International Journal of Advanced Trends in Computer Applications
www.ijatca.com

DATA MINING METHODS FOR OBESITY LEVEL RECOGNITION: A SYSTEMATIC REVIEW OF THE LITERATURE

[1]Rafael Luckert, [2]Kevin De Alba, [3]Jaime Sarmiento, [4]Karen Salas Viloria, [5]Alexis De la Hoz Manotas, [6]Fabio Mendoza Palechor
[1]Department of Computer Science and Electronics, Universidad de la Costa, Barranquilla, Colombia
[1]rluckert1@cuc.edu.co, [2]kdealba1@cuc.edu.co, [3]jsarmient37@cuc.edu.co, [4]ksalas@cuc.edu.co, [5]adelahoz6@cuc.edu.co, [6]fmendoza1@cuc.edu.co

Abstract: *Obesity in teenagers and adults has increased worldwide, with serious impact and consequences for health in the short and long term. Technology has allowed to discover new ways of treating diseases and problems with health issues, and data mining has become a relevant area of research and discovery, especially in recent years due to its precision and reliability analyzing datasets of patients to detect diseases and facilitate their prevention. The goal of this study was to identify the techniques and algorithms in data mining most commonly used, to detect several factors that favor the apparition of obesity issues and to determine the reliability of those methods, based on the results obtained from a data mining model. Data mining methods as simple regression and decision trees, are most commonly used to detect obesity levels, where the simple regression method was found in 19% of the articles reviewed and the decision trees method was used in 11% of them.*

Keywords: obesity, data mining, overweight.

I. INTRODUCTION

Overweight and obesity are defined as abnormal or excessive fat accumulation that may impair health. In 2016, WHO statistics accounted that 39% of adult people and over 18 years old were overweight, and 13% were obese, more than 340 million children and adolescents aged 5-19 were overweight or obese [1]. Overweight can be measured with the Mass Body Index (MBI), that ties weight with the height of a person, an increase in MBI can increase the risk to contract diseases such as diabetes and cardiac conditions.

Data mining has become widely used in the field of medicine, focusing its impact in the use of tools for analysis and classification of information. Many studies about obesity, use several data mining methods to identify possible causes and risks that are present in this condition, having variables such as: the presence of a high caloric intake, sedentarism, hormonal issues and others, the early detection of these factors, can improve methods for preventing obesity in population.

The analysis of obesity related problems is a topic of interest for researchers, with a significant number of authors producing solutions based on intelligent methods for early and efficient detection of this disease, especially contributions from studies as [2-5]. According to literature, a common trend for studies related with obesity are methods like logistic regression [9-10], support vector machines [24], Naïve Bayes [6] and decision trees [15][28][35].
Those data mining techniques and many others, are applied to datasets, collections of information from many and different patients, to determine patterns and models for detecting obesity problems, some of those datasets are PubMed-NCBI [12], ANZCTRN [13], STRIDE [25][34] and DEOL [43].

This systematic review focuses on answering questions such as: ¿which techniques or algorithms in data mining are commonly used to detect factors related to obesity?, ¿What is the reliability of the results shown by a data mining model to detect obesity issues?, to achieve that objective, a sample of articles was observed, all having proposals and implementations of analytics using machine learning and data mining for information validation and decision making.

Finally, the study is structured with the following sections: II. Previous Studies, contains the information of every contribution from authors, III. Methodology, covers all the stages that were made to accomplish the review, IV. Scientometric analysis resumes all the quantitative data in aspects such as number of articles by year, number of publications by journal and quartile of each publication, V. Technical Analysis, contains all the information associated to the number of studies and the pathologies related with cardiovascular diseases, methods of data mining and databases used by them. VI. Characterization and Results, covers the results obtained based on different validation metrics for the data mining methods implemented, VII: Discussion, presents a critical analysis of the results observed in the reviewed documents to close the research question, and conclusions can be found at VIII section.

II. PREVIOUS WORKS

Obesity is a condition that affects many people worldwide, regardless of gender and economic factor, for many reasons, has become a relevant field of scientific research, in this review you can find several proposals to solve the challenge of identifying the obesity level in people, with high precision and early detection, a brief review of each contribution follows.

In [6], the authors presented a study to create a prediction model for factors that could trigger obesity in young people, using Bayesian networks and what-if analysis. The study showed the remarkable improvement in the socioeconomic lifestyle worldwide has consolidated adolescent obesity, becoming a public health problem that cannot be ignored. Authors explored an approach using General Bayesian Networks (GBN) with What-If analysis looking to apply this solution to other areas of public health. The data used was provided by the 2017 Korean Youth Health Survey conducted by the centers for control and prevention of diseases in Korea, including 19 attributes and 11,206 individual data points.

In [7], the authors looked for solid evidence of causality between factors such as diabetes, cardiovascular risk and vitamin D concentrations. Their approach was based in an retrospective observational study using data mining analysis with Artificial Analysis of Neural Networks (ANN). They built an auto-contractive map (AutoCM) and semantic mapping followed by Activation and Competition system on data of workers referred by an ambulatory clinic of occupational health. The parameters analyzed included weight, height, waist circumference, mass body index (MBI), percentage of fat mass, glucose, insulin, glycated hemoglobin, creatinine, total cholesterol, low and high density lipoprotein cholesterol, triglycerides, uric acid, fibrinogen, homocysteine, c-reactive protein, diastolic and systolic blood pressure, and 25(OH)D.

In [8], the study had the objective of generating waist circumference to height ratio cut-off values for obesity categories in a model of the association between mass body index and waist circumference to height ratio. The values of the circumference of the waist and its proportion with height were compared with values prevalent in practice, derived from pragmatic criteria. The data collected included age, sex, height, weight, waist circumference, presence of diabetes, hypertension and cardiovascular disease, for 847 participants over 8 years old involved in a rural study conducted by the Australian Health Revision Clinic (DiabHealth). The authors concluded that the cut off values recommended for waist circumference to height ratio, provided a useful index to evaluate obesity stages and the risk of chronic diseases, improving the medical attention in the clinic practices.

In [9], the goal of the study was to research the FTO mRNA expression in human clear cell renal cell carcinoma and its clinical value. FTO mRNA expression and its prognostic value were investigated by bioinformatic analysis of the data from The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/). The Kaplan-Meier analysis showed that FTO mRNA expression in the lower quartile is significantly associated with poor survival in clear cell renal cell carcinoma patients ($P < 0.0001$). This study indicated that higher FTO mRNA expression may have a protective role and it may be a vital molecular marker in the prognosis of clear cell renal cell carcinoma patients.

In [10], the authors aimed to identify salivary biomarkers and other factors associated with obesity using an ensemble data mining approach. For a random cohort of over 700 subjects from 8137 Kuwait children (10.00 ± 0.67 years), four data mining methods were applied to identify important variables associated with obesity, including logistic regression by lasso regularization (Lasso), multivariate adaptive regression spline (MARS), random forests (RF), and boosting classification trees (BT). The study concluded that a data mining approach based on multiple algorithms is useful for identifying factors associated with phenotypes, especially in cases where relationships are not salient, and a consensus from multiple methods can help produce a more generalizable subset of features. The authors demonstrated that evaluation using the waist circumference includes association with high levels of salivary leptin, which is not seen with evaluation by BMI.

In [11], the study looked to determine the associations between miRNAs-target genes, miRNA-long ncRNAs (lncRNAs), and miRNAs-small molecules in human metabolic diseases, including obesity, type 2 diabetes and non-alcoholic fatty liver disease. The metabolic disease-related miRNAs were obtained from the Human MicroRNA Disease Database (HMDD) and miR2Disease database. A search on the databases Matrix Decomposition and Heterogeneous Graph Inference (MDHGI) and DisGeNET were also performed. The results showed a total of 20 miRNAs were revealed to be associated with metabolic disorders. Furthermore, the target genes of these miRNAs participate in several pathways previously associated with metabolic disease, and interactions between miRNA-lncRNA and miRNA-small molecules were also found, suggesting that some molecules can modulate gene expression via such an indirect way. Thus, the results of this data mining and integration analysis provided further information on the possible molecular basis of the metabolic disease pathogenesis as well as provide a path to search for potential biomarkers and therapeutic targets concerning metabolic diseases.

In [12], the study reported that microRNA-155 (miR-155) deficiency in ApoE-/- mice yields a novel metabolically healthy obese (MHO) model, which exhibited improved atherosclerosis but resulted in obesity, non-alcoholic fatty liver disease (NAFLD) without insulin resistance. Using experimental data mining approaches combined with experiments, the authors found, among 109 miRNAs, miR-155, and miR-221 are significantly modulated in all four hyperlipidemia-related diseases (HRDs), namely atherosclerosis, NAFLD, obesity and type II diabetes (T2DM). These findings led to a new classification of types I and II MHOs, which are regulated by miR-221 and miR-155, respectively. Western blots showed that the proinflammatory adipokine, resistin, is significantly increased in white adipose tissues (WAT) of the MHO mice, revealing on the study, miR-155-suppressed "secondary wave inflammatory state (SWIS)," characteristic of MHO transition to classical obesity (CO). Their conclusions showed that MHO may have heterogeneity in comorbidities, and is therefore classified into type I, and type II MHOs; and that increased expression of resistin in miR-155-/- white adipose tissues may be a driver for SWIS in MHO transition to CO. These findings provided novel insights into the pathogenesis of MHO, MHO transition to CO, hyperlipidemic pathways related to cancer, and new therapeutic targets.

In [13], the goal of the study was to identify the usual food choices for meals of overweight and obese volunteers for a weight-loss trial. A cross-sectional analysis was performed using screening diet history data from a 12-month weight-loss trial (the HealthTrack study). A descriptive data mining tool, the Apriori algorithm of association rules, was applied to identify food choices at meal occasions using a nested hierarchical food group classification system. Overall, 432 breakfasts, 428 lunches, 432 dinners and 433 others (meals) were identified from the intake data (n 433 participants). A total of 142 items of closely related food clusters were identified at three food group levels. Given the large number of foods available, having an understanding of eating patterns in which key foods drive overall meal content can help translate and develop novel dietary strategies for weight loss at the individual level.

In [14], authors developed a new comprehensive mobile architecture for tackling the challenging issues of obesity control, monitoring, and prevention. Their architecture system can also help individuals track food intake, lifestyle, calories intake, calories consumption, and exercise activities. The study analyzed the data collected from continuous monitoring using non-invasive sensors, in addition to the data collected from social communities created to propagate awareness and share appropriate information about the obesity problem and its solution. Authors also developed data mining algorithms and sentiment analysis algorithms and generate intelligent suggestions, warnings, and recommendations to control and mitigate the risk of obesity and its related diseases. Finally they evaluated the efficacy and scalability of the implemented system using a comprehensive cloud database including entered data, calculated data, sensory data, and social data of 50 underweight, overweight, normal, and obese volunteer subjects.

In [15], the study was looking to evaluate and identify the risk factors associated with metabolic syndrome by using a decision tree algorithm as a data mining tool. A total of 6578 individuals were included in the analysis using a body mass index (BMI) cutoff >= 25 kg/m2 for the definition of overweight, in accordance with International Diabetic Federation (IDF) criteria. Two models were validated. The results showed that in model I, serum fasted TG was the most important associated risk factor for metabolic syndrome. In model II, serum Hs-CRP was identified as a risk factor of metabolic syndrome. The findings in model I suggest that the IDF criteria are suitable for identifying individuals within the Iranian population into those with, or without MetS. Furthermore, model II showed that serum Hs-CRP concentrations were identified as a risk factor for metabolic syndrome within the Iranian population.

In [16], authors presented a study to examine the risk factors of developing functional decline and make probabilistic predictions by using a tree-based method that allows higher order polynomials and interactions of the risk factors. The conditional inference tree analysis, a data mining approach, was used to construct a risk stratification algorithm for developing functional limitation based on BMI and other potential risk factors for disability in 1,951 older adults without functional limitations at baseline (baseline age 73.1 ± 4.2 y). The study also analyzed the data with multivariate stepwise logistic regression and compared the two approaches (e.g., cross-validation). Over a mean of 9.2 ± 1.7 years of follow-up, 221 individuals developed functional limitation. Higher BMI, age, and comorbid disease were consistently identified as significant risk factors for functional decline among older individuals across all approaches and analyses.

In [17], the goal of the authors was to (1) build, from geotagged Twitter and Yelp data, a national food environment database and (2) to test associations between state food environment indicators and health outcomes. This was a cross-sectional study based upon secondary analyses of publicly available data. Using Twitter's Streaming Application Programming Interface (API), they collected and processed 4,041,521 food-related, geotagged tweets between April 2015 and March 2016. Using Yelp's Search API, the study collected data on 505,554 unique food-related businesses. In linear regression models, the authors examined associations between food environment characteristics and state-level health outcomes, controlling for state-level differences in age, percent non-Hispanic white, and median household income. A one standard deviation increase in caloric density of food tweets was related to higher all-cause mortality (+46.50 per 100,000), diabetes (+0.75%), obesity (+1.78%), high cholesterol (+1.40%), and fair/poor self-rated health (2.01%). More burger Yelp listings were related to higher prevalence of diabetes (+0.55%), obesity (1.35%), and fair/poor self-rated health (1.12%).

In [18], the goal of the study was to assess the effects of diet, physical activity and behavioral interventions for the treatment of overweight or obese adolescents aged 12 to 17 years. The authors performed a systematic literature search in: CENTRAL, MEDLINE, Embase, PsycINFO, CINAHL, LILACS, and the trial registers ClinicalTrials.gov and ICTRP Search Portal. The date of the last search was July 2016 for all databases. The study selected randomized controlled trials (RCTs) of diet, physical activity and behavioral interventions for treating overweight or obesity in adolescents aged 12 to 17 years. Two review authors independently assessed risk of bias, evaluated the overall quality of the evidence using the GRADE instrument and extracted data following the guidelines of the Cochrane Handbook for Systematic Reviews of Interventions. The authors included 44 completed RCTs (4781 participants) and 50 ongoing studies. The number of participants in each trial varied (10 to 521) as did the length of follow-up (6 to 24 months). Participants ages ranged from 12 to 17.5 years in all trials that reported mean age at baseline.

In [19], the goal of the study was to develop and validate predictive models for detecting undiagnosed diabetes using data from the Longitudinal Study of Adult Health (ELSA-Brasil) and to compare the performance of different machine-learning algorithms in this task. After selecting a subset of 27 candidate variables from the literature, models were built and validated in four sequential steps: (i) parameter tuning with tenfold cross-validation, repeated three times; (ii) automatic variable selection using forward selection, a wrapper strategy with four different machine-learning algorithms and tenfold cross-validation (repeated three times), to evaluate each subset of variables; (iii) error estimation of model parameters with tenfold cross-validation, repeated ten times; and (iv) generalization testing on an independent dataset. The models were created with the following machine-learning algorithms: logistic regression, artificial neural network, naïve Bayes, K-nearest neighbor and random forest.

In [20], authors assessed risk estimates of liver injury associated with antibiotic use in children and adolescent outpatients. A large, multi-database, population-based, case-control study was performed in people <18 years of age from two European countries (Italy and The Netherlands) during the period 2000-2008. All potential cases of liver injury were automatically extracted from three databases and then manually validated based on Council for International Organizations of Medical Sciences (CIOMS) criteria and by exclusion of all competing causes for liver injury. Multivariate conditional logistic regression analyses were applied to calculate odds ratios (ORs) as a measure of the association (with 95% confidence interval [CI]), and 938 cases liver injury were identified, concluding that antibiotic-induced liver injury in children is heterogeneous across the use of individual antibiotics.

In [21], the study described a non-targeted metabolomics platform based on UPLC-UHR-QToF-MS(/MS) for the assessment of plasma non-polar metabolites. The method was applied to a longitudinal mouse obesity study comparing mice on control and high fat diet (HFD), respectively. Plasma metabolites were assessed 2, 4, 8 and 16 weeks after initiation of feeding. Multivariate analysis of the metabolite dataset showed clear differentiation of the feeding groups after 8 weeks when the HFD-fed mice exhibited clear signs of insulin resistance. The discrimination of the groups was due to changes in various metabolic pathways including, among others, glycerophospholipid, sphingolipid and cholesterol metabolism. Thirteen of these observed metabolites are known key metabolites to diabetes or its secondary diseases like diabetic nephropathy and neuropathy.

In [22], the authors conducted a series of bibliometric analyses on the related literature, including papers' production trends in the field and the trend of each paper's co-author number, the distribution of core institutions and countries, the core literature distribution, the related information of prolific authors and innovation paths in the field, a keyword co-occurrence analysis, and research hotspots and trends for the future. The study found the following: (a) In the early stage, researchers from the United States, the People's Republic of China, the United Kingdom, and Germany made the most contributions to the literature associated with healthcare big data research and the innovation path in this field. (b) The innovation path in healthcare big data consists of three stages: the disease early detection, diagnosis, treatment, and prognosis phase, the life and health promotion phase, and the nursing phase. (c) Research hotspots are mainly concentrated in three dimensions: the disease dimension (e.g., epidemiology, breast cancer, obesity, and diabetes), the technical dimension (e.g., data mining and machine learning), and the health service dimension (e.g., customized service and elderly nursing).

In [23], the study sought to quantify (1) the surgical risk, and (2) the costs associated with complications after THA in patients who were morbidly obesity (BMI $\geq$ 40 kg/m2) or super obese (BMI $\geq$ 50 kg/m2). This was a retrospective study of patients, using Medicare hospital claims data, who underwent THA. Twelve complications occurring during the 90 days after THA were analyzed using multivariate Cox models adjusting for patient demographic, comorbidities, and institutional factors. The conclusions showed that patients who are super obese are at increased risk for serious complications compared with patients with morbid obesity, whose risks are elevated relative to patients whose BMI is less than 40 kg/m2.

In [24], the authors presented a method of identifying obesity automatically using text mining techniques and information related to body weight measures and obesity comorbidities. They used a dataset of 3015 de-identified medical records that contain labels for two classification problems. The first classification problem distinguishes between obesity, overweight, normal weight, and underweight. The second classification problem differentiates between obesity types: super obesity, morbid obesity, severe obesity and moderate obesity. The study implemented two approaches: a hierarchical method and a nonhierarchical one. The authors used Support Vector Machine and Naïve Bayes together with ten-fold cross validation to evaluate and compare performances. In general, the results showed that Support Vector Machine obtained better performances than Naïve Bayes for both classification problems.

In [25], their goal was to investigate the possibility that metformin, the primary oral hypoglycemic agent in use worldwide, may influence the progression of AAA disease. Preoperative AAA patients with diabetes were identified from an institutional database. After tabulation of individual cardiovascular and demographic risk factors and prescription drug regimens, odds ratios for categorical influences on annual AAA enlargement were calculated through nominal logistical regression. Experimental AAA modeling experiments were subsequently performed in normoglycemic mice to validate the database-derived observations as well as to suggest potential mechanisms of metformin-mediated aneurysm suppression.

In [26], the study had the objective to define the natural history of patients with Isolated metabolic syndrome. Metabolic syndrome (MS) is associated with increased risk of cardiovascular mortality. Isolated MS patients are a subset of MS patients who don't meet the diagnostic criteria of hypertension (HTN) and diabetes mellitus(DM). Data was collected prospectively on a population-based random sample of 1042 Olmsted County, Minnesota, residents aged 45 years or older who underwent clinical evaluation, medical record abstraction, and echocardiography (Visit 1 January 1,1997-December 31, 2000). Isolated MS was associated with increased risk for the development of hypertension, diabetes and obesity, but not increased mortality or heart failure over an eight year period compared to healthy controls. Future studies should determine if aggressive management of risk factors in Isolated MS will prevent progression to MS.

In [27], the goal of this study was to understand research trends and collaboration patterns together with scholarly impact within the domain of global obesity research. The authors developed and analyzed bibliographic affiliation data collected from 117,340 research articles indexed in Scopus database on the topic of obesity and published from 1993-2012. They found steady growth and an exponential increase of publication numbers. The highest publication output was from the USA - 42% of publications had at least one author from the USA. Overall, this study provides one of the most comprehensive longitudinal bibliometric analyses of obesity research. This should help in understanding research trends, spatial density, collaboration patterns and the complex multi-disciplinary nature of research in the obesity domain.

In [28], the authors studied the deposits of fat on the surroundings of the heart are correlated to several health risk factors such as atherosclerosis, carotid stiffness, coronary artery calcification, atrial fibrillation and many others. These deposits vary unrelated to obesity, which reinforces its direct segmentation for further quantification. However, manual segmentation of these fats has not been widely deployed in clinical practice due to the required human workload and consequential high cost of physicians and technicians. In this work, the authors proposed a unified method for an autonomous segmentation and quantification of two types of cardiac fats. They compared the performance of several classification algorithms on this task, including neural networks, probabilistic models and decision tree algorithms. Experimental results of the proposed methodology showed that the mean accuracy regarding both epicardial and mediastinal fats is 98.5% (99.5% if the features are normalized), with a mean true positive rate of 98.0%.

In [29], the study indicated that heart disease is the leading cause of death globally and a significant part of the human population lives with it. A number of risk factors have been recognized as contributing to the disease, including obesity, coronary artery disease (CAD), hypertension, hyperlipidemia, diabetes, smoking, and family history of premature CAD. The authors described and evaluated a methodology to extract mentions of such risk factors from diabetic clinical notes, which was a task of the i2b2/UTHealth 2014 Challenge in Natural Language Processing for Clinical Data. The methodology was knowledge-driven and the system implemented local lexicalised rules (based on syntactical patterns observed in notes) combined with manually constructed dictionaries that characterize the domain. A part of the task was also to detect the time interval in which the risk factors were present in a patient.

In [30], the authors had the goal of studying coronary artery disease (CAD), which is the leading cause of death in both the UK and worldwide. The detection of related risk factors and tracking their progress over time is of great importance for early prevention and treatment of CAD. The study described an information extraction system that was developed to automatically identify risk factors for heart disease in medical records while the authors participated in the 2014 i2b2/UTHealth NLP Challenge. Their approaches relied on several nature language processing (NLP) techniques such as machine learning, rule-based methods, and dictionary-based keyword spotting to cope with complicated clinical contexts inherent in a wide variety of risk factors. The system achieved encouraging performance on the challenge test data with an overall micro-averaged F-measure of 0.915, which was competitive to the best system (F-measure of 0.927) of this challenge task.

In [31], the study described a rule-based system developed using a combination of regular expressions, concepts from the Unified Medical Language System (UMLS), and freely-available resources from the community. With a performance (F1=90.7) that is significantly higher than the median (F1=87.20) and close to the top performing system (F1=92.8), it was the best rule-based system of all the submissions in the 2014 i2b2 challenge. The authors also used this system to evaluate the utility of different terminologies in the UMLS towards the challenge task. Of the 155 terminologies in the UMLS, 129 (76.78%) have no representation in the corpus. The Consumer Health Vocabulary had very good coverage of relevant concepts and was the most useful terminology for the challenge task. While segmenting notes into sections and lists has a significant impact on the performance, identifying negations and experiencer of the medical event results in negligible gain.

In [32], the study focused on identifying risk factors for heart disease (specifically, Cardiac Artery Disease) in clinical narratives. The authors used a "light" annotation paradigm to annotate a set of 1304 longitudinal medical records describing 296 patients for risk factors and the times they were present. They designed the annotation task for this track with the goal of balancing annotation load and time with quality, so as to generate a gold standard corpus that can benefit a clinically-relevant task. Then, they applied light annotation procedures and determined the gold standard using majority voting. On average, the agreement of annotators with the gold standard was above 0.95, indicating high reliability. The resulting document-level annotations generated for each record in each longitudinal EMR in this corpus provide information that can support studies of progression of heart disease risk factors in the included patients over time. These annotations were used in the Risk Factor track of the 2014 i2b2/UTHealth shared task. Participating systems achieved a mean micro-averaged F1 measure of 0.815 and a maximum F1 measure of 0.928 for identifying these risk factors in patient records.

In [33], the authors expressed in the United States about 600,000 people die of heart disease every year. The annual cost of care services, medications, and lost productivity reportedly exceeds 108.9 billion dollars. Effective disease risk assessment is critical to prevention, care, and treatment planning. Recent advancements in text analytics have opened up new possibilities of using the rich information in electronic medical records (EMRs) to identify relevant risk factors. The 2014 i2b2/UTHealth Challenge brought together researchers and practitioners of clinical natural language processing (NLP) to tackle the identification of heart disease risk factors reported in EMRs. The authors participated in this track and developed an NLP system by leveraging existing tools and resources, both public and proprietary. Their system was a hybrid of several machine-learning and rule-based components. The system achieved an overall F1 score of 0.9185, with a recall of 0.9409 and a precision of 0.8972.

In the study [34], the authors quantified the trade-off between the speed and simplicity of dictionary-based term recognition and the richer linguistic information provided by more advanced natural language processing (NLP) among text processing systems that make different trade-offs between speed and linguistic understanding. They tested both types of systems in three clinical research tasks: phase IV safety profiling of a drug, learning adverse drug-drug interactions, and learning used-to-treat relationships between drugs and indications. The study benchmarked the accuracy of the NCBO Annotator and REVEAL in a manually annotated, publically available dataset from the 2008 i2b2 Obesity Challenge. The authors then applied the NCBO Annotator and REVEAL to 9 million clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE) and used the resulting data for three research tasks. The study concluded there is no significant difference between using the NCBO Annotator and REVEAL in the results of the three research tasks when using large datasets. In one subtask, REVEAL achieved higher sensitivity with smaller datasets.

In the study [35], the composition of the intestinal microbiota of 92 healthy Japanese men was measured following consumption of identical meals for 3 days; terminal restriction fragment length polymorphisms were then used to analyze the DNA content of their feces. The obtained operational taxonomic units (OTUs) were further analyzed using seven restriction enzymes: 516f-BslI and -HaeIII, 27f-MspI and -AluI, and 35f-HhaI, -MspI and -AluI. Subjects were classified by their body mass index (BMI) as lean (<18.5) or obese (>25.0). OTUs were then analyzed using data mining software. Pearson correlation coefficients on data mining results indicated only a weak relationship between BMI and OTU diversity. Specific OTUs attributed to lean and obese subjects were further examined by data mining with six groups of enzymes and closely related accession numbers for lean and obese subjects were successfully narrowed down. 16S rRNA sequences showed Bacillus spp., Erysipelothrix spp. and Holdemania spp. to be present among 30 bacterial candidates related to the lean group. Fifteen candidates were classified Firmicutes, one was classified as Chloroflexi, and the others were not classified. 45 Microbacteriaceae, 11 uncultured Actinobacterium, and 3 other families were present among the 119 candidate OTUs related to obesity. The study concluded that the presence of Firmicutes and Actinobacteria may be related to the BMI of the subject.

In the study [36], the authors were interested to determine if the community food environment may contribute to obesity by influencing food choice. The purpose of the study was to develop and validate reduced-item food environment audit tools for stores and restaurants. Nutrition Environment Measures Surveys for stores (NEMS-S) and restaurants (NEMS-R) were completed in 820 stores and 1,795 restaurants in West Virginia, San Diego, and Seattle. Data mining techniques (correlation-based feature selection and linear regression) were used to identify survey items highly correlated to total survey scores and produce reduced-item audit tools that were subsequently validated against full NEMS surveys. Regression coefficients were used as weights that were applied to reduced-item tool items to generate comparable scores to full NEMS surveys. Data were collected and analyzed in 2008-2013. There were no significant differences in median scores for varying types of retail food outlets when compared to the full survey scores. Median in-store audit time was reduced 25%-50%. Their conclusion was reduced-item audit tools can reduce the burden and complexity of large-scale or repeated assessments of the retail food environment without compromising measurement quality.

In [37], the authors looked to enlighten the complex etiology beneath obesity by analyzing data from a large nutrigenetics study, in which nutritional and genetic factors associated with obesity were recorded for around two thousand individuals. These data have been analyzed using artificial neural network methods, which identified optimized subsets of factors to predict one's obesity status. These methods did not reveal though how the selected factors interact with each other in the obtained predictive models. For that reason, parallel Multifactor Dimensionality Reduction pMDR was used to further analyze the pre-selected subsets of nutrigenetic factors. Within pMDR, predictive models using up to eight factors were constructed, further reducing the input dimensionality, while rules describing the interactive effects of the selected factors were derived. In this way, it was possible to identify specific genetic variations and their interactive effects with particular nutritional factors, which are now under further study.

In [38] the aim of this study was to identify risk patterns for type 2 diabetes incidence using association rule mining (ARM). A population of 6647 individuals without diabetes, aged ≥ 20 years at inclusion, was followed for 10-12 years, to analyze risk patterns for diabetes occurrence. Study variables included demographic and anthropometric characteristics, smoking status, medical and drug history and laboratory measures. In the case of women, the results showed that impaired fasting glucose (IFG) and impaired glucose tolerance (IGT), in combination with body mass index (BMI) ≥ 30 kg/m2, family history of diabetes, wrist circumference > 16.5 cm and waist to height ≥ 0.5 can increase the risk for developing diabetes. For men, a combination of IGT, IFG, length of stay in the city (> 40 years), central obesity, total cholesterol to high density lipoprotein ratio ≥ 5.3, low physical activity, chronic kidney disease and wrist circumference > 18.5 cm were identified as risk patterns for diabetes occurrence. The authors showed that ARM is a useful approach in determining which combinations of variables or predictors occur together frequently, in people who will develop diabetes. The ARM focuses on joint exposure to different combinations of risk factors, and not the predictors alone.

In this study [39], an automated case-finding EHR-derived algorithm was designed to identify women of child-bearing age having outpatient encounters in an 85-site, integrated health system. The algorithm simultaneously cross-referenced multiple discrete data fields to identify selected preconception factors (obesity, hypertension, diabetes, teratogen use including ACE inhibitors, multivitamin supplementation, anemia, renal insufficiency, untreated sexually transmitted infection, HIV positivity, and tobacco, alcohol or illegal drug use). Concordance was assessed between the algorithm output, survey results, and manual data abstraction. Using the patient survey results as a reference point, health-factor agreement was similar comparing the algorithm (85.8 %) and the chart abstraction (87.2 %) results. Incorrect or missing data entries in the EHR encounters were largely responsible for discordances observed. Preconception screening using an automated algorithm in a system-wide EHR identified a large group of women with potentially modifiable preconception health conditions. The issue most responsible for limiting algorithm performance was incomplete point of care documentation. Accurate data capture during patient encounters should be a focus for quality improvement, so that novel applications of system-wide data mining can be reliably implemented.

In [40], This study examined clustering of factors associated with low fitness in adolescents in order to best target public health interventions for young people. 1147 children were assessed for fitness, had blood samples, anthropometric measures and all data were linked with routine electronic data to examine educational achievement, deprivation and health service usage. Factors associated with fitness were examined using logistic regression, conditional trees and data mining cluster analysis. Focus groups were conducted with children in a deprived school to examine barriers and facilitators to activity for children in a deprived community. Unfit adolescents are more likely to be deprived, female, have obesity in the family and not achieve in education. There were 3 main clusters for risk of future heart disease/diabetes (high cholesterol/insulin); children at low risk (not obese, fit, achieving in education), children ‘visibly at risk’ (overweight, unfit, many hospital/GP visits) and ‘invisibly at risk’ (unfit but not overweight, failing in academic achievement). The authors concluded that low fitness in the non-obese child can reveal a hidden group who have high risk factors for heart disease and diabetes but may not be identified as they are normal weight. In deprived communities low fitness is associated with non-achievement in education but in non-deprived communities low fitness is associated with female gender.

In [41], the authors expressed that numerous clinical epidemiology and laboratory studies have demonstrated a close association between inflammation or immunity and the occurrence of hypertension. Previous studies have shown that upon intraperitoneal injection of lipopolysaccharide (LPS) or zymosan (Zym) into pregnant rats, the offspring of the experimental group exhibited higher blood pressure levels compared with the control group, and these may be associated with the effects of an accumulation of gene expression changes on female rats during pregnancy. The study used the Affymetrix GeneChip® Rat Genome 230 2.0 Array to examine gene expression profile changes in the embryos of experimental pregnant rats intraperitoneally injected with LPS or Zym vs. the untreated controls. These changes in gene expression may affect the developmental and metabolic status of the offspring, thereby increasing their susceptibility to hypertension and obesity.

In the study [42], Health eTools for Schools was developed to assist school nurses with routine entries, including height and weight, on student health records, thus providing a readily accessible data base. Data-mining techniques were applied to this database to determine if clinically significant results could be generated. Body mass index (BMI) data collected and entered in eTools by school nurses from 657,068 students attending 1156 schools in 49 of 67 Pennsylvania counties during 2005-2009 were analyzed. Students in each BMI category were sorted; regression was used to model mean and percentage trends. A chi-square test of individually matched BMI percentages was computed and migration across normal, overweight, and obese states determined. The highest percentage of obese students occured in middle school. The mean trends for obesity and overweight had increasing slopes of 0.189 and 0.227, respectively; with regression slope for overweight >59%. Within groups, substantial percentages of individually matched BMIs changed significantly (p < .0001) over 2 years, migrating between normal weight, overweight, and obese. The means trends for both overweight and obesity were greater in 2009 than in 2005, increasing steadily to 2008 and slightly declining to 2009. The dominant overall pattern flows from overweight to obese. If continued unabated, percentage of students who are obese will dominate over time.

III.METHODOLOGY

Literacy review is a key process for research, showing new insights for future studies, for the literacy analysis was needed to determine research question, keywords, search strings, inclusion and exclusion criteria, quality validation and data recollection. In Figure 1, you can see the methodology used for the recollection information stage.

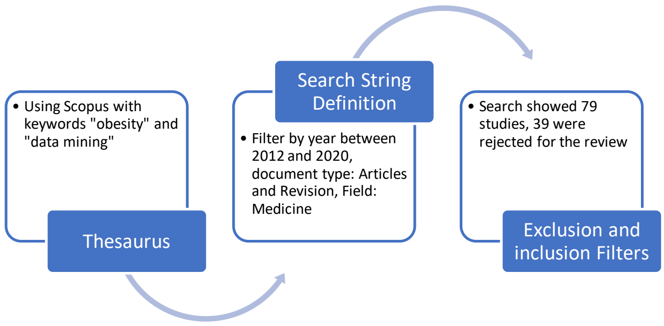
Using Scopus with keywords "obesity" and "data mining"
Thesaurus
Search String Definition
Filter by year between 2012 and 2020, document type: Articles and Revision, Field: Medicine
Search showed 79 studies, 39 were rejected for the review
Exclusion and inclusion Filters

Figure 1: Methodology applied for the literacy review

The use of thesaurus allowed to determine the correct terminology for the information search and obtain the right results for the literacy review. The thesaurus used was provided by UNESCO, to choose the most appropriate keywords such as: obesity, nutritional disease, data analysis, and data processing.

After choosing the relevant keywords for the theme, the search string was defined as you can see in Table 1.

Table 1: Search string for the literacy review

(TITLE-ABS-KEY (obesity) AND TITLE-ABS-KEY ("data mining")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re")) AND (LIMIT-TO (SUBJAREA , "MEDI")) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010)) AND (EXCLUDE (PUBYEAR , 2012) OR EXCLUDE (PUBYEAR , 2011) OR EXCLUDE (PUBYEAR , 2010))

The inclusion and exclusion criteria for the studies in the literacy review, is presented in Table 2.

Table 2: Inclusion and exclusion criteria

Inclusion Criteria	**Exclusion Criteria**
Studies include the application of a data mining technique or algorithm for detecting factors related to obesity problems or overweight.	Studies do not include any application of a data mining technique or algorithm for detecting factors related to obesity problems or overweight.

The selection of the studies obtained 40 scientific articles, with diverse topics but all applied data mining techniques or algorithms to detect obesity issues.

After selection of the documents, the compilation and the quality validation was performed, the later stages of the methodology included: scientometric analysis, technical analysis, characterization and results.

IV. SCIENTOMETRIC ANALYSIS

Scientometric analysis quantifies some relevant aspects inside the target documents such as: number of publications yearly, number of articles per journal, number of articles per quartile and number of articles based on publication type (Book – Journal – Proceedings).

In Figure 2, you can see the number of articles related to the theme by year, showing an increase in the number of publications between 2015 and 2017, which represents this research field has been supported by many authors since 5 years ago, looking for opportunities to create solutions using data mining tools in this area.

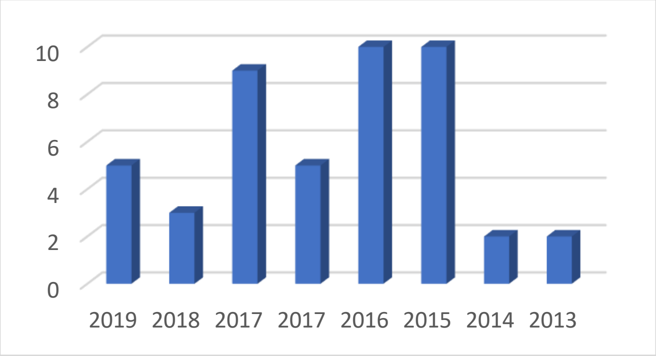

Figure 2: Number of articles by year

In Figure 3, you can see the number of articles by journal, to identify trends in authors to publish in certain journals or publications based on the field of study. Journal of Biomedical Informatics (ISSN: 1532-0464) and Molecular Medicine Reports (Online ISSN: 1791-3004) have the highest number of publications in the analysis.

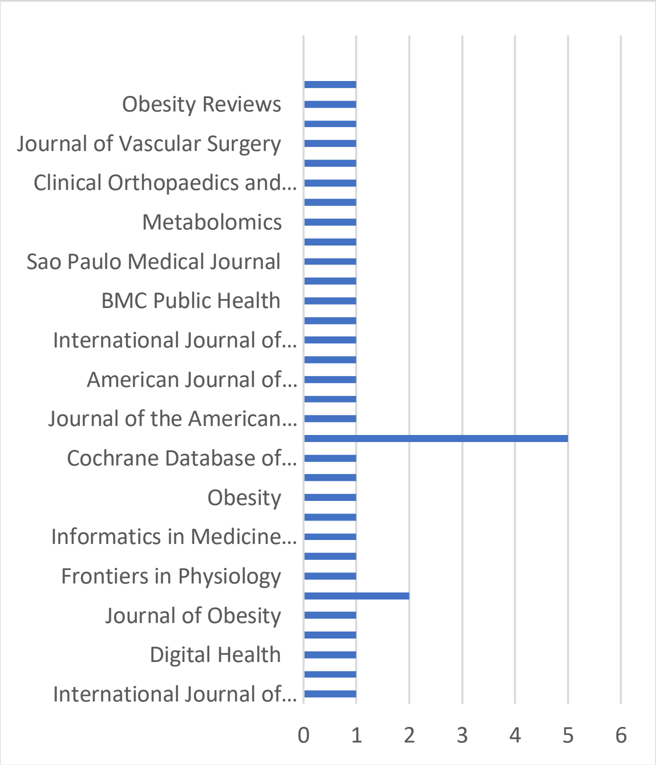

Figure 3: Number of articles by publisher

In Figure 4, you can see the number of articles by quartile, according to Scopus, this information indicates the quality or impact factor of the study, in this scientific field specifically. The analysis showed that 54% of the articles have a high impact factor, located in the Q1 quartile, also 20% can be found in the Q2 quartile and 26% of the studies are found in the Q3 Quartile for this literacy review.

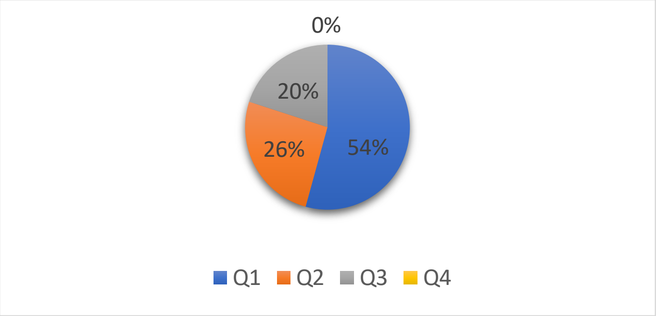

Figure 4: Number of articles by quartile (Scopus)

V. TECHNICAL ANALYSIS

Technical analysis quantifies some relevant aspects inside the target documents such as: number of publications by disease type, data mining methods implemented, and the datasets used for authors in the studies compiled by the literacy review.

In Figure 5, you can see the number of studies with diseases related to obesity issues, there is a clear trend to implement solutions for obesity but there is a group of studies involved with diseases derived from obesity too.

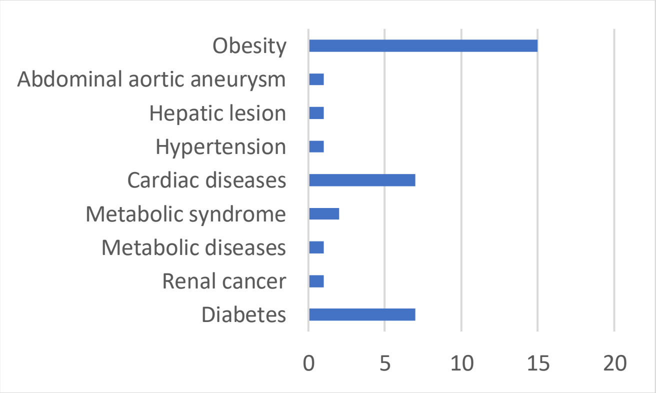

Figure 5: Number of articles by condition (disease)

In Table 3, you can see the methods commonly used to implement early and precise detection of obesity issues, according to this, Logistic Regression is the most common method so far in the studies, appearing in 6 articles, followed by Decision Trees, used in 3 studies, Neural Networks, Simple Regression, Random Forest and Cox Regression were implemented in 2 studies respectively.

Table 3: Data mining methods used by studies in the literacy review

Methods	Studies
Bayesian General Network (GBN)	[6]
Neural Networks (ANN)	[7] [28]
Simple Regression	[8] [9]
Logistic Regression	[9] [10] [19] [20] [25] [26]
Multivariant Adaptive Regression Spline (MARS)	[10]
Random Forest (RF)	[10] [19]
Boosting Decision Trees (BT)	[10]
Apriori Association	[13]
Decision Trees	[15] [28] [35]
Conditional Inference Tree	[16]
Machine Learning	[17]
K-Nearest Neighbor	[19]
Cox Regression	[23] [26]
Support Vector Machines	[24]
Exponential Regression	[27]
Multifactor Dimensionality Reduction	[37]

In Table 4, you can see the datasets used by the authors in the literacy review, most studies include a database created by the authors, accounting by 10 studies with these feature, DCT and STRIDE databases are used two times , and finally there are other databases like PubMed-NCBI, ANZCTRN, LILACS, ELSA STUDY BASELINE, HFD, HGGB EMR SYSTEM, NEMS and SAIL DATABANK.

Table 4: Datasets commonly used by studies in the literacy review

Datasets	Studies
Proprietary Database	[8] [9] [17][20] [23] [35] [39] [42] [6] [7]
PubMed-NCBI	[12]
ANZCTRN	[18]
LILACS	[18]
ELSA STUDY BASELINE	[19]
HFD	[21]
HGGB EMR SYSTEM	[24]
STRIDE	[25] [31]
DCT	[32] [33]
NEMS	[36]
SAIL DATABANL	[40]

VI. DISCUSSION

According to the research questions for this study, the relevance of the systematic review falls in recognizing the techniques or algorithms in data mining more important for detecting factors related to overweight or obesity problems.

This section of the paper pretends to answer the next questions:
¿What are the techniques or algorithms in data mining commonly used to detect factors that affect the appearance of obesity issues?

In this review, you can find more than 16 techniques or algorithms in data mining, highlighting Simple Regression, a technique used in seven articles ([8],[9], [10], [19], [20], [25]), this method explains the relationship between a response variable Y and one single explanatory variable X. Another technique found in the review is Decision Trees, in three articles ([15],[28],[35]), this analytic method uses an schematic representation to facilitate decision making processes. These two techniques are widely used, but the information is insufficient to indicate one of them is better for detecting obesity factors. Nevertheless, Simple Regression is the most used of the two of them.

¿What is the reliability of the results found by a data mining model to detect obesity problems?

In general terms, data mining techniques are defined as processes for extracting hidden patterns in the data, these techniques can have a predictive or descriptive focus, with very specific tasks, producing quite different studies and in consequence different results. Although, using them properly and setting the terms and patterns for them, you can obtain a high reliability for factors involved with obesity issues in the target population.
Most studies used their own dataset; ¿this situation affects the credibility of their research?

There is no relation between the use of their own dataset and the credibility of their studies, in other hand, you can infer that creating their own dataset, shows the dedication of the authors with their research.

VII. CONCLUSION

This systematic review of the literacy had the goal to identify and suggest the techniques and algorithms of data mining most commonly used to detect factors involved or related to the appearance of obesity issues, and to analyze the reliability of these methods in a data mining model. Data mining is widely used in the field of medicine, to analyze risk factors for obesity, especially due their precision and practicality in data analysis. Many authors have designed solutions based on these techniques and predictive models for other diseases and conditions, including obesity.

The technical analysis of the study showed that Simple Regression and Decision Trees, are the most common methods to recognize obesity levels, 19% of the reviewed articles used Simple Regression and 11% of them, used Decision Trees. Also, the results identified that 25% of the studies were related to cardiac diseases and 25% were related to diabetes, conditions that a person with obesity can contract easily and represent a considerable risk.

VIII.FUTURE WORKS

In future works, the authors propose a systematic review of the literacy to identify the most effective methods to detect diseases acquired in presence of obesity. These diseases would be divided in cardiac diseases and other kind of conditions as diabetes type 1 and diabetes type 2. This works would contribute in the advances of detecting these two types of diseases that affect a vast number of people suffering from obesity.

REFERENCES

[1] OMS. (2020, 1 de abril). Obesidad y sobrepeso (Notas descriptivas) [Online]. En: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

[2] P. Werner, E. Meiss, L. Scheja, J. Heeren y M. Fischer, “Metabolite profiling: development and application of an UHR-QTOF-MS(/MS) method approach for the assessment of metabolic changes in high fat diet fed mice”, P. Werner, E. Meiss, L. Scheja, J. Heeren y M. Fischer, Metabolomics, Volume 13, Issue 4, 1 April 2017, Article number 44.

[3] [M. Meller MD, PhD, N. Toossi MD, M. H. Gonzalez MD, PhD, M. Son PhD, E. C. Lau MS, N. Johanson MD, “Surgical Risks and Costs of Care are Greater in Patients Who AreSuperObese and Undergoing THA”, Clinical Orthopaedics and Related Research, Volume 474, Issue 11, 1 November 2016, Pages 2472-2481

[4] R. L. Figueroa y C. A. Flores, “Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures”, Journal of Medical Systems, Volume 40, Issue 8, 1 August 2016, Article number 191.

[5] Karayianni, KN a , Grimaldi, KA a , Nikita, KS a , Valavanis, IK b, International Journal of Bioinformatics Research and Applications, Volume 11, Issue 3, 2015, Pages 233-246, 2015

[6] C. Kim, F. Costello, K. Lee, Y. Li and C. Li, "Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis", International Journal of Environmental Research and Public Health, vol. 16, no. 23, p. 4684, 2019.

[7] L. Vigna., A. Silvia, E. Grossi, S. Turolo, L. Tomaino, F. Napolitano, "Directional Relationship Between Vitamin D Status and Prediabetes: A New Approach from Artificial Neural Network in a Cohort of Workers with Overweight-Obesity", Journal of the American College of Nutrition, vol. 38, no. 8, pp. 681-692, 2019.

[8] H. Jelinek, A. Stranieri, A. Yatsko and S. Venkatraman, "Personalised measures of obesity using waist to height ratios from an Australian health screening program", DIGITAL HEALTH, vol. 5, p. 205520761984436, 2019.

[9] L. Wen, Y. Yu, H. Lv, Y. He and B. Yang, "FTO mRNA expression in the lower quartile is associated with bad prognosis in clear cell renal cell carcinoma based on TCGA data mining", Annals of Diagnostic Pathology, vol. 38, pp. 1-5, 2019.

[10] [P. Shi and J. Goodson, "A Data Mining Approach Identified Salivary Biomarkers That Discriminate between Two Obesity Measures", Journal of Obesity, vol. 2019, pp. 1-7, 2019.

[11] T. Assmann, F. Milagro and J. Martínez, "Crosstalk between microRNAs, the putative target genes and the lncRNA network in metabolic diseases", Molecular Medicine Reports, 2019.

[12] C. Johnson, C. Drummer, A. Virtue, T. Gao, S. Wu, M. Hernandez, L. Singh, H. Wang, X. Yang, "Increased Expression of Resistin in MicroRNA-155-Deficient White Adipose Tissues May Be a Possible Driver of Metabolically Healthy Obesity Transition to Classical Obesity", Frontiers in Physiology, vol. 9, 2018.

[13] V. Guan, Y. Probst, E. Neale, M. Batterham and L. Tapsell, "Identifying usual food choices at meals in overweight and obese study volunteers: implications for dietary advice", British Journal of Nutrition, vol. 120, no. 4, pp. 472-480, 2018.

[14] S. Harous, M. El Menshawy, M. Serhani and A. Benharref, "Mobile health architecture for obesity management using sensory and social data", Informatics in Medicine Unlocked, vol. 10, pp. 27-44, 2018.

[15] M. Tayefi, M. Saberi-Karimian, H. Esmaeili, A. Zadeh, M. Ebrahimi, M. Mohebati, A. Heidari-Bakavoli, M. Reza, M. Heshmati, M. Safarian, S. Reza, G. Ferns, M. Ghayour-Mobarhan, "Evaluating of associated risk factors of metabolic syndrome by using decision tree", Comparative Clinical Pathology, vol. 27, no. 1, pp. 215-223, 2017.

[16] F. Cheng, X. Gao, L. Bao, D. Mitchell, C. Wood, M. Sliwinski, H. Smiciklas-Wright, C. Still, D. Rolston, G. Jensen, "Obesity as a risk factor for developing functional limitation among older adults: A conditional inference tree analysis", Obesity, vol. 25, no. 7, pp. 1263-1269, 2017.

[17] Q. Nguyen, H. Meng, S. Kath, M. McCullough, D. Paul, P. Kanokvimankul, T. Nguyen, F. Li, "Social media indicators of the food environment and state health outcomes", Public Health, vol. 148, pp. 120-128, 2017.

[18] L. Al-Khudairy, E. Loveman, J. Colquitt, E. Mead, R. Johnson, H. Fraser, J. Olajide, M. Murphy, R. Marian, C. O’Malley, L. Azevedo, L. Ells, M. Metzendorf, K. Ress, "Diet, physical activity and behavioural interventions for the treatment of overweight or obese adolescents aged 12 to 17 years", Cochrane Database of Systematic Reviews, 2017.

[19] A. R. Olivera, V. Roesler, C. Iochpe, M. I. Schmidt, Á. Vigo, S. M. Barreto, “Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil: accuracy study”, Sao Paulo Medical Journal, Volumen 135, Número 3, mayo-junio de 2017, páginas 234-246.

[20] C. Ferrajolo, K. M. C. Verhamme, G. Trifiro, G. W. ‘t Jong. G. Picelli, C. Giaquinto, G. Mazzaglia, B. H. Stricker, F. Rossi. A. Capuano y M. C. J. M. Sturkenboom. “Antibiotic-Induced Liver Injury in Paediatric Outpatients: A Case-Control Study in Primary Care Databases”, Drug Safety, Volume 40, Issue 4, 1 April 2017, Pages 305-315.

[21] P. Werner, E. Meiss, L. Scheja, J. Heeren y M. Fischer, “Metabolite profiling: development and application of an UHR-QTOF-MS(/MS) method approach for the assessment

of metabolic changes in high fat diet fed mice", Metabolomics, Volume 13, Issue 4, 1 April 2017, Article number 44.

[22] D. Gu, J. Li, X. Li, C. Liang, "Visualizing the knowledge structure and evolution of big data research in healthcare informatics", International Journal of Medical Informatics, Volume 98, 1 February 2017, Pages 22-32.

[23] M. Meller MD, PhD, N. Toossi MD, M. H. Gonzalez MD, PhD, M. Son PhD, E. C. Lau MS, N. Johanson MD, "Surgical Risks and Costs of Care are Greater in Patients Who Are Super Obese and Undergoing THA", Clinical Orthopaedics and Related Research, Volume 474, Issue 11, 1 November 2016, Pages 2472-2481

[24] R. L. Figueroa y C. A. Flores, "Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures", Journal of Medical Systems, Volume 40, Issue 8, 1 August 2016, Article number 191.

[25] N. Fujimura, J. Xiong, PhD, E. B. Kettler, H. Xuan, K. J. Glover, M. W. Mell, B. Xu, and R. L. Dalman, "Metformin treatment status and abdominal aortic aneurysm disease progression", Journal of Vascular Surgery, Volume 64, Issue 1, 1 July 2016, Pages 46-54.e8.

[26] P. A. Patel, C. G. Scott, R. J. Rodeheffer and H. Chen, "The Natural History of Patients with Isolated Metabolic Syndrome", Mayo Clinic Proceedings, Volume 91, Issue 5, 1 May 2016, Pages 623-633.

[27] A. Khan, N. Choudhury, S. Uddin, L. Hossain, and L. A Baur, "Longitudinal trends in global obesity research and collaboration: a review using bibliometric metadata", Obesity Reviews, Volume 17, Issue 4, 1 April 2016, Pages 377-385.

[28] É.O. Rodrigues, F.F.C. Morais, N.A.O.S. Morais, L.S. Conci, L.V. Neto y A. Conci, "A novel approach for the automated segmentation and volume quantification of cardiac fats on computed tomography", Computer Methods and Programs in Biomedicine, Volume 123, 1 January 2016, Pages 109-128.

[29] G. Karystianis, A. Dehghan, A. Kovacevic, J. A. Keane, and G. Nenadic, "Using local lexicalized rules to identify heart disease risk factors in clinical notes," J. Biomed. Inform., vol. 58, pp. S183–S188, 2015, doi: 10.1016/j.jbi.2015.06.013.

[30] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," J. Biomed. Inform., vol. 58, pp. S171–S182, 2015, doi: 10.1016/j.jbi.2015.09.006.

[31] C. Shivade, P. Malewadkar, E. Fosler-Lussier, and A. M. Lai, "Comparison of UMLS terminologies to identify risk of heart disease using clinical notes," J. Biomed. Inform., vol. 58, pp. S103–S110, 2015, doi: 10.1016/j.jbi.2015.08.025.

[32] A. Stubbs, Ö. Uzuner, "Annotating risk factors for heart disease in clinical narratives for diabetic patients," J. Biomed. Inform., vol. 58, pp. S78–S91, Dec. 2015, doi: 10.1016/j.jbi.2015.05.009.

[33] M. Torii, J. Fan, T. Lee, T. Wiley, D. Zisook, Y. Huang, "Risk factor detection for heart disease by applying text analytics in electronic medical records," J. Biomed. Inform., vol. 58, pp. S164–S170, 2015, doi: 10.1016/j.jbi.2015.08.011.

[34] K. Jung, P. LePendu, S. Iyer, A. Bauer-Mehren, B. Percha, and N. H. Shah, "Functional evaluation of out-of-the-box text-mining tools for data-mining tasks," J. Am. Med. Informatics Assoc., vol. 22, no. 1, pp. 121–131, 2015, doi: 10.1136/amiajnl-2014-002902.

[35] T. Kobayashi, T. Osaki, and S. Oikawa, "Use of T-RFLP and seven restriction enzymes to compare the faecal microbiota of obese and lean Japanese healthy men," Benef. Microbes, vol. 6, no. 5, pp. 735–745, 2015, doi: 10.3920/BM2014.0147.

[36] S. N. Partington, T. J. Menzies, T. A. Colburn, B. E. Saelens, and K. Glanz, "Reduced-item food audits based on the nutrition environment measures surveys," Am. J. Prev. Med., vol. 49, no. 4, pp. e23–e33, 2015, doi: 10.1016/j.amepre.2015.04.036.

[37] K. N. Karayianni, K. A. Grimaldi, K. S. Nikita, and I. K. Valavanis, "Mining nutrigenetics patterns related to obesity: Use of parallel multifactor dimensionality reduction," Int. J. Bioinform. Res. Appl., vol. 11, no. 3, pp. 233–246, 2015, doi: 10.1504/IJBRA.2015.069194.

[38] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, "An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database," Int. J. Endocrinol. Metab., vol. 13, no. 2, 2015, doi: 10.5812/ijem.25389.

[39] H. Straub, M. Adams, and R. K. Silver, "Can an electronic health record system be used for preconception health optimization?," Matern. Child Health J., vol. 18, no. 9, pp. 2134–2140, 2014, doi: 10.1007/s10995-014-1461-8.

[40] R. Charlton, M. Gravenor, A. Rees, G. Knox, R. Hill, M. Rahman, K. Jones, D. Christian, J. Baker, G. Straton, S. Brophy, "Factors associated with low fitness in adolescents - A mixed methods study," BMC Public Health, vol. 14, no. 1, 2014, doi: 10.1186/1471-2458-14-764.

[41] J. Zhou, X. Zhang, H. Zhang, Y. Jia, Y. Liu, Y. Tuang, X. Li, "Use of data mining to determine changes in the gene expression profiles of rat embryos following prenatal exposure to inflammatory stimulants," Mol. Med. Rep., vol. 8, no. 1, pp. 95–102, 2013, doi: 10.3892/mmr.2013.1498.

[42] A. H. Youssefagha, D. K. Lohrmann, and W. P. Jayawardene, "Use of Data Mining to Reveal Body Mass Index (BMI): Patterns Among Pennsylvania Schoolchildren, Pre-K to Grade 12," J. Sch. Health, vol. 83, no. 2, pp. 85–92, 2013, doi: 10.1111/josh.12002.

[43] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in brief, 25, 104344.