Deep Neural Network Training with Synthetic Data for BMI Classification

Velázquez Dodge Charles Fernando [**]
Xochimanca Tapia Eduardo [*] **Ayala Coapango Ana Luisa** [*]
Torres Gómez Georgina [*] **Leyva Varela Paola** [*]

[*] *Universidad Stratford de Cuautla, Av Insurgentes 250, Emiliano Zapata, 62744 Cuautla, Mor.*
[**] *Instituto Tecnológico de Cuautla, Libramiento Cuautla-Oaxaca S/N Juan Morales, 62745 Cuautla, Mor. (e-mail: charlesdodge@ieee.org)*

Abstract:
This research explores the use of synthetic data to train a deep neural network for Body Mass Index (BMI) categorization. After training, the model undergoes evaluation using data sourced from actual individuals. The classification based on BMI changes according to an individual's age, and World Health Organization provides a percentile table for the classification of youths, whereas adults and the elderly are categorized using other classifications. This model is taught to classify individuals into underweight, healthy, overweight, and obese. The concept of using synthetic data arises from the challenges associated with obtaining authentic information, particularly data pertaining to children, since parental consent is required according to Health Insurance Portability and Accountability Act (HIPAA) U.S. Department of Health and Human Services, Office for Civil Rights (2003) and *Ley Federal de Protección de Datos Personales en Posesión de los Particulares* (LFPDPPP) Ley Federal de México (2010). A further drawback is the inflexibility of the obtained data, and that's why we suggested this strategy to train a neural network, since synthetic data is more available, can be balanced, and may be created for particular classifications according to the needs. After completing the neural network training, we proceed to evaluate the model using real-world data. As a result, the model produces an acceptable classification, which could be readily improved by creating more synthetic data.

Keywords: Synthetic Data, Deep Learning, BMI Classification, Obesity, Health Prediction

1. INTRODUCTION

The Body Mass Index (BMI) is a widely used statistic for determining whether a person is healthy, obese, underweight, or overweight. It is vital to monitor patients in hospitals and avoid various health complications. It is typical in the literature to find the usage of large datasets based on real-world patients Yi et al. (2024); nevertheless, the dataset information may not always suit the particular study goals. Another problem is the privacy of patients, particularly minors, whose data requires parental consent according to U.S. Department of Health and Human Services, Office for Civil Rights (2003) and Ley Federal de México (2010). Synthetic data is a beneficial alternative since it is constantly available, can be focused on the information required, is simple to collect, and is less expensive Smith et al. (2022) Klein and Michels (2024). In this context, we are employing metrics based on Mexican standards; therefore, the data created will be provided by the IMSS online calculator, which is based on World Health Organization percentile data Rodríguez et al. (2011).

Several sectors, including healthcare, have effectively used deep learning. To get effective results in neural network training, the dataset is critical, and synthetic data will meet the data requirements. However, we test the trained model's performance using a real-world situation using data from a group of Mexicans.

The purpose of this project is to investigate the use of synthetic data for training deep neural networks to categorize BMI categories in adults, as well as BMI percentiles for newborns, children, and adolescents.

The neural network uses three input neurons for sex, age, and BMI. The following layer includes eight neurons, the next ten neurons, and finally another layer of eight neurons and four output neurons. The layers employ the RELU activation function, with the exception of the output, which uses the sigmoid function. The learning rate is dynamic and determined by the derivative of the error.

2. STATE OF THE ART

The next works presented help to prove the benefits of the use of synthetic data and show related works with obesity and BMI that use different algorithms.

2.1 The use of synthetic data in biology and medicine

The work of Smith et al. (2022) proves that the synthetic data can be used to share sensitive health information without compromising privacy and that its use accelerates access to clinical data without the lengthy ethical approval processes.

Other studies that use the synthetic data published by Klein and Michels (2024) recognize the high cost of real data and the challenges of generating synthetic data. Proposes the synthetic generation to train a machine learning model that optimizes the process to detect disease in tomato plants with lower cost. They conclude that the use of synthetic data lets them generate a “good” training data with lower cost and time.

On the other side, we discovered research that utilizes synthetic data but is unrelated to obesity or BMI Almeida et al. (2022). This research highlights the importance of using synthetic data in therapeutic settings since it is more accessible, less expensive, and does not need personal information. The research used synthetic data to substitute for a genuine diabetes patient, which was then analyzed by two certified experts in the field.

2.2 Algorithms used on BMI

For this research, we identified a thorough paper review authored by Yi et al. (2024), which included 40 publications about the usage of deep learning and obesity. None of the 40 publications included in the study discuss the use of synthetic data; the majority are datasets from actual patients, but other types of datasets include genetic and phenotypic databases, pictures, mobile app data, and electronic health records. Traditional feedforward networks, CNNs, LSTMs, and hybrid models have all been utilized.

Other obesity-related papers, such as López Díaz et al. Díaz et al. (2023), were discovered by other methods. They utilized KNN and random trees in teens with an accuracy of 80-87%.

Other publications on baby obesity give a categorization utilizing fuzzy logic with ID3Suca et al. (2016); the findings of this piece were excellent, with an accuracy of 83.65% in males and 76.13% in women. Other similar study used fuzzy rules in Mexican childrenAlva et al. (2020); they obtained a prediction of 93% to 97% in children aged 10 to 14 years old.

2.3 Conclusions

The current state of the art shows the advantages of utilizing synthetic data, including cost reduction, time efficiency, and the preservation of patient privacy. There are many algorithms used for BMI and obesity; however, those utilizing synthetic data are limited, particularly in application to a younger population.

3. METHODOLOGY

This section describes the method used to generate our data to train the model and the architecture of the neural network.

3.1 Generation of data

To generate the synthetic data, high-level students from the areas of nursing and nutrition generated, based on their experience, a list of age, gender, weight, and height. Using the IMSS’s online calculator (*https://imss.gob.mx/salud-en-linea/calculaimc*), they created a balanced list with the same number of men and women, varying ages, and the same number of underweight, overweight, obese, and healthy classes. Finally, they validated the consistency.

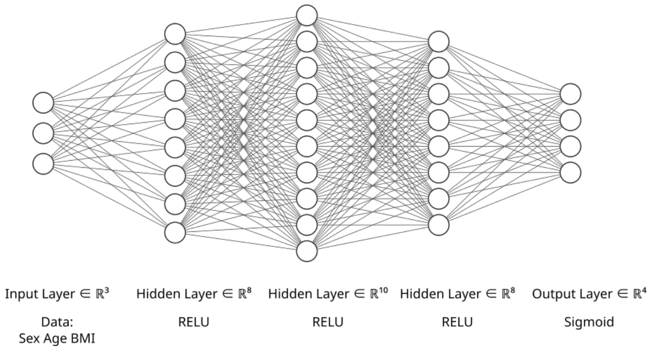

Fig. 1. Neural network architecture

The final dataset obtained has 100 registers, with 24 representing underweight, 27 for healthy, 27 for overweight, and 22 for obese. The data were also normalized, with 100 representing the maximum age and 40 representing the maximum BMI.

3.2 Neural Network Architecture

The neural network receives input from three neurons: BMI, age, and sex. The output is four neurons for classification: underweight, healthy, overweight, and obese.

The first hidden layer contains eight neurons with the RELU activation function, the second has ten neurons with the RELU activation function, and the final hidden layer has eight neurons with the RELU activation function as well. The output layer employs the sigmoid activation function. All layers are fully connected, and similar models were utilized for deep learning in BMI applications Varela-Tapia et al. (2022). The RELU function is employed to avoid the vanishing gradient Hochreiter (1991).

The implemented deep neural network consists of an input layer with n neurons corresponding to the input features and three hidden layers with ReLU activation:

$$h_j = \max(0, \sum_i w_{ij}x_i + \theta_j) \tag{1}$$

An output layer with sigmoid activation:

$$y_k = \frac{1}{1 + e^{-\sum_j w_{jk}h_j - \theta_k}} \tag{2}$$

3.3 Training

Cost function: Mean Squared Error (MSE) was used:

$$E = \frac{1}{n}\sum_{p=1}^{n}\sum_{k=1}^{m}(t_{pk} - y_{pk})^2 \tag{3}$$

Weight update using backpropagation:

$$w_{ij} = w_{ij} + \eta \cdot \delta_j \cdot x_i + \alpha \cdot \Delta w_{ij} \tag{4}$$

where

$$\delta_j = \frac{\partial E}{\partial h_j} \tag{5}$$

During training, we discovered a difficulty reducing error. So, we modified the learning rate equation to make the value dynamic:

$$\eta = \frac{\eta_{\max}}{1 + \lambda \left(\frac{dE}{dt}\right)^2} \tag{6}$$

The Table 1 shows the parameters and architecture information:

Parameter	Value
Epochs	50,000
Batch size	100
Input neurons	3
Hidden layer 1	8
Hidden layer 2	10
Hidden layer 3	8
Output neurons	4
Bias initialization	0.08
Initial learning rate	1×10^{-5}
Minimum learning rate	1×10^{-9}
Maximum learning rate	1×10^{-4}
Learning rate decay	0.9
Momentum (α)	0.1
Momentum decay	0.99
Tolerance	1×10^{-4}
Activation function (hidden layers)	ReLU
Activation function (output layer)	Sigmoid

Table 1. Hyperparameters and architecture of neural network.

The learning rate changes every epoch acording to the equation (6)

4. RESULTS

To validate the model's performance, real data from young individuals and children was used to assess the network's classification accuracy. The data set of the Table 4 was obtained by nutrition students using the online aplication given by IMSS.

The performance of the neural network was assessed using 48 test patterns. The predicted and actual outputs of each pattern were compared. A pattern was judged properly categorized if all of the predicted outputs matched the actual results.

Out of the 48 test patterns, 35 were properly categorized and 13 were incorrectly classified. This yields an overall categorization accuracy of around 72.9%. Table 2 displays the final classification results in a confusion matrix where the main diagonal shows the correct classifications, while Table 3 displays the model's performance with different metrics.

	A	B	C	D
A	2	1	0	0
B	2	19	7	0
C	0	0	7	3
D	0	0	0	7

Table 2. Confusion Matrix of the Neural Network. Classes: A=Underweight, B=Healthy, C=Overweight and D=Obesity.

Class	Precision	Recall	F-Measure
Underweight	50.00	66.67	57.14
Healthy	90.48	73.08	80.85
Overweight	70.00	100.00	82.35
Obesity	100.00	100.00	100.00

Table 3. Metrics of the Neural Network's Performance

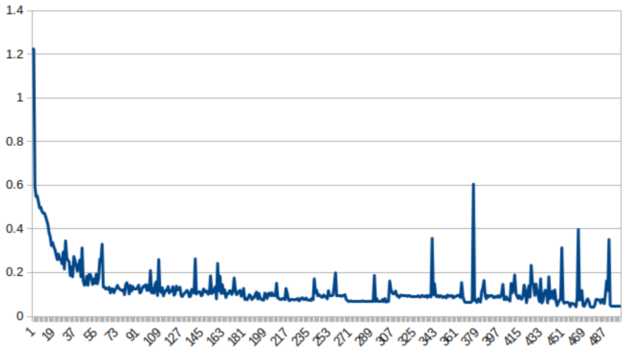

Fig. 2. Mean Squared Error (MSE)

5. CONTRIBUTIONS

The contributions of this study are the following:

- The use of the IMSS government calculator to generate synthetic to train a deep artificial neural network that is faster and easier than conventional methods while protecting patient privacy. This is an improvement to BMI data acquisition and makes the data more adaptable to the needs.
- The use of deep learning for BMI prediction in a mexican young population is a novel approach. Given the difficulty of the bureaucratic process, parental permission is required, as well as consideration of local law depending on the country.
- A dataset of teenagers groups from Mexico obtained by nutrition students.

In this study, we demonstrate the acceptable performance of a model trained with synthetic data utilizing IMSS standards and an online calculator. This technique is a unique approach to training a deep neural network for BMI classification in mexican adolescents. The protection of minors' biometric and personal information, addressing ethical concerns.

6. DISCUSSION

The accuracy of 72.9% is a good starting point to investigate the use of synthetic data. We consider that an acceptable starting point, and it may be improved in future works. The results for underweight are lower than the others, so more data is needed to improve the significance and quality of the neural network. We attempted to generate diverse data in age, gender, and BMI classification; however, it is necessary to increase the infant data because it is nonlinear when compared to adult data classification. That's because the children's classification is based on statistical tables. Other changes that could improve the results include expanding the classification with more variables, such as weight and height. Perhaps it is redundant, but there may be a pattern in the additional information that leads to improved classification performance, especially on newborns that are classified with other metrics like the head size and the proportion of their body Rodríguez et al. (2011).

The limitation of this study was the volume of data. Generating synthetic data is easier than getting the data from real people, but it requires validation from experienced people in the area. We aimed to gather high-quality data to train, however the data set is too small to train the model. To improve the results, a larger dataset is needed, as well as modifications to the model architecture, such as incorporating dropout and experimenting with other training algorithms, like Adam.

7. CONCLUSIONS

Using synthetic data for BMI is a simple and successful method that eliminates the need for patients to provide their information, especially for children. Our performance is acceptable, but we believe it might be improved by using additional synthetic data for training. The train performed well (95% accuracy), but had some trouble identifying infants correctly. The IMSS categorization use height and body proportions rather than BMI, as outlined in the IMSS guideline Rodríguez et al. (2011). This ambiguity in the model leads to negative consequences for babies and early children.

At the start of the experiments, we used data generated by ChatGPT combined with the IMSS calculator, but we discovered discrepancies between the values calculated with the IMSS online application. The training performance was around 70%, so we decided to delete the ChatGPT data and used the online calculator created by the government health authority.

REFERENCES

Almeida, A.R., Déniz, A., Fabelo, H., Ortega, S., Soguer, C., Wägner, A., and Callicó, G.M. (2022). Validación clínica y análisis de datos sintéticos de mujeres con diabetes generados con técnicas de inteligencia artificial: prueba de concepto. *Endocrinol Diabetes Nutr.*, 69, Espec Cong 1, 210.

Alva, R., Laria, J., Ibarra, S., Castán, J., and Teran, J. (2020). Propuesta de un modelo difuso para determinar sobrepeso y obesidad en niños y adolescentes. *Revista Chilena de Nutrición*. SCIELO.

Díaz, S.L., Sibaja, J.P., Martínez, A.F., and Vázquez, S.J. (2023). Algoritmos de clasificación para la detección de obesidad en adolescentes: Un estudio comparativo entre knn y árboles de decisión. *Revista de Investigación en Tecnologías de la Información.*, 11, 23.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen*. Ph.D. thesis, Technische Universität München.

Klein, Waller, P.P.T. and Michels (2024). Synthetic data at scale: a development model to efficiently leverage machine learning in agriculture. *PubMed*. URL `https://pubmed.ncbi.nlm.nih.gov/39351023/`.

Ley Federal de México (2010). Ley federal de protección de datos personales en posesión de los particulares. Accessed: 2025-05-13.

Rodríguez, M.R., Ruano, M.D., Blanno, A.G., Hernández, C.Q., Gómez, V.G., and Márquez, R.H. (2011). Abordaje diagnóstico y seguimiento del paciente pediátrico con talla baja. Catálogo Maestro de Guías de Práctica Clínica. IMSS-510-11.

Smith, A., Lambert, P.C., and Rutherford, M.J. (2022). Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Medical Research Methodology*, 22(1), 176. doi:10.1186/s12874-022-01654-1. URL `https://doi.org/10.1186/s12874-022-01654-1`.

Suca, C., Córdova, A., Condori, A., and Cayra, J. (2016). Modelo difuso para la predicción de casos de obesidad empleando el árbol gfid3 generalizado. In *Research in Computing Science*.

Id	Sex	Age	Height	Weight	BMI	Class
1	M	15	1.76	72.6	23.4375	overweight
2	M	15	1.9	84.1	23.2964	overweight
3	M	15	1.62	55.8	21.2620	healthy
4	M	14	1.66	65.4	23.7335	overweight
5	F	16	1.59	57.8	22.8630	healthy
6	M	15	1.74	61.5	20.3131	healthy
7	F	15	1.61	60	23.1473	healthy
8	F	15	1.6	49.5	19.3360	healthy
9	F	15	1.51	80.7	35.3932	obesity
10	M	17	1.9	87.5	24.2382	healthy
11	M	17	1.74	70.4	23.2527	healthy
12	M	17	1.8	90.7	27.9938	overweight
13	M	17	1.7	96.5	33.3910	obesity
14	M	17	1.7	102.9	35.6055	obesity
15	F	17	1.53	56.9	24.3069	healthy
16	F	17	1.68	68	24.0930	healthy
17	F	17	1.6	49.4	19.2969	healthy
18	F	17	1.61	90.08	34.7517	obesity
19	F	17	1.53	44.2	18.8816	healthy
20	F	17	1.65	65.9	24.2057	healthy
21	M	17	1.8	81	25	overweight
22	M	16	1.7	80	27.6817	overweight
23	M	16	1.69	53.8	18.8369	healthy
24	F	16	1.57	47.2	19.1488	healthy
25	M	16	1.71	84.7	28.9662	obesity
26	F	16	1.49	40.3	18.1523	healthy
27	F	16	1.63	64.9	24.4270	overweight
28	F	14	1.55	56.7	23.6004	overweight
29	F	14	1.59	62.3	24.6430	overweight
30	M	14	1.57	51.1	20.7311	healthy
31	F	14	1.53	52.8	22.5554	healthy
32	F	14	1.5	43.1	19.1556	healthy
33	M	14	1.45	37.4	17.7883	healthy
34	M	14	1.72	81.9	27.6839	overweight
35	M	14	1.74	63.4	20.9407	healthy
36	M	14	1.64	70	26.0262	obesity
37	F	14	1.44	40.5	19.5313	healthy
38	F	12	1.6	44.2	17.2656	healthy
39	F	12	1.53	45.7	19.5224	healthy
40	M	12	1.44	42.2	20.3511	overweight
41	M	12	1.48	37.2	16.9832	healthy
42	M	12	1.61	46.9	18.0934	healthy
43	F	12	1.53	61.9	26.4428	obesity
44	F	12	1.43	37.9	18.5339	healthy
45	M	13	1.48	34.65	15.8190	healthy
46	M	13	1.71	61	20.8611	overweight
47	F	13	1.45	36.2	17.2176	healthy
48	M	13	1.63	47.4	17.8403	healthy

Table 4. Real-world data used to validate.

U.S. Department of Health and Human Services, Office for Civil Rights (2003). Ocr privacy brief: Summary of the hipaa privacy rule. Accessed: 2025-05-13.

Varela-Tapia, E., Acosta-Guzmán, I., Acosta-Varela, C., Marcillo-Sanchez, P., Pérez, D., and Bravo, J. (2022). Intelligent predictive model of bmi in nutritionists' patients using machine learning algorithms: Logistic regression and neural networks. In *Proceedings of the 20th LACCEI International Multi-Conference for Engineering, Education and Technology: "Education, Research and Leadership in Post-pandemic Engineering: Resilient, Inclusive and Sustainable Actions"*. URL `https://doi.org/10.18687/laccei2022.1.1.791`.

Yi, X., He, Y., Gao, S., and Li, M. (2024). A review of the application of deep learning in obesity: From early prediction aid to advanced management assistance. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 18, 103000.