colmenarejo [19], respectively. Triantafyllidis *et al.* [18] did a systematic literature review of Computerized Decision Support (CDS) and Machine Learning (ML) applications for the prevention and treatment of childhood obesity. The authors reviewed 8 studies on CDS interventions and 9 studies utilizing ML algorithms and concluded that for self-management or remote medical management of childhood obesity, CDS tools can be useful. Further, for prediction purposes, ML algorithms such as decision trees and artificial neural networks could be utilized. This review included a very limited number (n=9) of studies on ML and with it being a systematic review did not elaborate much on factors, methods, etc. One interesting find was that ML algorithms were not used in Computer Decision Support intervention systems in a clinical setting.

Colmenarejo [19] did a comprehensive and critical review of Machine Learning models to predict childhood and adolescent obesity and the related outcomes. This paper was more of a comparison between Statistical and Machine Learning prediction models and it suggested the use of ML models in most situations. Although their paper was quite thorough, it lacked information on feature selection methods and the incorporation of interpretability in ML models. The paper made a case for why Machine Learning models should be preferred over Statistical ones, reasons being their excellent predictive power; ability to model complex, nonlinear

TABLE 2. Summary and analysis of the previous machine learning reviews on childhood and adolescent obesity.

Article	Conclusions drawn	Limitations
Adnan et al. [14] (2010)	1. For ANN, the environment is an important factor. 2. Naive Bayes focused more on a child's personal attributes. 3. Decision Trees are easier to build than the other two methods.	Very few studies and methods were discussed.
Triantafyllidis et al. [18] (2020)	1. For self-management or remote medical management of childhood obesity, Computer Decision Support (CDS) tools can be useful. 2. For prediction purposes, ML algorithms such as decision trees and artificial neural networks could be helpful.	1. Considered studies using CDS interventions or having advanced data analytics through ML algorithms. 2. Had very few ML studies included due to strict inclusion criteria
Colmenarejo [19] (2020)	1. Statistical models are probably more oriented towards earlier ages, where the number of impacting factors do not vary. 2. Multidomain factors are more appropriate. 3. DL/ML are appropriate tools for the modeling of Electronic Health Records (EHR). 4. ML has good predictive performance but less interpretability.	1. Mostly compared Statistical and ML models. 2. Limited information on datasets, feature selection, etc.
LeCroy et al. [20] (2021)	1. 0-24 months weight history of the child, parental overweight or obesity were key risk factors. 2. Middle and late childhood/adolescence social factors and were important factors. 3. Race/ethnic-specific models may be needed to accurately predict obesity from middle childhood onward.	Only sociodemographic and behavioral factors were considered for selecting studies.

TABLE 3. Differences between previously published surveys and our survey.

Article	Previous Survey	Our Survey
Adnan et al. [14] (2010)	Summarized papers only on models -ANN, NB, and DT.	Broad range of both traditional machine learning (e.g.: DT, RF, KNN) and deep learning models are covered (e.g.: ANN, RNN, CNN).
Triantafyllidis et al. [18] (2020)	Only 9 ML studies were covered and very few details provided about methods, factors, etc.	27 ML studies are covered. Information on datasets, factors, methods, performance (when available) is provided.
Colmenarejo [19] (2020)	Very comprehensive and thorough review, but compared statistical models with ML models mostly.	Provides a taxonomy of the models based on methods, datasets used, outcome predicted, factors used, and purpose of developing the models. Predictor-focused and prediction-focused models have been covered in detail.
LeCroy et al. [20] (2021)	Analysis was done according to age groups: early, middle, and late childhood.	Analysis done based on the developed taxonomy which identifies new research areas to work on.

relationships between variables; and capacity to deal with high-dimensional data.

The last and the most recent paper published in 2021 [20] was a narrative review of 15 peer-reviewed studies that used machine learning to predict childhood obesity using a combination of socio-demographic and behavioral risk factors. This paper talked about the methods as well as the determinants used for prediction in the three age groups: early, middle, and late childhood. The authors observed that between birth and 24 months of age, the child's weight history and parental overweight/obesity were key risk factors. For middle and late childhood/adolescence, social factors and physical inactivity were important factors. The authors suggested that race/ethnic-specific models may be needed to accurately predict obesity from middle childhood onward.

A. OUR CONTRIBUTIONS

Table 3 summarizes the differences between our survey and previously published surveys. The contributions of our survey (using 27 research articles) over existing survey articles on childhood and adolescent obesity prediction [14], [18]–[20] are as follows:

1) Providing a **taxonomy** of the models based on methods, datasets used, outcome predicted, factors used, and purpose of developing the models. This has not been previously done in any of the existing review papers [14], [18]–[20].
2) Including models that **rank predictor variables, create profiles** of groups at higher risk of development of obesity, and models focused on improving the accuracy of identifying children with obesity.
3) Providing a detailed review of studies that rank predictors (predictor-focused studies) and studies that aim to correctly identify children and adolescents at higher risk of developing obesity (prediction-focused studies).
4) Highlighting the **gap between predictor-focused and prediction-focused** studies. Prediction-focused studies do not utilize the findings of predictor-focused models to their full potential.
5) Including studies based on **computer vision methods** for children and adolescent obesity prediction, and

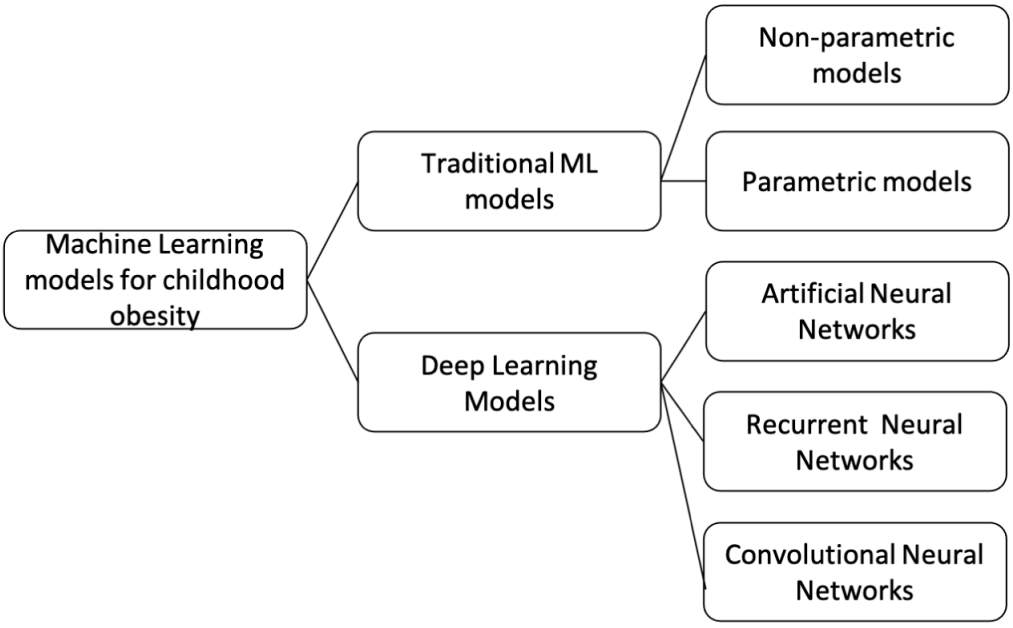

FIGURE 1. Categorization of the literature on machine learning models for childhood and adolescent obesity based on the type of the model used.

TABLE 4. Summary of the existing studies on machine learning methods used in childhood and adolescent obesity models.

Methods	Reference to Existing Studies
Parametric	[5] [6] [7] [8] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35]
Non-parametric	[36], [9], [37], [10], [11], [38], [39], [12], [13], [5], [40] , [24], [26] [27], [29], [41], [34] [35]
ANN	[26] [27], [31], [35]
RNN	[28]
CNN	[33]

interpretable/explainable models that help in understanding the reasoning behind the decisions made by the models.

III. TAXONOMY OF MACHINE LEARNING MODELS FOR OBESITY PREDICTION

Taxonomy of existing literature in surveys from other fields [21]–[23] has provided a comprehensive view of the state-of-the-art and the critical analysis based on it has opened up new research avenues. In this section, we will discuss the taxonomy of the existing literature on obesity prediction. The existing literature on obesity prediction using ML is categorized into methods used, type of datasets, predicted outcome, factors used, and purpose of developing the model. All these categories are discussed as follows:

A. METHODS

Figure 1 shows the categorization of methods used for developing childhood and adolescent obesity models. These methods can broadly be divided into traditional Machine Learning models and Deep Learning models. Traditional models can be further divided into Non-parametric and Parametric. Artificial Neural Networks, Recurrent Neural Networks, and Convolutional Neural Networks are branches of Deep Learning Models. Next, we will discuss each of these approaches. Table 4 summarizes the existing literature based on the method used to develop the model.

1) **Traditional Machine Learning models:** The current studies that use machine learning models can be broadly classified into two subcategories: Parametric and Non-parametric, based on the type of ML model used. Parametric models (such as Naive Bayes, Support Vector Machines, and Neural Networks) are based on learning parameters from the training data and non-parametric models (such as Decision Trees and K-nearest Neighbours) use the instances as it is for decision making [42].
   - **Parametric models:** The existing obesity papers that used parametric machine learning models such as Naive Bayes and LASSO (Least Absolute Shrinkage and Selection Operator) are [5]–[8], [24]–[35].
   - **Non-Parametric models** Almost all of the predictor focused studies [5], [9]–[13], [36]–[40] for childhood/adolescent obesity use this category of models as predictor importance is easier

to gauge. Non-parametric models such as Decision Trees (DT), k-Nearest Neighbors (KNN), and Random Forest (RF) were a popular choice for predictor-based methods [24], [26], [27], [29], [34], [35], [41].

2) **Deep Learning models:** Deep Learning models are engineered systems inspired by the biological brain [42]. For childhood and adolescent obesity, three types of deep learning models have been used until now: Artificial Neural Network (ANN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN).
   - **Artificial Neural Networks (ANN):** Artificial Neural Network is one of the simplest forms of neural networks. It can be represented as a group of multiple perceptrons/neurons at each layer. They are also known as Universal Function Approximators as they are capable of learning any non-linear function. This non-linearity is introduced by using activation functions. Studies in [26], [27], [31], and [35] have used Multi-Layered Perceptron for obesity prediction.
   - **Recurrent Neural Networks (RNN):** Recurrent neural networks are a family of neural networks for processing sequential data. These types of networks can scale to much longer sequences than would be practical for networks without sequence-based specialization [42]. Study in [28] used RNN for obesity prediction.
   - **Convolutional Neural Networks (CNN):** Convolutional Neural Networks (CNNs) are a very popular choice in the deep learning community right now. Filters or kernels form the basic building block of CNNs and they are learnt automatically without explicit intervention. These kernels extract features from the input data with the help of convolution operation. Like RNNs, CNNs also follow the concept of parameter sharing i.e. a single filter is applied across different parts of an input to generate a feature map. This method was used by Guan *et al.* [33].

B. TYPE OF DATASETS

Figure 2 shows taxonomy based on datasets used for developing childhood and adolescent obesity models. The earlier models on childhood obesity mostly used cohort studies which are a type of longitudinal study. Later in 2010 along with cohort studies, EHRs, as well as image datasets, have been used for obesity prediction mainly for two reasons: Firstly, with the introduction of EHRs, streamlined data is available to be used, and secondly, efficient deep learning models could be utilized to tap into these massive EHR and image datasets.

1) **Surveys:** Studies in [6]–[8], [10], [26], [27], [32], [37], [39] used data that is collected using survey from a selected sample of individuals. Data was collected through various means such as in-person interviews, telephone interviews, mailed and online questionnaires.
2) **Cohort Study:** Most of the earlier studies on predicting childhood and adolescent obesity using machine learning models use existing cohort studies as datasets. Cohort studies are an example of longitudinal studies in which research participants are followed over a period of time. Usually, some of the participants are exposed to a specific risk factor and then the impact of this variable or risk factor is studied by monitoring the outcome over a period of time. In this way, cohort studies help researchers in understanding the factors impacting the likelihood of the development of a disease. This type of cohort study is called a prospective cohort.
   As these studies often span years, there's a high chance that participants may drop out thereby increasing attrition bias. In cohort studies where the participants know that they are being observed and studied, they may act differently than they normally do. This type of behavior is known as the 'Hawthorne effect' [43] and can affect different habits such as dietary, hygiene practices, etc. Another cohort very frequently used is the retrospective cohort in which the participants already have a known disease or outcome. In this type of cohort, the researcher starts the study when the follow-up has already been completed. Archived or self-report data is investigated to find out if the risk of disease was different between exposed and non-exposed groups. There are a few cons associated with this method as well. Retrospective cohort studies might have a bias because of sampling methods (more chance of missing data). Another disadvantage of using retrospective data is that the data might be of poor quality as it was not designed for that particular study.
   Most of the papers covered in this survey use cohorts that were not designed for studying overweight/obesity but they did contain variables such as anthropometric, behavioral, demographics, etc. that are known to be associated with obesity. The majority of the studies that focus on finding patterns and associations have used cohorts [9], [36] [5], [13], [38], [40].
3) **Electronic Health Records:** Electronic Health Records contain data on a patient's medical history, prescriptions, allergies, treatment data, radiology images, etc. consolidated into one single digital database. EHR data is updated in real-time and therefore it can be accessed for descriptive or predictive analysis at all times. EHR contains both structured and unstructured data. HIMSS (Healthcare Information and Management Systems Society)[9] defines structured data as that can be "organized into specific fields as a part of a schema, with each field having a defined purpose."

[9]https://www.himss.org/

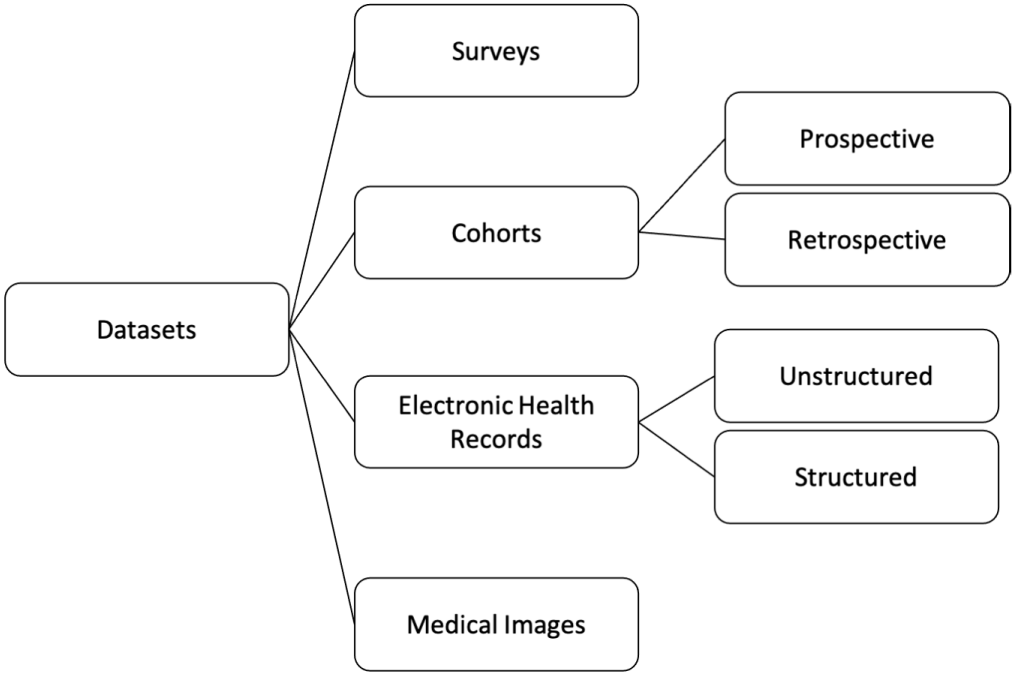

FIGURE 2. Categorization of the datasets used in existing literature on obesity based on the type: Surveys, Cohorts, Electronic Health Records, and Medical images.

This can include patient name, contact information, demographic information, lab values, etc. On the other hand, unstructured data is defined as data that ''cannot be easily organized using pre-defined structures.'' Natural Language Processing is used to process unstructured text data.

Nau *et al.* [11] used EHR for identifying a combination of community features that are most important predictors of obesogenic and obesoprotective environments while others used EHR for training models for identifying overweight or obese children [24], [25], [28], [29], [34], [35], [41].

4) **Image Datasets:** There has been a steady increase in the use of images for disease prediction with the introduction of deep learning-based Convolutional Neural Networks. As healthcare image data is very sensitive and it is very difficult to acquire them for research studies. Therefore, there are only a handful of studies that use image data for obesity prediction for adults. Few adult obesity prediction or rather diagnosis studies have used face images but for children, there are no such studies because of the unavailability of the publicly available dataset. For children, obesity/overweight studies mostly use MRI (Magnetic Resonance Imaging) image datasets [30], [33]. Figure 3 shows sample images from different types of datasets used in obesity studies.

C. OUTCOME PREDICTED

For most of the ML models, the outcome predicted is: overweight, obesity, or both combined. Table 5 summarizes the existing literature based on the predicted outcome. There is no one universal criterion for classifying children into these three classes but the three most commonly used criteria are: World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and International Obesity Task Force (IOTF). These three classification criteria differ in certain aspects [47]–[49]. Gonzalez-Casanova *et al.* [49] found a lack of consistency among the three systems in assessing overweight and obesity in children and adolescents. Depending on the system used appreciably different estimates of overweight and obesity with age and sex were observed in the nationally representative Colombian National Nutrition Survey of 2005. Studies have also modeled BMI percentile [9], BMI progression [9], [30], and even raw BMI [40]. In the case of raw BMI modeling, it is done when there's not much difference between the ages of the children as BMI calculation for children and adolescents is age and gender-specific. These can be broadly divided into two categories: Categorical and Numerical. Figure 4 depicts this categorization.

D. FACTORS USED IN MACHINE LEARNING MODELS

Factors from different domains have been used to study their associations with obesity as well as in predicting obesity. On a broad level, these factors can be divided into four categories but these shouldn't be treated as the ultimate division as certain factors overlap and can be included in more than one category. Figure 5 shows the taxonomy based on the factors used for predicting childhood and adolescent obesity.

1) **Individual:** Individual characteristics of the child such as age, weight, height, sex, birth weight, and height are covered under this category. Psychological and

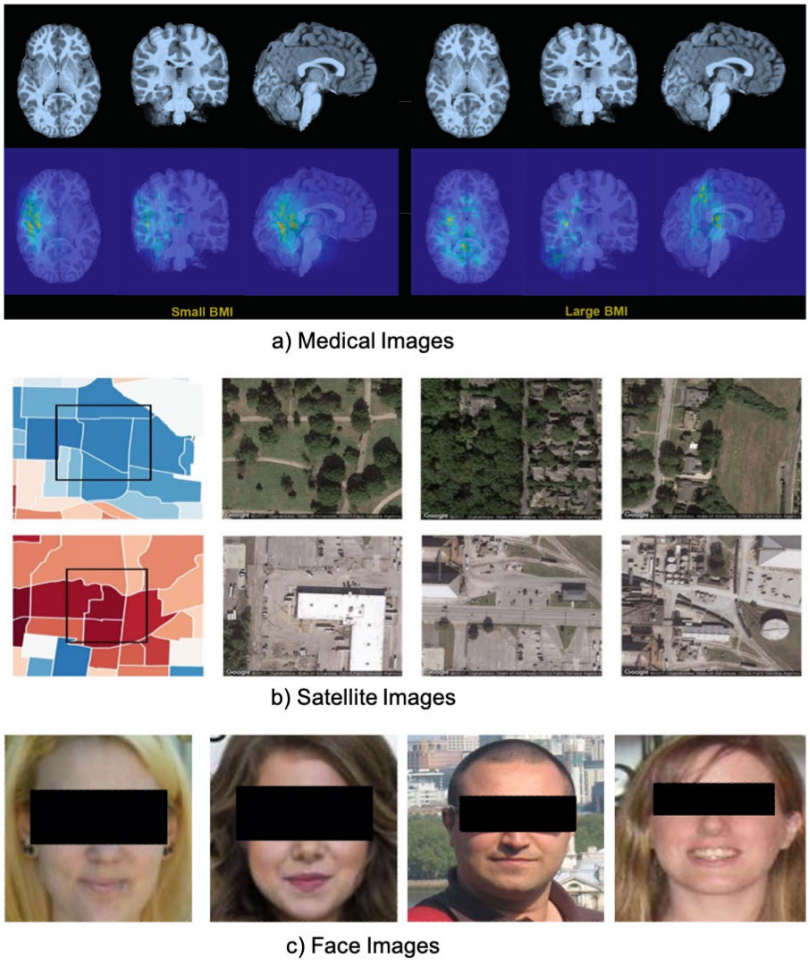

FIGURE 3. Different types of image datasets used in Obesity studies in general. (a) Example medical images from Adolescent Brain Cognitive Development (ABCD) dataset [44] (b) example satellite images used in [45], and example face images from VisualBMI dataset [46].

TABLE 5. Grouping of studies according to the outcome predicted i.e., Categorical or Numerical.

Type	Reference to Existing Studies
Categorical (Obese, Overweight, Normal)	[36] [9] [37] [10] [38] [39] [12] [13] [6] [7] [8] [24] [25] [26] [27] [28] [29] [32] [41] [34] [35]
Numerical (BMI value)	[9] [5] [40] [28] [30] [31] [32] [33]

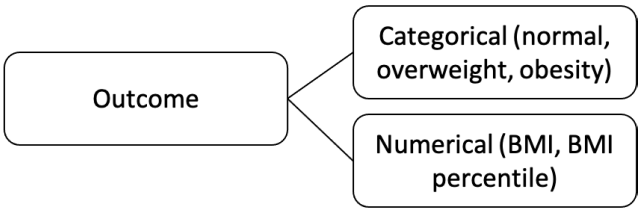

FIGURE 4. Outcomes predicted using childhood and adolescent obesity model was either Categorical or Numerical.

behavioral factors such as early life trauma, lack of social networks, academic performance may also be included in this category. All of the studies in this survey included variables from this category [5]–[13], [24]–[41].

2) **Familial:** This refers to the factors related to the home and family of the child such as BMI of parents, socioeconomic status of the family, working parents, ethnicity of parents, etc. Included in this type of factor is also the pre-pregnancy health condition as well as the pregnancy lifestyle of the mother. Genetics is one of the important factors considered as well. Lot of research has been done to study the causal effect of genetics in relation to obesity [50], [51]. Most of the studies included data from familial category [5]–[9], [12], [13], [24]–[26], [29], [32], [34], [36]–[40].

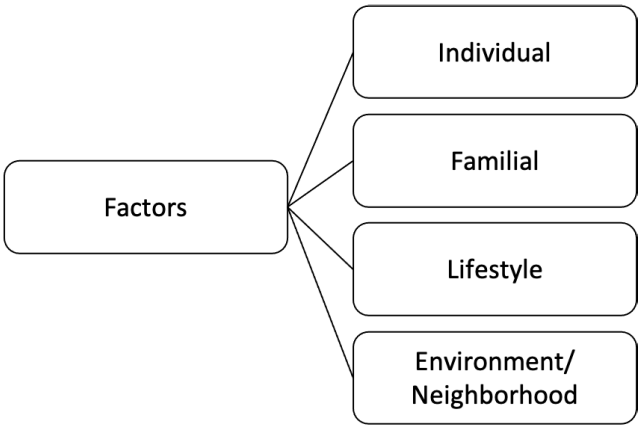

FIGURE 5. Factors used for childhood and adolescent obesity prediction models could be categorized into Individual, Familial, Lifestyle, and, Environment/ Neighborhood.

3) **Lifestyle:** This category includes sleep duration, physical activity, screen time, smoking habits, etc. Lifestyle factors up to a certain age are influenced by the activities of the parents. Young children depend on their parents for nutritional intake, and parents, in turn, provide food depending on the financial condition and also the amount of time that they can invest in preparing healthy and nutritious food for the child. The following studies had variables from lifestyle domain: [5]–[10], [13], [26], [27], [32], [37]–[40].
4) **Environment:** The environment can be of different types: home, school, childcare, school neighborhood, and home neighborhood. Food preferences and eating habits of children are often established when the child is young. Home, school, childcare, and neighborhood contribute a lot to the development of the eating habits of children. For example, in childcare centers, meals can either be provided by the center or by the parents themselves. Environments pertaining to childhood obesity can be of different types - home, childcare (type: informal versus formal care, duration - number of years spent in childcare, intensity - number of hours per week and timing - age of onset of childcare), school, neighborhood environment, etc. The environments that aid in the development of obesity are called obesogenic and those that prevent it are obesoprotective environments. Only a few studies included environment variables [5], [11], [12], [38], [39].

In addition to the aforementioned factors, certain factors may also exacerbate or mitigate the influence of other factors in the development of childhood obesity, therefore these correlated factors are often used in combination to study the influence.

E. PURPOSE

ML child and adolescent obesity models can be broadly divided into two categories based on the purpose they serve (a) identifying potential risk factors contributing to obesity, and (b) those focused on the prediction to improve various metrics such as accuracy, sensitivity, and specificity. There are a good number of machine learning studies that stress on both, the predictors or risk factors associated with child and adolescent obesity [5], [9]–[13], [36]–[40] and predicting childhood and adolescent obesity [6]–[8], [24]–[35], [41]. These two approaches have been discussed in detail as it is extremely important to understand the generalization ability of existing obesity prediction models (Prediction-focused) along with the important factors and associations attributed to obesity need to be determined (Predictor-focused). We will discuss both these approaches in this section. Figure 6 shows the division based on the purpose of developing the models.

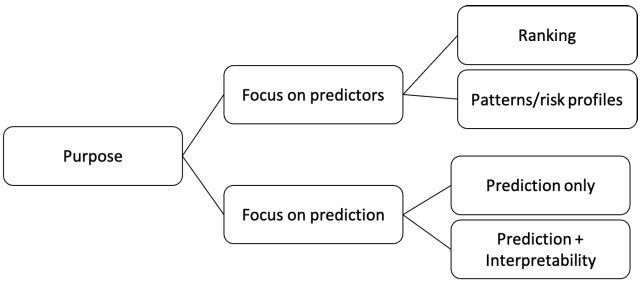

FIGURE 6. Categorization of ML models for obesity based on the purpose they serve i.e., (a) identifying potential risk factors (key determinants) contributing to obesity, and (b) predicting obesity.

1) FOCUS ON PREDICTORS

Most of the studies that cover the association of variables with overweight/obesity are statistical in nature. In this section, we will discuss 11 papers that focus on finding patterns in data and the associations with a child's weight status. These types of studies can be further subdivided into two types that focus on - (a) predictors (attributes) ranking, and (b) finding patterns or building risk profiles.

In the next few paragraphs, we will briefly discuss studies under both categories and then analyse them based on certain factors such as diversity in datasets, factors, and the methods used to develop the models.

a: PREDICTOR RANKING

Predictors (attributes) ranking for predicting childhood and adolescent obesity is done to find out the most important factors contributing to the predicted outcome. The output variable BMI or BMI categories have been used solely to find the predictor importance. The studies on predictor ranking most commonly used Decision Tree [5], [9]–[13] and Gradient Boosting Methods (GBM) [40]. Predictor ranking models can be used for intervention programs, diagnostic and prediction tools.

Rehkopf *et al.* [9] focused on finding the best predictors associated with female BMI change from age 9 to 19 utilizing already 41 established risk factors from an initial list of 142 potential risk factors from the NHGS (NHLBI Growth and Health Study) dataset (with a sample size of 2150).

Random Forest was used for ranking of predictor variables based on the mean decrease in the obesity prediction accuracy. 20 out of 41 predictor variables consistently and significantly associated with BMI percentile ($p < 0.05$) in the decreasing order of importance were: body dissatisfaction, drive for thinness, physical appearance, income, parent education, perfectionism, bulimia scale, anxiety, emotional eating index, interoceptive awareness, ineffectiveness, number of siblings, race, eating breakfast, time to eat, parent depression, parent BMI, frequency of snacks, soda on the table and self-worth.

Life survey data from the 2011 National Youth Risk Behavior Survey (YRBS) was used by Pochini *et al.* [10] in 2014 to study the risk factors associated with overweight and obesity among adolescents. The sample size was 15,425 and the mode of data collection was questionnaires. This dataset did not have any information on anthropometric measurements or medical examinations. Only the variables related to the lifestyle of high school students were used as factors in the study. Logistic Regression (LR) and Decision Trees (DT) were used to classify the target variable. The following 9 binary factor variables were used: whether the student 1.) had 1+ drinks past 30 days 2.) drank fruit juice past 7 days 3.) drank soda 1+ times/day past 7 days 4.) ate breakfast on all of the past 7 days 5.) gets 8+ hours of sleep 6.) used any tobacco past 30 days 7.) ate 5+ fruits/vegetables/day 7 days 8.) was physically active 7 of past 7 days 9.) watches TV or plays video/computer games over 3 hours each day. Frequently doing physical activities and having breakfast every day were found to be protective factors against being overweight and obese by both decision tree and logistic regression. On the other hand, smoking and drinking sugar-sweetened beverages were found to be associated with an increased risk of obesity. The two categories overweight and obesity were mutually exclusive in this study.

Electronic health records from the Geisinger Health System of children aged 10–18 years collected during 2010 (the sample size is equal to $22,497$) were used to study obesogenic and obesoprotective environments with the help of Conditional Random Forest (CRF) by Nau *et al.* [11] in 2015. The aim was to find pre-established risk factors that render a community obesogenic or obesoprotective. Obesogenic and obesoprotective communities were classified based on the distribution of average BMI in each community. 44 features linked to obesity from multiple domains (social factors, food availability, and physical activity-related features including land use characteristics and physical activity establishments) were collected from secondary data sources. Of all the 44 features, only 13 (mix of social, food, land use, and physical activity features) contributed consistently to the classification.

Hinojosa *et al.* [12] aimed at specifically studying social school environments and neighborhoods in 2018. The authors were of the opinion that health tracking via public schools can be done to identify at-risk populations, the reason being the substantial amount of time spent in school. Schools can also work as a platform via which a larger population can be reached out for preventive measures. The dataset for this study was gathered from participants of the Physical Fitness Test (PFT) 2003 through 2007 (5th, 7th, and 9th graders). In addition to individual, race, and gender factors, which are already known to be associated with obesity, certain school and neighborhood factors were found to be positively and negatively associated. Violent crime, English learners, socioeconomic disadvantage, fewer physical education (PE) and fully credentialed teachers, and diversity index were found to be positively associated whereas academic performance index, PE participation, mean educational attainment and per capita income were negatively associated with obesity.

Decision tree models were used by Lee *et al.* [13] (2019) to predict two BMI categories: Overweight and Normal-weight between 24 and 80 months. The dataset for this study was from a South Korean longitudinal cohort (Korean National Health Insurance database) which contains qualification data, national health checkup data (for infants, children, and adults), and health insurance claims data. A total of 21 predictors from 4 different categories (Socioeconomic status, maternal, paternal, and child-related factors) were used in this study. The best predictors were maternal history of obesity before pregnancy and paternal obesity, and the second-best was socioeconomic status.

Gray *et al.* [5] used Adolescent Brain Cognitive Development (ABCD) dataset to study factors associated with obesity belonging to four different groups: demographics, psychological health, lifestyle behaviors, and cognition. Ridge, LASSO, and ElasticNet regression were used to predict the percentage of 95th BMI percentile. This study also analysed race/ethnicity-specific models. The best performing model was LASSO which selected 25 features and the most important features were no stimulant medication use, Hispanic ethnicity, nonwhite race, male sex, lower socioeconomic status (SES), and unmarried caregiver. Besides stimulant and demographic factors, other factors were: attention problems and matrix reasoning (inversely associated), social problems, screen time, and reward responsiveness. For sex-based analysis, LASSO did best for girls, and for boys, ElasticNet performed the best. The authors also performed exploratory analyses on Hispanic, Black non-Hispanic, and White non-Hispanic participants. White non-Hispanics had the lowest percentage of sex and age-specific BMI percentile followed by Hispanic, and Black, non-Hispanic had the highest.

Marcos-Pasero *et al.* [40] used machine learning models to rank predictor variables from different domains for BMI estimation in 6-9 years old in 2010. The sample size consisted of 221 children from Spain and the number of multi-domain predictor variables was 190. The predictor variables were grouped by domain: characteristics of school children (3); genetics (1); physical and leisure activities (24); diet, food, and nutrients (80); risk factors of pregnancy and birth (39); social, health and demographic factors (43). Both Random

Forest (RF) and Gradient Boosting Methods (GBM) were used for BMI prediction and the ranking of predictor variables was done through an iterative process that involved permutation and multiple imputations. The top four predictor variables in decreasing order of importance were: Familiar Nutri-status perception (Perception of the person completing the questionnaire about child's nutritional status), Relation TEI-TEE (%) (Percentage of difference between Total Energy Intake (TEI) and Total Energy Expenditure (TEE)), BMI of the mother, and BMI of the father.

b: PATTERNS OR RISK PROFILES

In these studies, groups which are at greater development of obesity are determined. Mostly tree-based methods have been used for these studies as well [36]–[39].

Kitsantas and Gaffney in 2010 [36] focused on building the risk profile of overweight/obese children using decision trees based classifiers. Data from the ECLS-B (Early Childhood Longitudinal Study, Birth Cohort) cohort was used in this study and only non-Hispanic black, non-Hispanic white, and Hispanic mothers were included in the analysis (with a sample size of 6540). The classification results revealed four preliminary risk profiles: Group 1- Children overweight/obese at 2 years old were at high risk. Group 2- Children born to mothers with a normal pre-pregnancy BMI were at higher risk if they belonged to lower socioeconomic status and had $\geq$ 4000g birthweight. Group 3- Normal weight children (at 2 years) whose mothers were overweight/obese and of Hispanic origin were more likely to be overweight/obese at age of four years than those children who did not have that profile. Group 4- Children of overweight/obese (before pregnancy) white or black mothers, had been breastfed 2.5 months or less, and had less than two prior pregnancies were at an increased risk of developing overweight/obesity. Out of the 12 features, seven features i.e., overweight/obesity at age two, mothers overweight/obese before pregnancy, low Socioeconomic status (SES), birthweight $\geq$ 4000g, maternal Hispanic ethnicity, mother with less than two previous pregnancies, and with less than 2.5 months of breastfeeding were associated with overweight/obesity at 4 years. Maternal age, marital status, smoking status, gestational age at birth, and child's sex did not contribute significantly to the prediction of obesity.

Lazarou *et al.* [37] used the CYKIDS (Cyprus Kids Study) dataset from Cyprus for analyzing dietary patterns. The sample size consisted of 634 subjects in the age group 9-13 years for those with height, weight, and waist circumference measurements available. The data was collected through questionnaires distributed to children and certain variables such as parents' educational level, income, and occupation were collected through questionnaires distributed to parents. The Food Groups Frequency Questionnaire (FGFQ) which is a 4-point scale evaluating the frequency of consumption of food groups was used for analysis. 15 variables were used: fried food, grilled food, fish and seafood, meat, etc. Decision tree (C4.5) was the main method of analysis and logistic regression (PCA - Principal Component Analysis was used to extract the main factors of diet composition from the 15 variables) was used for comparison. Results for both the methods differed vastly. Decision tree revealed that fried food, delicatessen meat, sweets, junk food, and soft drinks were associated with an increased risk of obesity (detrimental factor). Further, frequent consumption of fish is more likely to be associated with normal weight (protective factor). PCA revealed that increased consumption of milk and cereals was related to lower obesity levels, but only in girls.

Van Hulst *et al.* [38] used CART (Classification and Regression Tree) to study individual, familial, and neighborhood environment factors using QUALITY (Quebec Adipose and Lifestyle Investigation in Youth) dataset (512 non-Hispanic white 8-10 years old) in 2015. This study used a cohort that was collected to study the natural history of Obesity and Cardiovascular risk in Quebec youth. It is one of the few studies that used datasets specifically gathered for studying obesity. The variables used for this study were: sugar-sweetened beverage intake, meeting Physical Activity guidelines (Individual), number of BMI-defined obese parents, number of parents with abdominal obesity, parental education, household income (familial), and disadvantage, prestige, and presence of $\geq$ 1 park, fast food restaurant, and convenience store (neighborhood). The interaction between various pre-established factors was studied.

Wiechmann *et al.* [39] published a paper in 2017 that aimed at understanding the strongly correlated factors with childhood obesity among 2-5 years old children of Hispanic ethnicity. This study collected its own dataset (238 children with Hispanic parents) instead of using some pre-existing study to decipher social and epidemiological family conditions associated with barriers that challenge healthy eating. The following variables from 10 different domains were used with the C4.5 decision tree algorithm: demographics, caregiver feeding style, feeding practices, home environment, dietary information, beverage consumption, social support, family life, integrated behavior model, and spousal support.

Table 6 summarizes the different studies that focus on finding the most important factors for predicting overweight and obesity. Majority of the studies used datasets from United States [9], [36] [5], [10]–[12], [39], and the rest of the studies used datasets from Cyprus [37], Canada [38], South Korea [13], and Spain [40]. China, India, the USA, Indonesia, and Brazil are the top five countries that have been projected to have the highest number of children and adolescents with obesity by 2030. Amongst these five countries, datasets from only USA have been explored extensively using Machine Learning algorithms. Existing studies suggest that obesity prediction models are not equitable across age groups, ethnicity/race, and region.

Collectively, the papers exhibit diversity in factors but individually they have factors from a limited number of domains. Rehkopf *et al.* [9] used familial, dietary, behavioral, psychological, and social risk factors; Kitsantas and Gaffney [36] used maternal and child factors; Lazarou *et al.* [37] focused on dietary factors; Pochini *et al.* [10] on lifestyle factors;

TABLE 6. List of studies on machine Learning models that focus on ranking predictors for obesity prediction.

Article	Dataset (Country, Type)	Original dataset purpose	Dataset Size, Method, Predictors	Outcome Studied (Standard used)	Key Determinants
Kitsantas et al. [36] (2010)	ECLS-B (US, Cohort)	To provide information on health, development, care and education.	6540, CART, 12	Normal, Overweight, Obesity, 0-4 years (CDC)	1. Overweight/Obese at 24 months. 2. Elevated birthweight. 3. Low/middle SES. 4. Maternal Hispanic ethnicity. 5. Low breastfeeding duration. 6. Low parity 7. High birthweight
Rehkopf et al. [9] (2011)	NHGS (US, Cohort)	To investigate racial differences in dietary, physical activity, family, and psychosocial factors associated with the development of obesity from pre-adolescence through maturation between African-American and White Girls.	221, RF, 190	BMI Percentile, Overweight, Obesity, 9-19 years (CDC)	BMI percentile: 1. Body dissatisfaction, 2. Drive for thinness, 3. Physical appearance, 4. Income, 5. Parent education Overweight/obesity: 1. Income, 2. Ineffectiveness 3. Race
Lazarou et al. [37] (2012)	CYKIDS (Cyprus, Cross-sectional Survey)	To assess dietary patterns and behaviors	634, DT, LR, 5	Overweight 10-12 years (IOTF)	1. Frequent consumption of fried food, delicatessen meat, sweets, junk food, and soft drinks is associated with an increased risk for obesity. 2. Frequent consumption of fish is more likely to be associated with normal weight (i.e. Protective factor).
Pochini et al. [10] (2014)	YRBS (US, Surveys)	Monitors six categories of health-related behaviors- 1. Unintentional injuries and violence. 2. Sexual behaviors 3. Alcohol and other drug use 4. Tobacco use 5. Unhealthy dietary behaviors 6. Inadequate physical activity also measures the prevalence of obesity.	15425, DT, LR, 9	Overweight, Obesity 14-18 years (CDC)	1. Frequently doing physical activities and having breakfast everyday were found to be protective factors. 2. Smoking and drinking sugar-sweetened beverages were found to be associated with an increased risk of obesity.
Nau et al. [11] (2015)	Geisinger Health System (US, EHR)	Healthcare provider in Pennsylvania. Their EHR database contains information on more than three million patients.	22497, CRF, 44	Community-level BMI-z score 10-18 years (CDC)	1. Unemployment 2. Population density 3. Social disorganization. 4. Less than high school degree 5. Population change 6. No car ownership 7. Public assistance 8. Snack stores 9. Fast-food chain 10. Community type 11. Poverty 12. Vehicle miles traveled 13. Physical activity

TABLE 6. *(Continued.)* **List of studies on machine Learning models that focus on ranking predictors for obesity prediction.**

Article	Dataset (Country, Type)	Original dataset purpose	Dataset Size, Method, Predictors	Outcome Studied (Standard used)	Key Determinants
Van Hulst et al. [38] (2015)	QUALITY (Canada, Cohort)	To investigate the natural history and determinants of childhood obesity and its cardiometabolic consequences.	512, CART, 11	Obesity 8-10 years (CDC)	1. More than 1 parent with obesity 2. Not meeting physical guidelines 3. 2 parents with abdominal obesity 4. Average to high neighborhood disadvantage. 5. Zero neighborhood parks. 6. No access to a convenience store.
Wiechmann et al. [39] (2017)	Collected own dataset using Questionnaires and Interviews (US)	Dataset gathered for this study.	238, DT, NA	Overweight, 2-5 years	Overall 1. How important the participant believes that she has to be especially careful to make sure that her child eats enough. 2. How important the participant's husband/partner eats healthy meals with them.
Hinojosa et al. [12] (2018)	Physical Fitness Test through FITNESSGRAM (US)	To help students in starting life-long habits of regular physical activity.	5265265, RF, 124	Obesity 5th, 7th, 9th graders (CDC)	1. Low academic performance. 2. High %age of individuals learning English as a second language. 3. Young age. 4. Having more crime that is violent. 5. High diversity index. 6. More teachers per student. 7. Hispanic ethnicity and male.
Lee et al. [13] (2019)	KNHI (South Korea, Cohort)	Health insurance data.	1001775, DT, 32or33(check)	Obesity 24-80 months (KCDC)	1. Parental obesity history 2. SES 3. Gestational hypertension and diabetes. 4. Older pregnancy. 5. Drinking during gestation. 6. Depression after delivery 7. Non-compliance with exclusive breastfeeding. 8. Sugar-sweetened beverage intake more than 200ml per day. 9. Irregular breakfast consumption
Gray et al. [5] (2019)	ABCD (US, Cohort)	Study of brain development and child health.	3847, Ridge, LASSO, ElasticNet, 43	% of 95th BMI percentile, 9-10 years (CDC)	1. Stimulant medications. 2. Hispanic ethnicity. 3. Non-white race. 4. Male sex. 5. Lower income. 6. Unmarried caregiver.
Marcos-Pasero et al. [40] (2021)	GENYAL (US, Interventional Study)	Detection of Genetic Polymorphisms associated with Obesity and its complications in schoolchildren within the Madrid community, and evaluation of health actions for reducing the Risk	221, GBM, RF, 190	Raw BMI, (IOTF, WHO)	1. Familiar nutri-status perception. 2. Relation TEI-TEE (%). 3. BMI of the mother. 4. BMI of the father.

NA = Not available

Nau *et al.* [11] on community factors (obesogenic and obesoprotective); Van Hulst *et al.* [38] used individual, familial, and neighborhood factors; Wiechmann *et al.* [39] used social and epidemiological factors but only Hispanics were studied; Hinojosa *et al.* [12] focused on school environments; Lee *et al.* [13] inter-generational factors (another name for using child, parental, socioeconomic status factors, etc.); Gray *et al.* [5] used demographic, psychological, behavioral, and cognitive variables; Marcos-Pasero *et al.* [40] used genetic, nutritional, exercise, social and health, lifestyle, birth, and pregnancy variables. Psychological, genetic, and childcare environment are three important factors that are under-explored.

The only study in [11] focused on the community as a whole as contributing to the development (obesogenic) or prevention (obesoprotective) of obesity. This type of study can be very useful in curbing obesity in areas where people of a single ethnicity/race reside. So instead of targeting individuals, the whole community can benefit from preventive measures which would bring down the cost associated with prevention strategies to a certain extent. Almost all of the studies except the one in [5] used some form of Decision Tree based model. The way decision trees are represented, different groups and risk profiles of children and adolescents susceptible to overweight and obesity can be identified. Tree-based methods are non-parametric methods and predictor importance is an inherent aspect of this model with the top-most node being the most important predictor. Ridge, LASSO and, ElasticNet were also used in one of the studies as they provide feature selection capability. In terms of race and ethnicity, these studies do not add anything new to the existing research. It is an already established fact that Hispanic and Black children and adolescents have a higher obesity prevalence rate. The existing studies merely confirmed this already established fact.

2) FOCUS ON PREDICTION

The studies discussed in this section stress more on prediction, so in addition to the factors, techniques for feature selection as well as the models used would also be included in the discussion. A total of 16 papers were included in this section, and we will briefly discuss each of these papers.

a: ONLY PREDICTION

The main focus of this kind of study is to accurately identify overweight or obese children.

In 2012, Adnan *et al.* published 3 research papers [6]–[8] on predicting child's weight status (normal-weight, overweight and obese) in quick succession using a Malaysian cohort of 140 subjects 9–11 years old children. In [8], the aim was to identify parameters for childhood obesity prediction using data mining methods that would increase the prediction accuracy. From the survey data collected by the authors, children factors (catch-up growth, adiposity rebound, premature birth, gender), lifestyle factors (duration of breastfeeding, duration of sleep, eating junk food, eating fried food, eating fruit, eating snacks in front of TV, duration of watching TV, eating warm meals for supper, physical activity, eating soup and sandwich bought outside home, eating snacks and chocolate bought outside home), and family/environment factors (mother BMI, father BMI, parental overweight/obesity, and number of children) were found to be important factors. Using the aforementioned selected features, Naive Bayes showed that using the selected features, an improvement of 21% in accuracy was obtained compared to parameters used in Zhang *et al.*'s study [17]. In 2012, Adnan *et al.* [6] used Genetic Algorithm (GA) optimization technique to mitigate the problem of zero value parameter when using Naive Bayes on a sample size of 180. Using GA, optimization prediction accuracy increased by 92%. In [7] two variable importance techniques for feature selection, CART and Euclidean distances were used. CART selected 8 features: watching-TV, father-BMI, fried-food, mother-BMI, sibling, fruit, average-sleep, and physical-activity and obtained better accuracy (normal: 37.5%, overweight: 95.83%, obesity: 75%) for obesity compared to classification using all the variables and euclidean distance (normal: 37.5%, overweight: 83.3%, obesity: 83.3%).

Dugan *et al.* [24] predicted obesity after 2 years using the data collected before the age of 2. The collected dataset is multi-ethnic and the majority of the subjects belong to minority and low-income groups. The sample for the study consisted of 7519 children and the criterion for selection was that at least one clinic visit before the second birthday and at least one BMI percentile after the second birthday. Six models were used: Random Tree, Random Forest, ID3, J48 (Java implementation of C4.5 algorithm), Naive Bayes, and Bayes Net (BN), and the full predictor set consisting of 167 variables. Three kinds of analysis were performed - entire dataset no resampling, entire dataset with resampling, and full fit and feature selection using an iterative process. The best accuracy of 85% was obtained by ID3 and the feature size went down from 167 to 87. Being overweight before 24 months was an important predictor in predicting obesity. Being very tall before 6 months stood out as a protective factor. In the analysis, the authors observed that belonging to the minority class increases the chance of being overweight after 2 years if the child was not overweight before 24 months.

Lingren *et al.* [25] used structured and unstructured data from EHR to identify children aged 1-5.99 years with severe early onset of childhood obesity. Rule-based and Machine Learning algorithms were developed using two EHR databases from Boston children's Hospital (BCH) and Cincinnati Children's Hospital and Medical Center (CCHMC). The predictor variables used were demographics, anthropometrics, ICD-9 diagnosis codes, and medications from structured data and features extracted using NLP from unstructured data. Rule-based algorithms performed better than Machine Learning (SVM-Support Vector Machine and NB-Naive Bayes) ones but one advantage of ML was the balancing of PPV (Positive Predictive Value) and sensitivity for selecting variable sets. This study is different from other

studies for two reasons. First, it uses genetic data for modeling and second it utilizes the unstructured data as well with the help of Natural Language Processing (NLP).

Abdullah *et al.* [26] used different machine learning methods (Bayes Net, Naive Bayes, Decision Tree, ANN, and SVM) to predict obesity amongst 12 year olds in a Malaysian cohort. Predictor variables from three domains: socio-demographic, physical activity, and diet were collected with the help of questionnaires. Different feature selection techniques (best first, genetic search, greedy step-wise, and linear forward) were explored to get an optimal set of attributes which was then evaluated on the aforementioned classifiers. Feature selection method of consistency combined with linear selection obtained the best accuracy of 82.72% for the Decision Tree (J48) classifier. A total of 29 features were used, some of which are: district, education level mother, age of mother, marital status of mother, education level father, age of the father, marital status of the father, and family income, etc.

Zheng and Ruggerio [27] used both risk and protective factors to build an adolescent obesity prediction model on YRBBS (sample size of 5127) 2015 survey data for the state of Tennessee. The following four machine learning models were compared and evaluated: logistic regression, decision tree, weighted k-nearest neighbor, and artificial neural network. The predictors were divided into three categories: energy intake (eating fruits/vegetables/breakfast, drinking soda/soft drinks in the past week), physical activity (at least 60 minutes of physical activity, and physical education classes in the past week), and sedentary behavior (hours of sleep during school days, time spent watching TV, using computer for non-school-related tasks or playing video games). The accuracy of the four models in the decreasing order were: 88.82% (weighted kNN), 84.22% (ANN), 80.23% (IDT-Improved Decision Tree), and 56.02% (Logistic Regression). The results suggested that engaging in physical activity and having breakfast everyday significantly reduced the risk of obesity whereas excessive computer use and consuming sugar-sweetened beverages increased the risk of developing obesity.

Hammond *et al.* [29] used EHR data from a safety health system in New York that contained data from 52, 945 children and 36244 mothers. Neighborhood data was included with the help of the 2015 American Community Survey 5-year estimates. Feature engineering was done to generate 19, 290 predictor variables belonging to the following categories: diagnosis, lab, medication, gender, ethnicity, race, vital, number of visits, zip code, census, maternal and newborn diagnosis, maternal ethnicity, primary and secondary insurance, maternal (race, nationality, language, marriage status, birthplace, delivery age, lab history, and produce history). Both regression and classification techniques were used for obesity prediction. For regression (LASSO, RF, GBM), median BMI was normalized and children were classified as obese or non-obese based on a threshold. For classification (logistic regression with L1 loss, RF, and GBM), class probabilities were used for predicting a child as obese/non-obese. For evaluating classification performance Area Under the ROC (Receiver Operating Characteristic) curve (AUC) which shows the performance of a classification model at different thresholds, was used. Separate models were developed for girls and boys. The best models have an AUC score of 81.7% for girls and 76.1% for boys.

Park *et al.* [30] used resting-state functional magnetic resonance imaging (rs-fMRI) to develop models that predict BMI progression of adolescents. The dataset consisted of 76 individuals from the Enhanced Nathan Kline Institute Rockland Sample (NKI-RS) database. The average age was 11.94 years and consisted of white and African preadolescents. 379 Degree-Centrality (DC) values of different parts of the brain were extracted using brain fMRI (sub-cortical volume and cortical surface both were used) from the first visit. LASSO used for prediction and feature selection retained only 6 DCs. The model obtained an Intra-class correlation of 0.70 for BMI progression and 0.98 for BMI, and AUC score of 0.82.

Singh and Tawfik [31] used a UK cohort for studying adolescent obesity. This dataset Millennium Cohort Study (MCS) followed every child born in the years 2000 and 2001. For predicting BMI at 14 years of age, the methods used were Linear SVM, linear regression, and ANN. The best accuracy was obtained by the ANN model (93.4%).

Kim *et al.* [32] studied the factors affecting adolescent obesity using a South Korean dataset in 2019. The authors used raw data from the 2017 Korean Youth Health Behavior Survey conducted by the Korean Centers for Disease Control & Prevention (KCDC). Three BMI categories were predicted: underweight, normal, and overweight with the help of 19 predictor variables (sample size is 11206). General Bayesian Network (GBN) was used for prediction and the results were compared with other ML methods. The best accuracy and AUC score were 53.7% and 0.76, respectively.

Only 200 data points and 11 predictors from the CHICA (Child Health Improvement through Computer Automation) dataset were used by Chatterjee *et al.* [35] to predict obesity in 3-5 years old children. SVM, KNN, and ANN were trained and evaluated on this data and the maximum accuracy of 96% was obtained by ANN.

b: PREDICTION AND INTERPRETABILITY/EXPLAINABILITY

Papers that covered these concepts were: [28], [33], [34], [40], [41]. The more the model is interpretable, the easier it is to identify cause and effect relationships between the inputs and outputs. On the other hand, explainability is associated with the internal logic and mechanisms of the model [52]. Decision Tree, Linear Regression, and Logistic Regression models can be seen as more interpretable than other models such as Random Forest and Convolutional Neural Networks. For deep learning models, the concept of saliency in images which refers to unique features, such as pixels or resolution of the image in the context of visual processing, is used for explaining predictions of these models.

Gupta *et al.* [28] used a more modern approach for building a prediction model using EHR data in conjunction with a Recurrent Neural Network (RNN) architecture with Long Short-term memory (LSTM) cells for capturing time-series data of various visits of the patient. For static data, such as sex, race, ethnicity, and zip code, a separate feed-forward neural network was used. The dataset used was massive, consisting of about 44 million rows with 68029 unique patients acquired from the Nemours Children Healthcare System, which contains pediatric health data from Delaware, Florida, Pennsylvania, and New Jersey. Some of the most important features identified at the population level were: BMI and previous obese/non-obese label, childhood obesity, morbid obesity, and obesity. In addition to these, achondroplasia and anomaly of the chromosome and chromosome21 were also very important predictors. Accuracy for different window sizes was in the range 0.75-0.92 and AUC was in the range 0.80-0.97. Embedding weights on input layer and softmax on LSTM layers were used to calculate importance of features and attention weights for each input timestamp thereby providing interpretability at both feature and timestamp level.

Pang *et al.* [41] used XGBoost to predict obesity in 2-7 years old children using data from birth up to 2 years in 2019. To explain the output of their model, the authors used a framework called SHAP (Shapley Additive exPlanations). The data used for this study was taken from the Pediatric Big data Repository (PBD). This repository contains EHRs of patients who visited the hospital in person between 2009 and 2016. The cohort size used for this study consisted of $27,203$ unique individuals who were divided into two training sets each containing 40% of the data and a test set with 20% data. The number of predictor variables used was 102. The best model (XGBoost) obtained an AUC of 0.81. Precision, F1, accuracy, and specificity were 30.9%, 44.6%, 66.14%, and 63.27%, respectively for a recall of 0.8. Features impacting the prediction model the most were identified via SHAP - weight for height, height, weight, race, ethnicity, care-site, head-circumference, body temperature, respiratory rate. Further analysis revealed that Black or Hispanic/Latino had a higher chance of developing obesity.

Guan *et al.* [33] used 3779 T1-weighted brain MRI (Magnetic Resonance Imaging) images of adolescents aged 9–10 from the Adolescent Brain Cognitive Development (ABCD) study to predict BMI in 2020. For predicting BMI, the authors proposed a deep multi-cue regression network (DMRN) which had three modules (1) MRI feature encoding, (2) multi-cue feature fusion (MRI feature vector and waist circumference), and (3) BMI regression. This model was compared with two conventional machine learning methods - support vector regressor (SVR) using ROI (Region of Interest) features, SVR using Waist Circumference (WC) as feature, and two deep learning models- AlexNet, and VoxCNN. The proposed model obtained the lowest MAE (Mean Absolute Error) of 2.87 and MSE (Mean Square Error) of 14.01 compared to the other models. Authors also aggregated saliency maps learned by the model on the whole cohort (grouped by high BMI $>$ 40 and low BMI $<$ 10) to visualize obesity-related patterns in MRIs. Heatmaps generated from the network helped in identifying important imaging biomarkers that may be associated with obesity.

Rossman *et al.* [34] used data extracted from EHR to predict obesity in 5-6 years old children based on the data from the first 2 years of life. These EHRs (from 2002 to 2018) were obtained from Israel's largest healthcare provider, Clalit Health Services. First BMI trajectories of $38,2132$ adolescents were analyzed to obtain the time period during which the largest annual increase in BMI percentile takes place. The greatest increase in BMI was noted between 2-4 years of age. Then a prediction model was developed to identify children who were at high risk of obesity right before the BMI acceleration time period i.e., data from 0-2 years was used to predict obesity at 5-6 years ($n = 136196$). The model obtained an AUC of 0.80 and was temporally and geographically validated. Shapley values were used to identify the most important features which were anthropometric measurements of the child and family, ancestry, and pregnancy glucose.

Table 7 provides a summary of all the studies, including those using interpretability/explainability, that focus on prediction or rather improving metrics that help in determining the best fit model. Compared to previously discussed studies focused on ranking predictors in section III-E1.a, these studies offer a lot more variety in the methods/ML models used, but not much in the datasets used. Predictor-focused models used a lot more diverse dataset in terms of domain with the exclusion of images which is not the case with prediction-based models. This gap needs to be filled to develop robust models that could help in predicting obesity and curbing this epidemic.

The studies under this category used both parametric as well as non-parametric models for making predictions. In addition to Decision trees [24], [26], [27], Random Forest [24], and Gradient Boosting Methods [34], [41], these studies used LASSO [29], [30], KNN [27], [35], ANN [26], [27], [31], [35], NB [6]–[8], [24]–[26], BN [24], [32], RNN [28], and CNN [33]. Some of the predictor-focused studies suggested that the gender-specific models (across boys and girls) would improve prediction performance because the manifestation of obesity is different across the gender [5], [37]. The study in [29] is the only one that developed gender-specific models separately for boys and girls. Most of the earlier prediction-based studies used datasets that were obtained from cohorts but the later models have predominantly made use of Electronic Health Records.

The models in this category have also used medical images for modeling BMI prediction models. Models on adult obesity have used facial analysis for obesity classification and BMI regression [53]–[56] but there are no such models for childhood obesity. This might be due to several reasons, the topmost being privacy. Another reason for this could be that for children, the growth rate is quite rapid i.e., there might be several changes within a short period and children also often fall sick which causes vast fluctuations in weight. Due to

TABLE 7. List of studies on machine learning models that focus on prediction of obesity.

Article	Dataset (Country, Type)	Original dataset purpose	Dataset Size, Method, Predictors	Outcome Studied (Standard used)	Important predictors	Results (if available)
Adnan et al. [8] (2012)	Malaysian dataset (Malaysia, Survey/ Questionnaire)	Authors collected data.	140, NB, 19	Overweight/ Obesity	All manually selected 1. Children - catch-up growth, adiposity rebound, premature birth, gender. 2. Lifestyle- Duration of breastfeeding, duration of sleep, eating junk food, eating fried food, eating fruit, eating snacks in front of TV, duration of watching TV, eating warm meals for supper, physical activity, eating soup and sandwich bought outside home, eating snacks and chocolate bought outside home. 3. Family/environment factors - Mother BMI, father BMI, parents' weight, number of children	Accuracy Obesity - 71%, Overweight - 50%
Adnan et al. [6] (2012)	Malaysian dataset (Malaysia, not reported)	Authors collected data.	180, NB, 19	Overweight/ Obesity	Same as above.	Overall accuracy not reported.
Adnan et al. [7] (2012)	Malaysian dataset (Malaysia, not reported)	Authors collected data.	320, NB, 8	Overweight/ Obesity	Watching TV, Father's BMI, fried-food, Mother's BMI, sibling, fruit, average sleep, physical-activity.	Accuracy Obesity - 70.8-83.3%, Overweight - 58.3-95.83%
Dugan et al. [24] (2015)	CHICA (US, EHR)	Data from Pediatric clinical decision support system. Contains data from 4 different community health centers.	7519, DT, RF, NB, BN, 167	Obesity at 2 years, (CDC)	Overweight before 24 months, very tall before 6 months (protective), race (non-White), depression in parent (Black, Asian pacific/islander, other/unknown), using a walker (White children), dental advice provided (protective in Hispanic).	Accuracy - 85% for DT-ID3
Lingren et al. [25] (2016)	BCH, CCHMC (US, EHR)	BCH is a comprehensive EHR (Cerner Corporation, Kansas City, MO) and CCHMC (Epic Systems, Verona, WI)	BCH - 3675, CCHMC - 2200, SVM, NB	Severe obesity 1- 5.99 years (CDC, WHO)	-	Precision (SVM) - 0.773 - 0.813
Abdullah et al. [26] (2017)	SEGAK + questionnaires (Malaysia, Assessment + Questionnaire	Physical Fitness Standard for Malaysian School children and questionnaires.	4245, BN, DT, NB, ANN, SVM	Obesity at 12 years	District, SEGAK grade, Fibre2, SFA, Sugar, blood group, edu level mother, age mother, marital status mother, GDM, health prob mother, family income cat, birthweight father, health prob father, age father, marital status father, edu level father, epilepsy, blood disease, eyesight prob, learning prob, obes history, heart history, high blood pressure history, high blood pressure history, diabetes history, diabetes history, cancer history	Accuracy - 82.72% for DT-J48

TABLE 7. *(Continued.)* **List of studies on machine learning models that focus on prediction of obesity.**

Article	Dataset (Country, Type)	Original dataset purpose	Dataset Size, Method, Predictors	Outcome Studied (Standard used)	Important predictors	Results (if available)
Zheng & Ruggiero [27] (2017)	YRBBS (US, Survey)	Monitors six categories of health-related behaviors- 1. Unintentional injuries and violence. 2. Sexual behaviors 3. Alcohol and other drug use 4. Tobacco use 5. Unhealthy dietary behaviors 6. Inadequate physical activity Also measures the prevalence of obesity.	5127, LR, DT, KNN, ANN, 9	Obesity at 14–18 years (CDC)	Frequent physical activity and having breakfast everyday (Protective), Frequent consumption of sugar-sweeten beverages, and excessive computer use (Detrimental)	Accuracy - 88.82% for KNN
Gupta et al. [28] (2019)	Nemours Children Health System (US, EHR)	Network of pediatric health in US (Delaware, Florida, Pennsylvania, NewJersey)	40817, LSTM, 1737	BMI and Obesity 3-20 years (CDC)	BMI [Percentile] Per age and gender, Obese/Non-Obese Label, Allergic urticaria, Childhood obesity, Morbid obesity, Suspected clinical finding, Achondroplasia, MCH (Entitic mass) by Automated count, Cholesterol in LDL/Cholesterol in HDL, Hearing loss, Abnormal weight gain, Anomaly of chromosome pair 21, Erythrocytes (#/volume) in body fluid, Obesity, Hyperactive behavior, Tachycardia, Requires respiratory syncytial virus, Vaccination, CO2l1712, pH of Blood	Accuracy 1 year window - 83%-93%, 2 year window - 78%-89%, 3 year window - 76%-88%
Hammond et al. [29] (2019)	NYU Langone Health (US, EHR)	EHR data from patients in a safety net health system that serves a racially and ethnically diverse urban community in New York City.	3449, LR, RF, GBM, LASSO, 19290	Obesity at 5 years (CDC)	Weight for length z-score, BMI between 19 and 24 months, last BMI measure recorded before age 2	AUC Girls - 81.7% Boys - 76.1%
Park et.al [30] (2019)	Enhanced Nathan Kline Institute Rockland Sample (US, Images)	Is an ongoing, institutionally centered endeavor aimed at creating a large-scale (N >1000) community sample of participants across the lifespan. Measures include a wide array of physiological and psychological assessments, genetic information, and advanced neuroimaging.	76, LASSO, 379	BMI	Prefrontal, Posterior cingulate, sensorimotor, and inferior parietal-cortices.	AUC - 0.82

TABLE 7. *(Continued.)* **List of studies on machine learning models that focus on prediction of obesity.**

Article	Dataset (Country, Type)	Original dataset purpose	Dataset Size, Method, Predictors	Outcome Studied (Standard used)	Important predictors	Results (if available)
Singh & Tawfik [31] (2019)	Millennium Cohort Study (UK, Cohort)	Follows the lives of around 19000 young people born across England, Scotland, Wales and Northern Ireland in 2000-02. Information on physical, socio-emotional, cognitive and behavioral development over time but also about their daily life, behavior, and experiences, economic circumstances, parenting, relationships and family life.	NA, Linear Regression, ANN, NA	BMI at 14 years	NA	MAE - 1.42 Accuracy - 93.4% (ANN)
Kim et al. [32] (2019)	Korea Centers for Disease Control & Prevention (South Korea, Survey)	2017 Korean Youth Health Behavior Survey in 15 areas.	11206, GBN, 19	BMI categories	Classification- Pocket money, sleep quality, sitting study time, academic performance, wealth. Increases obesity risk (What-if analysis) 1. With low-income background, pocket money between $60-$80. 2. Study time increase with mediocre performance, increased pressure and thereby obesity. 3. Father's education level increased, mother's decreased.	Accuracy - 53.7%
Pang et al. [41] (2019)	Pediatric Big Data Repository(US, EHR)	Normalized and anonymized clinical database derived from the EHR system at the Children's Hospital of Philadelphia.	27203, GBM, 102	Obesity at 2-7 years (CDC)	Weight, height, race, ethnicity, body temperature, respiratory rate.	AUC - 0.81
Guan et al. [33] (2020)	Adolescent Brain Cognitive Development (US, Images)	The ABCD Study is a landmark, longitudinal study of brain development and child health.	3771, DMRN, NA	BMI	Heat maps visualization was done to understand the features used.	MAE - 2.87
Rossman et al. [34] (2020)	Clalit Health Services (Israel, EHR)	Clalit serves as a non-profit integrated care organization comprising over four million patients in Israel.	136196, GBM, NA	Obesity at 5-6 years (CDC)	Anthropometric measurements of the child and family, ancestry, and pregnancy glucose.	AUC - 0.80
Chatterjee et al. [35] (2020)	CHICA system (US, EHR)	Data from Pediatric clinical decision support system. Contains data from 4 different community health centers.	200, SVM, KNN, ANN, 11	Obesity at 3-5 years	NA	Accuracy - 96%

NA = Not available

the COVID-19 pandemic, there were fewer problem-focused visits, with notably fewer infection-related visits for pediatric healthcare [57] which shows that during normal school sessions, children often fall sick. As can be seen in the table, some of the studies for childhood and adolescent obesity prediction covered in this survey [28], [33], [34], [40], [41] also tried to include concepts of interpretability and explainability. Decision Tree, Linear Regression, and Logistic Regression models are proven to be more interpretable than other models such as Random Forest and Convolutional Neural Networks. For deep learning models, the concept of saliency in images which refers to unique features, such as pixels or resolution of the image in the context of visual processing, is used for explaining predictions of these models. Guan *et al.* [33] used this technique to visualise the regions in the brain image linked to high and low BMI. In addition to these, traditional methods of sensitivity analysis are also used to test the impact of each variable on the models output by varying the inputs.

The most common performance metric used by the existing studies was accuracy with a range of [53.7%, 96%] [6]–[8], [24], [26]–[28], [31], [32], [35]. AUC was used by four studies [29], [30], [34], [41] and ranged from 0.76 to 0.82. Two studies [31], [33] used MAE $[1.42, 2.87]$ and one [25] used precision $[0.773, 0.813]$ for measuring the performance of the models. We could not do comparative analysis of the existing studies because of two reasons: a) dataset diversity due to different countries, age, predicted outcome, etc. and, b) inconsistency in selecting performance metrics across studies.

IV. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

In this article, we reviewed machine learning and deep learning based obesity prediction models for children and adolescents.

Analyzing the existing literature, predictor-focused models used a broader range of variables from different domains. Besides demographic, and anthropometric variables which are included in almost every study, *psychological, behavioral, and lifestyle factors* were found to be important predictors. These factors are modifiable and there is a room for more work using these factors especially concerning the influence of social media platforms. Prediction-focused models *do not use variables* from as many domains as predictor-focused models do. The reason for this could be the nature of the study. Most of the studies that rank predictors use *cross-sectional data whereas prediction models use longitudinal data* which was not meant for studying obesity and hence often suffers from missing data. Only a limited number of studies use longitudinal obesity-specific large datasets from multiple domains including images (medical images, satellite images of neighborhood). Only a *handful of studies propose gender-specific obesity prediction models*. Although few predictor-focused models suggested that the gender-specific models for boys and girls would greatly benefit because the manifestation of obesity across gender might be different. In terms of ethnicity, predictor-focused models have established that Hispanic and Black children and adolescents have a higher obesity prevalence rate. Race/ethnicity is also ranked as a top predictor in almost every prediction-focused study. Further, during the COVID-19 pandemic, disparity increased even further i.e., the prevalence of obesity in these two communities further increased. More work needs to be done to investigate protective and detrimental factors concerning race and ethnicity.

The majority of these existing studies have utilized datasets from the USA but to curb the obesity level at a global scale, more studies incorporating datasets from different countries should be conducted. Countries that have low levels of childhood and adolescent obesity should also be investigated to study the reasons attributing to low prevalence which can further help in controlling obesity rates elsewhere.

GIS (Geographic Information System) tools could be further utilized in studying obesogenic and obesoprotective neighborhoods. A study done by Maharana *et al.* [45] used Convolutional Neural for extraction of features of the built environment (both natural and modified elements of the physical environment) from satellite images. This information was then used to assess associations between the built environment and obesity prevalence. This assessment showed that physical characteristics of a neighborhood such as the presence of parks, highways, green streets, crosswalks, diverse housing types, can be associated with variations in obesity prevalence across different neighborhoods. This kind of approach can help in *designing neighborhood-level intervention programs such as access to healthy food stores, playgrounds, etc.* to curb obesity.

Most of the *existing studies calculate BMI from self-reported* height and weight. An integrative review done by Engstrom *et al.* [58] amongst 26 studies examined for accuracy of height and weight measurements in 39,244 women, 21 found that women overestimate height. Thirty-four studies reviewed the accuracy of self-reported weight in 57,172 women, and all 34 studies reported that women underestimated weight. Reported variables such as food and dietary ones might also suffer from recall bias. In public health research, bias is an important issue. For questionnaires and surveys for measuring lifestyle, psychological, and behavioral variables social desirability/conformity bias may come into play. This type of bias exists when people respond in a way that they think will make them look good. *Prospective cohorts in which measurements are taken by trained people and habits are meticulously monitored can help in mitigating certain biases to some extent*.

Reviews on interventional studies could help in better understanding the factors, such as between modifiable and non-modifiable factors. Prediction models offer a good range [53.7%, 96%] of accuracy but these should be interpreted carefully keeping in mind the population and the type of datasets used. *The pros and cons of releasing prediction tools* to the general public need to be analyzed for better monitoring of obesity. Most of these models have not been validated in clinical settings and those that were validated have not been

tested on a different population (age, race/ethnicity, country) than the one used for model development. For this reason, these tools are not being used in actual practice and hence there is no data on how effective these tools would be when put to use in childhood and adolescent obesity prevention programs. Studies addressing this issue would help in further analyzing Machine Learning models for childhood and adolescent obesity as they might need to be *tailored across demographic variations* such as age group, race/ethnicity, and gender of the child.

Future Research Directions: The future research directions emanating from the aforementioned critical analysis of the existing literature are as follows:

1) **Psychological variables and social media impact:** Including psychological variables in studying obesity, due to children's access to various smart devices and social media platforms, which contributes to assessing one's body image [59] would be important. Body image and dissatisfaction can be seen in children as young as 5-7 years of age [60] and their impact on children's obesity levels for all age groups needs to be evaluated as a part of future work.
2) **Role of childcare centers:** More studies are required in evaluating the impact of childcare centers on obesity, as they contribute to the development of dietary habits of children. As this factor is modifiable, investigating it can provide a good target area for intervention programs.
3) **GIS tools and Machine Learning:** More studies on using GIS tools combined with ML to capture features of the built environment are required. These kinds of studies would be useful in studying associations and curbing obesity in areas where people of a single ethnicity/race reside.
4) **Underexplored environment variables:** Studying the impact of underexplored environment variables, such as home, school, and neighborhood environment, on obesity development is required as a part of future work.
5) **Gender-and race-specific models:** More studies are required on developing and evaluating the impact of gender-specific models for boys and girls on obesity prediction. Shah *et al.* [61] deduced from the first Atlas of Childhood Obesity in 2019 released by WHO that for children aged between 5-9 years, 123 of 188 (65%) countries had a greater prevalence of obesity for boys than girls. This difference in the prevalence of obesity across gender needs to be investigated. Apart from gender, race-specific models should be developed to obtain equitable obesity prediction accuracy across race/ethnicity. Further, fairness of the facial images based obesity prediction models should be evaluated across race, gender and age. Bias mitigation strategies should be designed to reduce unequal accuracy rates as a part of future work.
6) **Newer cohorts:** There is a need for assembling and using new large-scale datasets/cohorts using multi-modalities at the international level. As most of the cohorts/datasets used in the existing studies are quite outdated. The evolution of obesity with respect to changing times needs to be looked into as with the rise of social media, apart from behavioral and psychological variables, dietary and lifestyle changes are witnessed. The impact of social media on obesity is an under-explored area and needs thorough investigation.
7) **Advanced computer vision techniques and Multi-modal analysis:** Handful of studies have used computer vision techniques for predicting obesity from MRIs at children and adolescent level [30], [33]. There is a need for advanced computer vision techniques for predicting obesity using image data such as facial images and MRI. Further, fusion models are required that can combine different sources and formats of data (such as text, image, and sensor) for enhanced prediction accuracy and analysis. Further, compressed deep learning models should be investigated to facilitate on-device deployment and inference on resource constraint smartphones for obesity prediction and monitoring.

As can be seen from the analysis, studies on childhood and adolescent obesity are very diverse in terms of datasets, methods, factor domains, etc., used for model development. Previous surveys on childhood and adolescent obesity offer insightful conclusions but they do not provide a thorough breakdown of studies into the type of datasets used, predicted outcome, different domains of predictor variables, methods used, and the purpose of developing the models.

Although extensive studies have been carried out to analyze as well as predict childhood and adolescent obesity but still, the rates of childhood and adolescent obesity are going up as is evident from the numbers in Section I. Categorization of existing literature is needed for proper understanding, critical analysis, and the future research directions.

Our survey provides a taxonomy of the existing literature for understanding the state-of-the-art. It also highlights the gap (Prediction-focused studies do not utilize the findings of predictor-focused models to their full potential) that needs to be filled in for developing better obesity prediction models. In addition, future research directions such as the impact of social media, using GIS tools for studying obesogenic and obesoprotective environments, developing gender and race-specific models, and using advanced computer vision techniques to study childhood and adolescent obesity are provided for further advancement of the state-of-the-art in this field.

In summary, our survey provides a comprehensive view of the existing literature, identify the existing gaps in the literature, and propose future research directions which would help in further advancing the state-of-the-art in child and adolescent healthcare. The findings and analysis from this review could be used by researchers from different fields such

as public health, social sciences, etc. to study and develop better prediction and obesity intervention programs.

ACKNOWLEDGMENT

All the authors are part of the Institute for Health Equity Advances at Wichita State University.

REFERENCES

[1] M. Guillaume, "Defining obesity in childhood: Current practice," *Amer. J. Clin. nutrition*, vol. 70, no. 1, pp. 126S–130S, 1999.
[2] T. J. Cole, M. C. Bellizzi, K. M. Flegal, and W. H. Dietz, "Establishing a standard definition for child overweight and obesity worldwide: International survey," *Bmj*, vol. 320, no. 7244, p. 1240, 2000.
[3] M. de Onis, "Development of a WHO growth reference for school-aged children and adolescents," *Bull. World Health Org.*, vol. 85, no. 9, pp. 660–667, Sep. 2007.
[4] R. J. Kuczmarski, *2000 CDC Growth Charts for United States: Methods and Development*, no. 246. Atlanta, GA, USA: Centers for Disease Control and Prevention, Department of Health and Human Services, 2002.
[5] J. C. Gray, N. A. Schvey, and M. Tanofsky-Kraff, "Demographic, psychological, behavioral, and cognitive correlates of BMI in youth: Findings from the adolescent brain cognitive development (ABCD) study," *Psychol. Med.*, vol. 50, no. 9, pp. 1539–1547, Jul. 2020.
[6] M. H. B. M. Adnan and W. Husain, "A hybrid approach using naïve Bayes and genetic algorithm for childhood obesity prediction," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, vol. 1, 2012, pp. 281–285.
[7] M. H. M. Adnan and W. Husain, "Hybrid approaches using decision tree, naive bayes, means and Euclidean distances for childhood obesity prediction," *Int. J. Softw. Eng. Appl.*, vol. 6, no. 3, pp. 99–106, 2012.
[8] M. Adnan, W. Husain, and N. Rashid, "Parameter identification and selection for childhood obesity prediction using data mining," in *Proc. 2nd Int. Conf. Manage. Artif. Intell.*, vol. 35, 2012, pp. 75–80.
[9] D. H. Rehkopf, B. A. Laraia, M. Segal, D. Braithwaite, and E. Epel, "The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls," *Int. J. Pediatric Obesity*, vol. 6, nos. 2–2, pp. e233–e242, Jun. 2011.
[10] A. Pochini, Y. Wu, and G. Hu, "Data mining for lifestyle risk factors associated with overweight and obesity among adolescents," in *Proc. 3rd Int. Conf. Adv. Appl. Informat. (IIAI)*, Aug. 2014, pp. 883–888.
[11] C. Nau, H. Ellis, H. Huang, B. S. Schwartz, A. Hirsch, L. Bailey-Davis, A. M. Kress, J. Pollak, and T. A. Glass, "Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments," *Health Place*, vol. 35, pp. 136–146, Sep. 2015.
[12] A. M. O. Hinojosa, K. E. MacLeod, J. Balmes, and M. Jerrett, "Influence of school environments on childhood obesity in California," *Environ. Res.*, vol. 166, pp. 100–107, Oct. 2018.
[13] I. Lee, K.-S. Bang, H. Moon, and J. Kim, "Risk factors for obesity among children aged 24 to 80 months in Korea: A decision tree analysis," *J. Pediatric Nursing*, vol. 46, pp. e15–e23, May 2019.
[14] M. H. B. M. Adnan, W. Husain, and F. Damanhoori, "A survey on utilization of data mining for childhood obesity prediction," in *Proc. 8th Asia–Pacific Symp. Inf. Telecommun. Technol.*, 2010, pp. 1–6.
[15] B. Novak and M. Bigec, "Application of artificial neural networks for childhood obesity prediction," in *Proc. 2nd New Zealand Int. Two-Stream Conf. Artif. Neural Netw. Expert Syst.*, 1995, pp. 377–380.
[16] B. Novak and M. Bigec, "Childhood obesity prediction with artificial neural networks," in *Proc. 9th IEEE Symp. Comput. Med. Syst.*, Jun. 1996, pp. 77–82.
[17] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane, "Comparing data mining methods with logistic regression in childhood obesity prediction," *Inf. Syst. Frontiers*, vol. 11, no. 4, pp. 449–460, Sep. 2009.
[18] A. Triantafyllidis, E. Polychronidou, A. Alexiadis, C. L. Rocha, D. N. Oliveira, A. S. da Silva, A. L. Freire, C. Macedo, I. F. Sousa, E. Werbet, E. A. Lillo, H. G. Luengo, M. T. Ellacuría, K. Votis, and D. Tzovaras, "Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101844.
[19] G. Colmenarejo, "Machine learning models to predict childhood and adolescent obesity: A review," *Nutrients*, vol. 12, no. 8, p. 2466, Aug. 2020.
[20] M. N. LeCroy, R. S. Kim, J. Stevens, D. B. Hanna, and C. R. Isasi, "Identifying key determinants of childhood obesity: A narrative review of machine learning studies," *Childhood Obesity*, vol. 17, no. 3, pp. 153–159, Apr. 2021.
[21] N. Poh, A. Rattani, and F. Roli, "Critical analysis of adaptive biometric systems," *IET Biometrics*, vol. 1, no. 4, pp. 179–187, Dec. 2012.
[22] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf. Fusion*, vol. 52, pp. 187–205, Dec. 2019.
[23] S. F. Ardabili, B. Najafi, S. Shamshirband, B. M. Bidgoli, R. C. Deo, and K. Chau, "Computational intelligence approach for modeling hydrogen production: A review," *Eng. Appl. Comput. Fluid Mech.*, vol. 12, no. 1, pp. 438–458, 2018.
[24] T. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs, "Machine learning techniques for prediction of early childhood obesity," *Appl. Clin. Informat.*, vol. 6, no. 3, p. 506, 2015.
[25] C. Brady, V. Thaker, C. Brady, B. Namjou, S. Kennebeck, J. Bickel, N. Patibandla, Y. Ni, S. L. Van Driest, L. Chen, and A. Roach, "Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers," *Appl. Clin. Informat.*, vol. 7, no. 3, pp. 693–706, 2016.
[26] F. S. Abdullah, N. S. A. Manan, A. Ahmad, S. W. Wafa, M. R. Shahril, N. Zulaily, R. M. Amin, and A. Ahmed, "Data mining techniques for classification of childhood obesity among year 6 school children," in *Proc. Int. Conf. Soft Comput. Data Mining*. Cham, Switzerland: Springer, 2016, pp. 465–474.
[27] Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 2132–2138.
[28] M. Gupta, T.-L. T. Phan, T. Bunnell, and R. Beheshti, "Obesity prediction with EHR data: A deep learning approach with interpretable elements," 2019, *arXiv:1912.02655*.
[29] R. Hammond, R. Athanasiadou, S. Curado, Y. Aphinyanaphongs, C. Abrams, M. J. Messito, R. Gross, M. Katzow, M. Jay, N. Razavian, and B. Elbel, "Predicting childhood obesity using electronic health records and publicly available data," *PLoS ONE*, vol. 14, no. 4, Apr. 2019, Art. no. e0215571.
[30] B.-Y. Park, C.-S. Chung, M. J. Lee, and H. Park, "Accurate neuroimaging biomarkers to predict body mass index in adolescents: A longitudinal study," *Brain Imag. Behav.*, vol. 14, no. 5, pp. 1682–1695, Oct. 2020.
[31] B. Singh and H. Tawfik, "A machine learning approach for predicting weight gain risks in young adults," in *Proc. 10th Int. Conf. Dependable Syst., Services Technol. (DESSERT)*, Jun. 2019, pp. 231–234.
[32] C. Kim, F. J. Costello, K. C. Lee, Y. Li, and C. Li, "Predicting factors affecting adolescent obesity using general Bayesian network and what-if analysis," *Int. J. Environ. Res. Public Health*, vol. 16, no. 23, p. 4684, Nov. 2019.
[33] H. Guan, E. Yang, L. Wang, P.-T. Yap, M. Liu, and D. Shen, "Linking adolescent brain MRI to obesity via deep multi-cue regression network," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2020, pp. 111–119.
[34] H. Rossman, S. Shilo, S. Barbash-Hazan, N. S. Artzi, E. Hadar, R. D. Balicer, B. Feldman, A. Wiznitzer, and E. Segal, "Prediction of childhood obesity from nationwide health records," *J. Pediatrics*, vol. 233, pp. 132–140, 2021.
[35] K. Chatterjee, U. Jha, P. Kumari, and D. Chatterjee, "Early prediction of childhood obesity using machine learning techniques," in *Proc. Adv. Commun. Comput. Technol.* Singapore: Springer, 2021, pp. 1431–1440.
[36] P. Kitsantas and K. F. Gaffney, "Risk profiles for overweight/obesity among preschoolers," *Early Hum. Develop.*, vol. 86, no. 9, pp. 563–568, Sep. 2010.
[37] C. Lazarou, M. Karaolis, A.-L. Matalas, and D. B. Panagiotakos, "Dietary patterns analysis using data mining method. An application to data from the CYKIDS study," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 706–714, Nov. 2012.
[38] A. Van Hulst, M.-H. Roy-Gagnon, L. Gauvin, Y. Kestens, M. Henderson, and T. A. Barnett, "Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors," *Int. J. Behav. Nutrition Phys. Activity*, vol. 12, no. 1, pp. 1–9, Dec. 2015.
[39] P. Wiechmann, K. Lora, P. Branscum, and J. Fu, "Identifying discriminative attributes to gain insights regarding child obesity in hispanic preschoolers using machine learning techniques," in *Proc. IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2017, pp. 11–15.
[40] H. Marcos-Pasero, G. Colmenarejo, E. Aguilar-Aguilar, A. R. de Molina, G. Reglero, and V. Loria-Kohen, "Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, Dec. 2021.

[41] X. Pang, C. B. Forrest, F. Le-Scherban, and A. J. Masino, "Understanding early childhood obesity via interpretation of machine learning model predictions," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1438–1443.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[43] E. A. Spencer and K. Mahtani, "Hawthorne effect," *Catalogue Bias*, 2017. [Online]. Available: https://catalogofbias.org/biases/hawthorne-effect/

[44] N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway, and B. G. Deeds, "The conception of the ABCD study: From substance use to a broad NIH collaboration," *Develop. Cognit. Neurosci.*, vol. 32, pp. 4–7, Aug. 2018.

[45] A. Maharana and E. O. Nsoesie, "Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity," *JAMA Netw. Open*, vol. 1, no. 4, 2018, Art. no. e181535.

[46] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber, "Face-to-BMI: Using computer vision to infer body mass index on social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2017, vol. 11, no. 1, pp. 1–4.

[47] M. Shields and M. S. Tremblay, "Canadian childhood obesity estimates based on WHO, IOTF and CDC cut-points," *Int. J. Pediatric Obesity*, vol. 5, no. 3, pp. 265–273, May 2010.

[48] C. Pedrosa, F. Correia, D. Seabra, B. M. Oliveira, C. Simoes-Pereira, and M. D. Vaz-de-Almeida, "Prevalence of overweight and obesity among 7–9-year-old children in aveiro, portugal: Comparison between IOTF and CDC references," *Public Health Nutrition*, vol. 14, no. 1, pp. 14–19, Jan. 2011.

[49] I. Gonzalez-Casanova, O. L. Sarmiento, J. A. Gazmararian, S. A. Cunningham, R. Martorell, M. Pratt, and A. D. Stein, "Comparing three body mass index classification systems to assess overweight and obesity in children and adolescents," *Revista Panamericana de Salud Pública*, vol. 33, no. 5, pp. 349–355, May 2013.

[50] C. Bouchard, "Childhood obesity: Are genetic differences involved?" *Amer. J. Clin. Nutrition*, vol. 89, no. 5, pp. 1494S–1501S, 2009.

[51] A. Herbert, N. P. Gerry, M. B. McQueen, I. M. Heid, A. Pfeufer, T. Illig, H.-E. Wichmann, T. Meitinger, D. Hunter, F. B. Hu, and G. Colditz, "A common genetic variant is associated with adult and childhood obesity," *Science*, vol. 312, no. 5771, pp. 279–283, Apr. 2006.

[52] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[53] H. Siddiqui, A. Rattani, D. R. Kisku, and T. Dean, "AI-based BMI inference from facial images: An application to weight monitoring," 2020, *arXiv:2010.07442*.

[54] M. Jiang, Y. Shang, and G. Guo, "On visual BMI analysis from facial images," *Image Vis. Comput.*, vol. 89, pp. 183–196, Sep. 2019.

[55] M. Jiang, G. Guo, and G. Mu, "Visual BMI estimation from face images using a label distribution based method," *Comput. Vis. Image Understand.*, vols. 197–198, Aug. 2020, Art. no. 102985.

[56] N. Yousaf, S. Hussein, and W. Sultani, "Estimation of BMI from facial images using semantic segmentation based region-aware pooling," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104392.

[57] K. Schweiberger, S. Y. Patel, A. Mehrotra, and K. N. Ray, "Trends in pediatric primary care visits during the coronavirus disease of 2019 pandemic," *Acad. Pediatrics*, vol. 21, no. 8, pp. 1426–1433, Nov./Dec. 2021.

[58] J. L. Engstrom, S. A. Paterson, A. Doherty, M. Trabulsi, and K. L. Speer, "Accuracy of self-reported height and weight in women: An integrative review of the literature," *J. Midwifery Women's Health*, vol. 48, no. 5, pp. 338–345, Sep. 2003.

[59] J. Fardouly and L. R. Vartanian, "Social media and body image concerns: Current research and future directions," *Current Opinion Psychol.*, vol. 9, pp. 1–5, Jun. 2016.

[60] M. Perez, A. M. K. Van Diest, H. Smith, and M. R. Sladek, "Body dissatisfaction and its correlates in 5-to 7-year-old girls: A social learning experiment," *J. Clin. Child Adolescent Psychol.*, vol. 47, no. 5, pp. 757–769, Sep. 2018.

[61] B. Shah, K. T. Cost, A. Fuller, C. S. Birken, and L. N. Anderson, "Sex and gender differences in childhood obesity: Contributing to the research agenda," *BMJ Nutrition, Prevention Health*, vol. 3, no. 2, p. 387, 2020.

HERA SIDDIQUI received the M.S. degree in computer science from San Diego State University, San Diego, CA, USA, in 2018. She is currently pursuing the Ph.D. degree in computer science with Wichita State University, Wichita, KS, USA.

She is also a Graduate Research Assistant at Wichita State University. Her research interest includes AI for healthcare with a focus on obesity prediction models.

AJITA RATTANI received the Ph.D. degree from the Department of Computer Science and Engineering, Michigan State University, USA, and the Ph.D. degree in computer science engineering from the University of Cagliari, Italy.

She is currently an Assistant Professor with the School of Computing, Wichita State University. She is also the Principal Investigator to three research projects funded by NSF, DOD, and WSU. Her research interests include computer vision, image analysis, deep learning, machine learning, and biometrics.

Dr. Rattani was a recipient of the Best Paper and Poster awards at IEEE IJCB 2014, IEEE HST 2017, 2019, and IAPR Biometric Summer School 2008. She is the Lead Editor of the Springer books titled *Adaptive Biometric Systems: Recent Advances and Challenges* and *Selfie Biometrics: Advances and Challenges*.

NIKKI K. WOODS received the bachelor's degree in health services management and community development from Wichita State University, in 2007, and the master's degree in behavioral science and public health and the Ph.D. degree in behavioral psychology from the University of Kansas.

She is currently an Associate Professor with the Department of Public Health Sciences, Wichita State University. She is a Maternal, Infant, and Child Health Researcher and an Educator with an emphasis on addressing health disparities. She established the Women's Health Network and the WSU Center for Health Disparities through securing external funding.

Dr. Woods was a recipient of the College of Health Professions Excellence in Research Award in 2017 (Wichita State University) and the National Public Health Thank You Day-Public Health Hero 2018 (Sedgwick County Division of Health).

LAILA CURE received the Ph.D. degree in industrial engineering from the University of South Florida, in 2011.

She is currently an Assistant Professor with the Department of Industrial and Manufacturing Engineering, Wichita State University. Her research focuses on the use of analytics and mathematical modeling techniques to support complex operations involving human behavior. Her research interests include analysis and design of complex work systems, data analytics, model-based decision support for the deployment of patient safety, quality, and population health interventions.

RHONDA K. LEWIS received the degree in psychology from Wichita State University, in 1991, the Master of Public Health degree from the University of Kansas School of Medicine, in 1996, and the Ph.D. degree in developmental and child psychology from the University of Kansas.

She is currently a Professor and the Chair of the Psychology Department, Wichita State University. She is a First-Generation Student. She is a Service-Learning Faculty Fellow at WSU. She is part of the Leadership Team for the NSF ADVANCE grant to increase Women and Underrepresented Minorities in STEM. She is a Co-PI for the President's Convergence Science Initiative to reduce health disparities in vulnerable populations. She uses behavioral and community research methodologies to promote health among adolescents and reduce health disparities. She has over 25 years of experience in community organizing, program development, and evaluation. She has over 60 publications and 100 presentations at regional, national, and international conferences.

Dr. Lewis is a member of the Society for Community Research and Action and the Association of Black Psychology. In 2020, she received the Legacy Award from Sistahs Can We Talk. In 2019, she received the Mental Health Professional of the Year award from the NAMI (National Alliance on Mental Illness). In 2017, she received the Wichita Business Journal's Diversity Leader's Award and the President's Distinguished Service Award at Wichita State University.

JANET TWOMEY received the master's and Ph.D. degrees in industrial engineering from the University of Pittsburgh, in 1992 and 1995, respectively.

She is currently an Associate Dean of the College of Engineering and a Professor of industrial, systems, and manufacturing engineering at Wichita State University. She has organized faculty development seminars focusing on research, assigns new faculty mentors, and travels to NSF with new faculty to meet program offices. As a Program Officer at the National Science Foundation, she together with two other NSF Program Officers, developed and delivered NSF Proposal Writing and CAREER Proposal Writing two workshops. Since leaving NSF, she has participated in over ten NSF sponsored CAREER proposal writing workshops. Her research interests include technology for environmental sustainability, intelligent computational methods, and industrial sustainability.

Dr. Twomey received the National Science Foundation (NSF) CAREER Award in 1998. She was elected as the Academic Vice President for Board of Trustees, Institute of Industrial Engineering, in 2011, 2013, and 2014.

BETTY SMITH-CAMPBELL received the master's degree in community health nursing from the University of Kansas and the Ph.D. degree in nursing from the University of Colorado.

She is currently a Professor Emeritus with the School of Nursing, Wichita State University.

Dr. Smith-Campbell received the Excellence in Leadership Award, the Sigma Theta Tau International–Epsilon Gamma-at-Large Chapter in 2015, and the American Association Nurse Practitioner (AANP) State Advocate Award in 2014.

TWYLA J. HILL received the master's and Ph.D. degrees in social sciences from the University of California at Irvine, Irvine, CA, USA, in 1993 and 1998, respectively.

She is currently a Professor of sociology with the Department of Sociology, Wichita State University. She is the Principal Investigator of the President's Convergence Science Initiative, Wichita State University. Her research interests include sociology of aging, sociology of families, research methods, sociology of law, and public policy.

Dr. Hill received a grant from the Regional Institute on Aging for an Interdisciplinary Project with a Faculty Member from Dance in 2020.