Int. j. adv. multidisc. res. stud. 2024; 4(3):1374-1379

Received: 29-04-2024
Accepted: 09-06-2024

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

A Novel Multi-class Classification of Obesity Level using Artificial Neural Network Machine Learning Model

[1] Okwori Anthony Okpe, [2] Odey John Adinya, [3] Oladunjoye John Abiodum
[1, 3] Department of Computer Science, Federal University Wukari, Wukari, Nigeria
[2] Department of Computer Science, University of Calabar, Calabar, Nigeria

Corresponding Author: **Okwori Anthony Okpe**

Abstract
Obesity which is the excessive accumulation of fat impairs an individual health status and thus contributes to numerous chronic diseases, including cancers, diabetes, metabolic syndrome cardiovascular diseases among others. To efficiently predict obesity using the obesity dataset needs a good computational technique such as machine learning techniques that have the power to eradicate data inconsistencies and detect patterns and behaviors of obesity occurrence. This paper aims at predicting the occurance of obesity using the obesity dataset obtained from the University of California Irvine (UCI) machine learning dataset repository. The dataseet was preprocessed by Identification and handling of missing values, Encoding the categorical variables as well as Scaling the dataset before splitting it for training and testing the artificial neural network**.** The Artificial Neural Networ model was built in python programming language using the Jupiter NoteBook as the programming environment along side with other python third-party libraries such as Numpy, Pandas, Sklearn, MatplotLib and Keras. The Artificial Neural Network model performce very well in the multiclass classification of obesity with prediction accuracy of 97%.

Keywords: Obesity, Chronic-Diseases, Predicting, Dataset, Machine-Learning

1. Introduction

Obesity is a medical condition resulted from abnormal or excessive fat accumulation in the body thereby impairing the health condition of an individual. It is usually caused by high level of energy imbalance between the amount of calories consumed and the amount of calories expended from the body. Obesity constitute a major risk factor for various chronic diseases such as cancers, hypertension, diabetes, metabolic syndrome, cardiovascular diseases (Sadaf *et al.*, 2021) [10] among others as a result, it has been identified as the fifth foremost reason for mortality worldwide (Salvador and Andreas, 2017) [11] as those with obesity are more vunerable to high low desity lipoprotine (LDL) cholesterol, low high density lipoprotine (HDL) cholesterol, or high levels of triglycerides (Dyslipidemia), Type II diabetes, Coronary heart disease, stroke, gallbladder disease, Osteoarthritis (a breakdown of cartilage and boe within a joint), sleep apnea breathing problems among others. All this can be unbearable for an obese victim and hence, lead to quick mortality as the patient might easily give up. Several Studies have practically demonstrated that obesity is not a simple problem but rather a complex health issue stemming from a combination of individual factors (such as genetics, learned behaviors) and substantial causes (such as unhealthy societal or cultural eating habits, food deserts) (Ellen *et al.*, 2015; Syahrul *et al., 2015*) [5, 14]. Some researchers also postulated that obesity is an "acquired" disease that, heavily depends on lifestyle factors (i.e., personal choices), such as low rates of physical activity and chronic overeating, despite its genetic and epigenetic influences. Although there is no effective, well-defined, evidence-based intervention for preventing obesity, its prevention requires a complex approach, including interventions at community, family, and individual levels (WHO, 2015; Chiolero, 2018) [15, 2]. Changes to diet and exercise are the main treatments recommended by health professionals. Diet quality can be improved by reducing the consumption of energy-dense foods, such as those high in fat or sugars, and by increasing the intake of dietary fiber (WHO, 2015) [15]. Medications can be used, along with a suitable diet, to reduce appetite or decrease fat absorption. If diet, exercise, and medication are not effective, a gastric balloon or surgery can be performed to reduce stomach volume or length of the intestines, leading to a reduced ability to absorb nutrients from food (*Colquitt et al., 2014*) [3]. To avoid these rigorous control measures, there is augent need to predic obesity at its various level of occurrence using machine learning techniques. To efficiently predict obesity, this paper proposed the implementation of Artificial Neural Network using obesity dataset obtained from UCI machine learning repository.

2. Related Litratures

Richa *et al*, (2021) [9] in their study proposed a thermal imaging method to evaluate childhood obesity based on machine learning techniques. Their objectives were to determine the potential of thermal imaging; to assess the difference in the thermal pattern in various body regions of the studied population; and to compare the performance of feature extraction, feature fusion, feature ranking, and feature dimension reduction (PCA) in the classification of obese and normal children using different Machine learning algorithms. The author obtained about 600 thermograms from various regions such as the abdomen, finger bed, forearm, neck, shank, and gluteal region for their studied population. The authors extracted fifteen statistical textual features from the six regional thermograms followed by implementing feature fusion with scale-invariant feature transform (SIFT) and Speeded up Robust Features (SURF) algorithm. As claimed by the authors, the Principal Component Analysis (PCA) method provides the best classification accuracy for Support Vector Machine (SVM) (98%) followed by Naïve Bayes and Random Forest (97%). They concluded that the regional thermography and computer-aided diagnostic tool with a machine learning classifier could be used as a basic non-invasive prognostic tool for the evaluation of obesity in children.

Casimiro *et al*, (2017) [1] proposed a machine learning approach for the prediction of obesity using a publicly available genetic profile and the manually curated database from the National Human Genome Research Institute Catalog. The authors used indexed genetic variants or Single Nucleotide Polymorphisms as inputs in the various optimized machine learning algorithms for the prediction of obesity. The body mass index status of participants was divided into two classes, normal class, and risk class. A dimensionality reduction task was performed to generate a set of principal variables - 13 SNPs - for the application of various machine learning methods. The models are evaluated using receiver operator characteristic curves and the area under the curve. The machine learning techniques optimized by the authors include gradient boosting, generalized linear model, classification and regression trees, k-nearest neighbors, support vector machines, random forest and multilayer perceptron neural network, which were comparatively assessed in terms of their ability to identify the most important factors among the initial 6622 variables describing genetic variants, age, and gender, to classify a subject into one of the body mass index related classes defined in their study. Their simulation results indicated that the support vector machine generated the highest area under the curve value of 90.5%.

Sri *et al.*, (2021) [13] proposed the prediction of obesity in adults using machine learning techniques. The authors optimized the Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes to identify the presence of obesity using publicly available health data as an attempt to go beyond the traditional prediction models and to compare the performance of three different methods. Their main objective was to establish a set of risk factors for obesity in adults via the studied variables. They further address data imbalance using Synthetic Minority Oversampling Technique (SMOTE) to predict obesity status based on risk factors available in the dataset. As claimed by the authors, the Logistic Regression method shows the highest performance. Nevertheless, kappa coefficients show only moderate concordance between predicted and measured obesity. Location, marital status, age groups, education, sweet drinks, fatty/oily foods, grilled foods, preserved foods, seasoning powders, soft/carbonated drinks, alcoholic drinks, mental-emotional disorders, diagnosed hypertension, physical activity, smoking, and fruit and vegetable consumptions are significant in predicting obesity status in adults as identified from their analytical experiment.

Dugan *et al.*, (2015) [4] in their study also proposed a machine learning technique for the prediction of early childhood obesity using data collected by a clinical decision support system called CHICA. The authors optimized six different machine learning methods namely: RandomTree, RandomForest, J48, ID3, Naïve Bayes, and Bayes trained on the CHICA dataset and identified that an accurate, sensitive model can be created. The result of their experiment proved ID3 as the best performing model with an accuracy of 85% and sensitivity of 89%. Additionally, the ID3 model had a positive predictive value of 84% and a negative predictive value of 88%.

Zhang *et al.*, (2009) [17] in their study on Comparing data mining methods with logistic regression in childhood obesity prediction did an excellent job of comparing the performance metrics of several machine learning techniques. Zhang's group analyzed the Wirral child database, which was limited to basic demographics (like sex) and biometrics (like height, weight, and BMI). By using a broader range of attributes collected by a production clinical decision support system. However, the best sensitivity noted by the author was the logistic regression with 62% sensitivity when looking at obesity prediction after the second birthday. Zhang *et al.*, (2009) [17] did not provide accuracy or specificity for this analysis.

Simmonds *et al*., (2020) [12] conducted a systematic review and thus combined with meta-analysis to examine whether BMI and similar measures can be used to calculate childhood obesity and hence predict adult obesity. Their review supported the conclusion that teenage obesity is a notable public health crisis because it often continues into adulthood. Accordingly, acting to reduce teen obesity can also reduce adult obesity. Early action is one of the most suitable approaches because once children have become overweight, this trend often persists through their adolescence and adulthood.

Hammond *et al*., (2019) [7] in predicting childhood obesity using electronic health record (EHR) data of 3449 children from birth to age 2 developed separate machine learning models for boys and girls to predict obesity at age 5 years and optimized the viability of several machine learning model to perform both binary classification and regression. In each of the separate models for boys and girls, the authors discovered that the weight for length z-score, BMI between 19 and 24 months, and the last BMI measure recorded before age two were the most important features for prediction. Hence, their best performing model "the LASSO regression model" was able to predict obesity with an Area Under the Receiver Operator Characteristic Curve (AUC) of 81.7% for girls and 76.1% for boys. Furthermore, the authors discovered that the optimized regression model outperforms its classification counterparts for the prediction of obesity.

Gupta *et al*., (2019) [6] in their study on obesity prediction adopted a long short-term memory (LSTM) based Recurrent Neural Network (RNN) architecture to predict obesity patterns of 3 consecutive years using a large unaugmented Electronic Health Record (HER) data from 3 previous years records of a large pediatric health system. Their target was to predict obesity for ages between 2-20 years. The authors compared the performance of the RNN model with LSTM cells and other machine learning methods that aggregate over sequential data and ignore temporality. To add interpretability, they included an attention layer to calculate the attention scores for the timestamps and rank features of each timestamp. The AUC of their models were 0.86, 0.78, and 0.78, respectively as claimed by the authors.

Lee *et al.*, (2019) [8] in an approach to identify the risk factors for obesity among children aged 24 to 80 months using longitudinal data from the South Korean National Health Insurance (KNHI) database (with several 1,001,775 families as study participants), applied a decision tree (DT) models to predict obesity as normal or underweight, with the intermediate group (overweight children) being excluded by the authors. They examined the Social-Economic Status **(**SES), parental and child-related factors and conducted a Descriptive statistics and Chi-squared Automatic Interaction Detection (CHAID) for a decision tree model and thus focused on interpreting decision tree (DT) structure to identify risk factors for childhood obesity were the best predictors reported were: Mothers being obese before conception, fathers being obese, non-medical aid beneficiaries, and mothers with hypertension during gestation. Lee *et al.*, (2019) [8] in conclusion discovered that child-related outcome predictors were noncompliance with exclusive breastfeeding, a sugar-sweetened beverage intake ≥200 ml per day, and irregular breakfast consumption.

Xueqin *et al.*, (2021) [16] in an attempt to predict early childhood obesity with machine learning and electronic health record data-optimized seven machine learning models to predict childhood obesity ranging from the age of 2 to 7 years using 2 years captured data record. The Electronic Health Record (EHR) dataset entails data from 860,510 patients with 11,194,579 healthcare encounters obtained from the Children's Hospital of Philadelphia. After applying stringent quality control to remove implausible growth values and including only individuals with all recommended wellness visits by age 7 years, 27,203 (50.78 % male) patients' record was utilized for model development. The model performance was evaluated by multiple standard classifier metrics and the differences among seven models were compared using Cochran's Q test and posthoc pairwise testing. Their result shows that XGBoost yielded 0.81 (0.001) AUC, which outperformed all other models. It also achieved statistically significant better performance than all other models on standard classifier metrics (sensitivity fixed at 80 %), precision 30.90 % (0.22 %), F1-score 44.60 % (0.26%), accuracy 66.14% (0.41%), and specificity 63.27% (0.41%).

3. Research Methods

This secsion describes the various steps taken to design and implement the Artificial Neural Network based obesity prediction model. It elaborately explain the dataset used, the various data preprocessing techniques adopted as well as the method of traing, testing and model evaluations.

3.1 Dataset Description

This study explores the Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from University of California Irvine (UCI) Machine Learning Repository. The dataset has 2111 records and 17 features. The records are labeled with the class variable "NObesity" (Obesity Level) which allows classification into 7 groups: "Insufficient Weight", "Normal Weight", "Overweight Level I", "Overweight Level II", "Obesity Type I", "Obesity Type II" and "Obesity Type III".

3.2 Reading the Textual File

To early identify the occurrence of obesity and perform the necessary analysis need to train a model, it is essential to optimize a dataset. This study utilized a secondary obesity dataset obtained from the UCI website. Pandas as a third-party library of python programming were utilized to read the experimental dataset. Pandas read data as data frames, which can be visualized in a tabular cell format. Fig 1 depicts the first 5 records from the obesity dataset after reading the dataset via Pandas framework.

```
df.head()
```

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRAI
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportati
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportati
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportati
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walki
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportati

Fig 1: Obesity Dataset

3.3 Data Preprocessing Steps

The data preprocessing steps used in this paper include:

1. **Identification and handling of missing values:** All missing values were identified and properly replaced.
2. **Encoding the categorical data:** All categorical variables were encoded to enhance optimal prediction performance.
3. **Data Scaling:** Normalization is an essential aspect of the machine learning task after all necessary data purification. Thus, it is essential to scale the dataset into a feature range, enhancing the model capabilities to predict obesity at an early stage effectively. To ensure feature scaling this study utilized the Standard Scaler functionalities from the Sklearn machine learning to scale the dataset as shown in figure 4.5 below. The essence was to bound the data records to a feature range between 0 and 1 and hence empower the model with great magnitude distances control among the datasets and thus enable it to convergence faster.

3.4 Splitting the dataset into training and test proportions

For the training and testing datasets, 70% of the sample dataset size was used as training dataset *S*, and the remaining 30% sample dataset as the testing dataset *T*.

Using mathematical specification, suppose that there is a total of *n* training data points in the training dataset *S*. And for each data point denoted as d^n, there exists a vector of predictive variables denoted as $x^n = x_1^n, x_2^n, \ldots x_m^n$ and a class label denoted as y^n ($y^n = 1$ if the subject met the definition of obesity; otherwise, $y^n = 0$).

3.5 Model building

An artificial neural network (ANN) is a computational model that emulates the biological neural system to conduct comprehensive data analysis. This study utilized a Multi-layer Perceptron Neural Networks model, which maps a set of input data onto a set of appropriate output data through three layers of neurons. The three layers are the input layer, hidden layer, and output layer. Further, there are multiple connections between adjacent layers with a weight assigned to each connection. Specifically, the input layer consists of neurons corresponding to predictive variables $(x^1x^2, \ldots \ldots x^k)$ which are connected to neurons in the hidden layer. Each neuron in the hidden layer sum the data received from the input layer through weighted connections, and then modifies the sum by a non-linear transfer function before passing the sum to the output layer.

To appropriately train the ANN model, the backpropagation algorithm was utilized, accompanied by the application of the rectified linear function (Relu) as the activation function to the input layer with 128 neurons, then a dense layer as the hidden layer was given 64 neurons and lastly, the final layer which corresponds to the output layer was given 7 neurons that equates the 7 class of obesity range classification. The output layer utilized softmax as the activation function. The learning rate and the momentum were set as 0.0001 and 0.7 respectively. At the cost of model compilation, the sparse_categorical_crossentropy was optimized as the loss function due to its suitability for multi-class classification problem

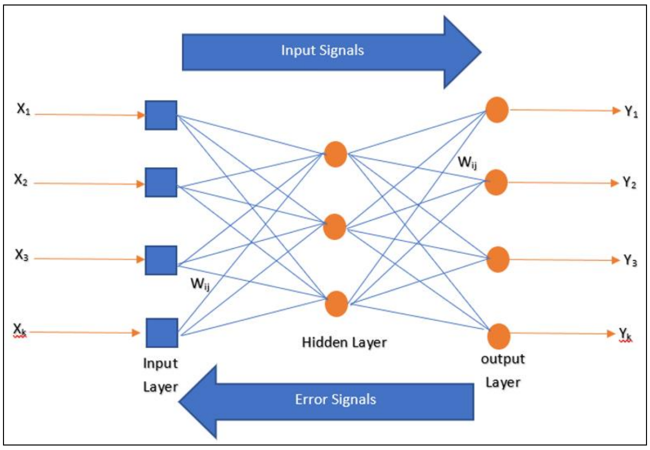

Fig 2: Atifitial Nueral Network Prediction Architecture

Algorithm 1: The Artificial Neural Network Prediction Algorithm

Step 1: Passed the input with some weight to the hidden layers $(x^1x^2, \ldots \ldots x^7)$

Step 2: Connect all the inputs to each neuron

Step 3: perform computation at the hidden layers

Step 3.1: Get the summation of all input with their weight

Step 3.2: Get bias.

Step 3.3: Get the threshold unit.

Step 4: Repeat step 3 for each of the hidden layers

Step 5: Pass the result to an output layer

Step 6: Get predictions from the output layers and hence calculate the performance metrics.

Step 7: Calculate error, i.e., the difference between the actual and predicted output.

Step 8: End.

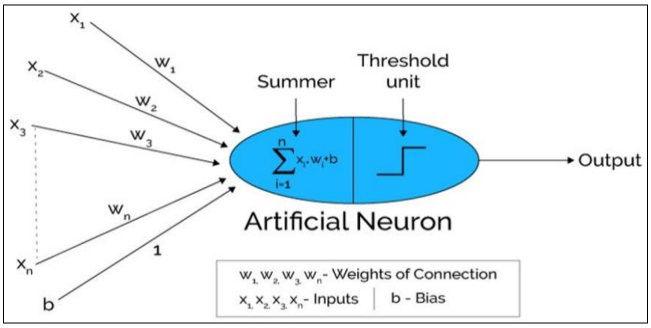

Fig 3: Algorithmic Working Model

3.5.1 Global Optimal Achievement Steps

Achieving global optimum is one task in a computational task that must be emphasized, hence to achieve global optimal, the flowing step was carried out:

Step 1: Randomly initialized weights values.

Step 2: Assign intercepts to the model while forward propagation and errors are calculated.

Step 3: Calculate the derivative of the error current weights (i.e gradient calculation).

Step 4: Calculate new weights using the below formula, where **a is the learning rate** which is the parameter also known as step size to control the speed or steps of the backpropagation.

$$W_n = W_x - a\left(\frac{d\,Error}{dW_x}\right) \quad (1)$$

Where, W_n is a new weight calculated,

W_x is the old weight,

'a' as the learning rate,

$\frac{d\,Error}{dW_x}$ as the derivative of the error concerning weight.

Note: The learning rate gives additional control on how fast the model move on the curve to reach global minima.

3.6 Performance Metrix used for Model Evaluation

The proposed system used the following evaluation parameters including precision, F-measure, recall, and accuracy score respectively.

Precision measures the percentage of the number of correctly predicted positive instances divided by the total number of predicted positive instances:

$$precision = \frac{TP}{TP+FP} \quad (2)$$

Recall measures the classifier's completeness. It is the percentage of correctly predicted positive instances to the actual number of positive instances on the dataset.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F-measure (or F-score) is defined as the harmonic mean of precision and recall, which combines recall and precision to output a single score. F-measure therefore might have the best value of 1 and the worst value of 0:

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Accuracy is one of the most important metrics of performance evaluation and is measured as a percentage of the number of correctly predicted instances to the total number of instances present in the dataset. Thus, the accuracy calculates the ratio of inputs in the test set correctly labeled by the classifier:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

4. Results and Discussion

The results obtained from the evaluation of Artificial Neural Network model for obesity prediction implemented in python programming language and other third-party python libraries such as Numpy, Matplotlib, Sklearn, Pandas, Tensorflow, and Keras to achieve the prediction model are summarized below. After the efficient Artificial Neural Network model training, the model prediction accuracy metric was 0.9669, which is equivalent to 97% when multiplied by 100 with a loss metric of 0.1419 deficiencies after 50 successful epochs with 32 batch-size per epoch. The precision, recall, f1_score for each of the seven classes, classification of obesity namely: "Insufficient Weight", "Normal Weight", "Overweight Level I", "Overweight Level II", "Obesity Type I", "Obesity Type II" and "Obesity Type III" as shown in Table 1 below.

Table 1: The multiclass classification results

	Precision	Recall	F1-score	Accuracy
Insufficient Weight	92%	100%	96%	97%
Normal Weight	95%	89%	96%	97%
Overweight Level I	99%	99%	99%	97%
Overweight Level II	98%	98%	98%	97%
Obesity Type I	100%	100%	100%	97%
Obesity Type II	95%	93%	94%	97%
Obesity Type III	98%	98%	98%	97%

5. Conclusion

This study implemented Artificial Neural Network with multi-layer perceptron as a suitable methodology for the detection of obesity patterns to predict its possible occurrence. The dataset utilized was obtained from the UCI website and the independent variable for classifying an individual record as having obesity or not, emanates from the seven multi-class prediction outputs. Artificial Neural Network with multi-layer perceptron was adopted due to its suitability for multi-class classification. The ANN model was implementation using python programming language with Jupiter NoteBook and other third-party python libraries such as Numpy, Matplotlib, Pandas, Tensorflow, Keras, and Sklearn. The model upon evaluation recorded an outstanding performace as it achieves high prediction accuracy of 97%.

6. References

1. Casimiro ACM, Paul F, Abir H, Dhiya A, Basma A, Jade H. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. International Joint Conference on Neural Networks (IJCNN), 2017, 2743-2750. Doi: 10.1109/IJCNN.2017.7966194.
2. Chiolero A. Why causality, and not prediction, should guide obesity prevention policy. The Lancet Public Health. 2018; 3(10):461-462. ISSN 2468-2667.
3. Colquitt JL, Pickett K, Loveman E, Frampton GK. Surgery for weight loss in adults. The Cochrane Database of Systematic Reviews (Meta-analysis, Review). 2014; 8(8):CD003641. Doi: 10.1002/14651858.CD003641.pub4. PMID 25105982.
4. Dugan TM, Mukhopadhyay S, Carroll A. Machine Learning Techniques for Prediction of Early Childhood Obesity. Applied Clinical Informatics. 2015; 6:506-520. Doi: http://dx.doi.org/10.4338/ACI-2015-03-RA-0036.
5. Ellen P, Williams MM, Karen W, Patricia MD, Sharon BW. Overweight and Obesity: Prevalence, Consequences, and Causes of a Growing Public Health Problem, 2015.
6. Gupta M, Phan TT, Bunnell T. Obesity Prediction with EHR Data: A Deep Learning Approach with Interpretable Elements, 2019. http://arxiv.org/abs/1912.02655.
7. Hammond R, Athanasiadou R, Curado S. Predicting Childhood Obesity Using Electronic Health Records and Publicly Available Data, Published Online First, 2019. Doi: https://doi.org/10.1371/journal.pone.0215571.
8. Lee I, Bang KS, Moon H. Risk factors for obesity among children aged 24 to 80 months in Korea: A decision tree analysis. Journal of Pediatric Nursing. 2019; 46:15-23. Doi: https://doi.org/10.1016/j.pedn.2019.02.004
9. Richa R, Snekhalatha U, Palani TK. Thermal imaging method to evaluate childhood obesity based on machine learning techniques, 2021. Doi: 10.1002/ima.22572
10. Sadaf I, Zuneera A, Aisha N, Mirza TB, Samina S, Ambreen H, *et al*. Overweight and obesity prevalence and predictors in people living in Karachi, J. Pharmaceut. Res. Int., 2021, 194-202.
11. Salvador C, Andreas R. Is the calorie concept a real solution to the obesity epidemic? Glob. Health Action. 2017; 10(1).
12. Simmonds M, Llewellyn A, Owen CG, Woolacott N. Predicting adult obesity from childhood obesity: A systematic review and meta-analysis, Obes. Rev. 2020; 17(2):95-107.
13. Sri AT, Dian SA, Hedi k, Armin L, Surdirman N. Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018, 2021. PMCID: PMC8255629. Doi: 10.3389/fnut.2021.669155
14. Syahrul SS, Abdul HAG, Mohd S, Mohamed SC, Norella TK. Steroid-induced diabetes mellitus in systemi c lupus erythematosus patients: Analysis from a

Malaysian multi-ethnic lupus cohort, Int. J. Rheum. Dis. 2015; 18(5):541-547.
15. World Health Organization. Obesity and Overweight fact Sheet, 2015. Retrieved from: https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight on 2, January, 2022.
16. Xueqin P, Christopher BF, Felice LS, Aaron JM. Prediction of early childhood obesity with machine learning and electronic health record data. International Journal of Medical Informatics. 2021; 150:104454. https://www.journals.elsevier.com/international-journal-of-medical-informatics
17. Zhang S, Tjortjis C, Zeng X, Qiao H, Buchan I, Keane J. Comparing data mining methods with logistic regression in childhood obesity prediction. Information Systems Frontiers. 2009; 11:449-460.