DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework

Ji-Hoon Jeong, *Associate Member, IEEE*, In-Gyu Lee, Sung-Kyung Kim, Tae-Eui Kam, Seong-Whan Lee, *Fellow, IEEE*, and Euijong Lee

Abstract—The global prevalence of childhood and adolescent obesity is a major concern due to its association with chronic diseases and long-term health risks. Artificial intelligence technology has been identified as a potential solution to accurately predict obesity rates and provide personalized feedback to adolescents. This study highlights the importance of early identification and prevention of obesity-related health issues. To develop effective algorithms for the prediction of obesity rates and provide personalized feedback, factors such as height, weight, waist circumference, calorie intake, physical activity levels, and other relevant health information must be taken into account. Therefore, by collecting health datasets from 321 adolescents who participated in Would You Do It! application, we proposed an adolescent obesity prediction system that provides personalized predictions and assists individuals in making informed health decisions. Our proposed deep learning framework, DeepHealthNet, effectively trains the model using data augmentation techniques, even when daily health data are limited, resulting in improved prediction accuracy (acc: 0.8842). Additionally, the study revealed variations in the prediction of the obesity rate between boys (acc: 0.9320) and girls (acc: 0.9163), allowing the identification of disparities and the determination of the optimal time to provide feedback. Statistical analysis revealed that the performance of the proposed deep learning framework was more statistically significant ($p<0.001$) compared to the other general models. The proposed system has the potential to effectively address childhood and adolescent obesity.

Ji-Hoon Jeong, In-Gyu Lee, and Euijong Lee are with the Department of Computer Science, Chungbuk National University, Chungbuk 28644, Republic of Korea (e-mail: jh.jeong@chungbuk.ac.kr; ingyu.lee@chungbuk.ac.kr; kongjjagae@cbnu.ac.kr).

Sung-Kyung Kim is with the Injewelme Corporation, Seoul 06764, Republic of Korea (e-mail: skkim@injewelme.com).

Tae-Eui Kam and Seong-Whan Lee are with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: kamte@korea.ac.kr; sw.lee@korea.ac.kr).

Index Terms—Childhood obesity prediction, digital healthcare, daily health informatics, artificial intelligence, deep learning.

I. INTRODUCTION

CHILDHOOD and adolescent obesity rates have become a growing concern worldwide in recent years. According to the World Health Organization (WHO), the number of overweight children and adolescents aged 5–19 years has increased from 32 million in 1990 to 42 million in 2013, globally [1], [2], [3], [4], [5], as shown in Table I [6]. Obesity in childhood and adolescence is associated with various health problems, such as cardiovascular disease, type 2 diabetes, and musculoskeletal disorders, and can lead to a higher risk of obesity and related health problems in adulthood [7], [8], [9].

Predicting obesity rates is crucial because early identification of individuals at risk can help prevent and manage obesity-related health problems [10], [11], [12]. Artificial intelligence (AI) technology has been widely used recently to implement digital healthcare using various approaches [13], such as implementing smart homes using the Internet of Things [14], [15], building multimodality interfaces for real-world applications [16], [17], [18], and providing feedback to users using biomedical sources [19], [20], [21]. Furthermore, AI technology is potentially beneficial to this field through its accurate and personalized predictions of obesity rates for adolescents. AI algorithms can be trained in large health information datasets and provide tailored feedback to individuals, allowing them to make informed health decisions [22], [23], [24], [25], [26].

The rising prevalence of childhood and adolescent obesity has become a significant public health concern. Recently, there has been an increasing focus on developing accurate and effective approaches to predict obesity rates in children and adolescents [27], [28]. A key factor in this effort is the collection and analysis of relevant health data [29]. Various health-related data, such as height, weight, waist circumference, calorie intake, physical activity levels, and other relevant health information, must be collected to accurately predict obesity rates in children and adolescents [30], [31], [32], [33], [34]. These data are essential to develop accurate algorithms to predict obesity rates and provide personalized feedback to individuals. The availability of such datasets can enable researchers and healthcare providers to develop more effective interventions and prevention strategies for childhood and adolescent obesity [11], [35], [36]. However,

TABLE I
GLOBAL OBESITY TRENDS FOR CHILDREN, ADOLESCENTS, AND ADULTS BY GENDER OVER THE PERIOD 2020–2035 [WORLD OBESITY FEDERATION]

	Children and Adolescents		Adults	
Year	Boy	Girl	Men	Women
2020	103 (10%)	72 (8%)	347 (14%)	466 (18%)
2025	140 (14%)	101 (10%)	439 (16%)	568 (21%)
2030	175 (17%)	135 (14%)	553 (19%)	693 (24%)
2035	208 (**20%**)	175 (**18%**)	690 (**23%**)	842 (**27%**)

*Notation: Number of obese individuals (million) and proportion of all genders
*Age: children, adolescents (5–19 years), adults (20 years and older)
*For children and adolescents, obesity is defined using the WHO classification of +2SD above median growth reference.

there are various challenges in collecting and analyzing these data, such as ensuring data quality, protecting privacy, and addressing ethical considerations [29], [37], [38].

Several studies have examined the use of AI technology to predict obesity rates among children and adolescents. Gupta et al. [24] trained a general long-term memory network (LSTM) using static and dynamic electronic health record data over a period of 1–3 years to predict obesity for individuals between 3–20 years. On average, they achieved an area-under-the-curve score of 0.88. Mondal et al. [35] employed a machine learning (ML) classifier to categorize individuals into three groups based on childhood health maintenance data: normal weight, overweight, and obese. The experimental results demonstrated the classification accuracies of 89%, 77%, and 89% for the three respective scenarios. In addition, Cheng et al. [39] used the Obesity Prediction in Early Life (OPEL) database as their dataset. After pre-processing the data, they divided the children who had clinical visits 2, 3, 5, and 8 times between ages 0 and 4 into male and female groups. They trained an LSTM model and obtained a mean absolute error of 0.98 and a lasso regression value of 0.72.

The primary objective of this study was to investigate the application of AI technology in predicting obesity rates in adolescents. The study evaluated the abilities of AI algorithms to predict obesity rates and provide personalized feedback to individuals. The overarching objective is to mitigate obesity among adolescents by introducing a model with demonstrated efficacy in predicting adolescent obesity. In addition, the study explored the importance of data collection, a crucial aspect in predicting obesity rates in adolescents. The types of data required for accurate predictions, the challenges involved in collecting and analyzing such data, and the potential advantages of data-driven approaches to combat childhood and adolescent obesity were examined.

The study also proposes a prediction system that uses an AI model to predict the likelihood of obesity in adolescents proactively and offers customized feedback based on these predictions. We adapt deep learning technology because previous research has not used it to predict obesity rates in adolescents. The proposed model, DeepHealthNet, outperformed other comparable machine learning-based models in accurately predicting obesity rates. DeepHealthNet effectively employs data augmentation techniques to train deep learning models, even when available daily health data are limited, resulting in improved predictive performance. Furthermore, the study investigated variations in the prediction of the obesity rate between boys and girls, to reveal any disparities and determine the optimal time to provide feedback. The implementation of the proposed system has significant potential to address the issue of childhood and adolescent obesity.

II. MATERIALS AND METHODS

A. Participants

Initially, 321 participants (aged 10–12 years, 133 males and 188 females), who were students of the same elementary school in Seoul, participated in the use of the Would You Do It! (WUDI!) mobile application [40]. Of these 321 participants, 187 (75 males and 112 females) officially underwent body measurements by the Korea Sports Promotion Foundation (KSPO) for the experiment, immediately before and after the experimental period. The overall experimental protocols and environments were reviewed and approved by the Institutional Review Board of Chungbuk National University (CBNU-202308-HR-0196).

The participants were healthy, with no neurophysiological abnormalities, musculoskeletal disorders, or growth hormone deficiencies. Before the experiments, they were briefed on how to use the mobile application and the smartwatch and sync the data using an animated tutorial. Parents and school personnel, including principal and class teachers, were informed of the experimental protocols, paradigms, and purpose. After ensuring that the parents and guardians of the participating students had understood the information, their written consent was obtained according to the Personal Information Protection Act of Korea, and their signature was obtained on a form that specified their consent to the anonymous public release of data. The physical and mental states of the participating students were evaluated to compare the effect of the application used on individual states. Furthermore, each participant was required to maintain normal daily routines and be in normal health during the experiments.

After submitting the signed consent form, each participant received a Samsung Galaxy Fit 2 smartwatch, prepared for the experiment to acquire detailed data on their activity and sleep for better accuracy and further analysis. As the application developer, Injewelme Co., Ltd., has officially partnered with Samsung Health, the data collected by the Samsung device were automatically synchronized from Samsung Health to the provided application WUDI! for this experiment, via application integration (Fig. 1).

Furthermore, prior to the experiment, each participant and their parents had to log in to WUDI!. First, parents had to agree to the service terms and enter basic information about their child, including height and weight. As consent for the use of the service is required from parents of minors under 14 years of age according to Korean law, participants could obtain the invitation code to activate their use of the WUDI! application only after the parents had completed the registration correctly. When the students installed WUDI!, they had to input the MMS invitation code and then watch an animated tutorial

Fig. 1.   Data configuration for acquiring daily health data, such as height, weight, body mass index (BMI), step count, sleep time, kcal intake, exercise, and burned kcal. Data were collected by physical examination, a smartwatch (Samsung Galaxy Fit 2), and the WUDI! application (Injewelme Co. Ltd.) from 321 participants.

that showed how to properly use the mobile application and sync the smartwatch.

B. Experimental Protocols for Data Acquisition

The body measurements of the participants were obtained twice. For this experiment, KSPO supported by providing measuring tools and two staff members in charge of the Songpa Fitness Certification Center to the participating elementary school immediately at the 3rd week of July 2021 and the 5th week of September 2021. Offline measurements consisted of three parameters: 1) height, 2) weight, and 3) waist size, to confirm and compare the body changes of the individuals objectively. A simple survey was administered to students asking what kind of gift they would like to receive as a surprise. The purpose of this survey was to prepare for a subsequent intervention to increase participation and attract more attention to WUDI!.

WUDI! is a mobile application that allows student participants to choose a character as an avatar, customize the avatar with items purchasable with game coins, and complete missions and mini-games to gain points to level up the avatar and coins to purchase game items and lottery tickets for the monthly draw. The missions consisted of three parts: "Play Well", "Sleep Well" and "Eat Well". Play Well tracked the activity of users, either by acquiring data from the smartwatch when they wore it or obtaining GPS data, step count, and moving minutes from the on-board sensors of the mobile phone when they did not wear the smartwatch. Additionally, the smart wearable device automatically recognizes exercise (swimming, football, jumping rope, running, etc.), measures burned calories, and takes photos of food on their smartphone to obtain calorie information from the food they ate. Sleep Well tracked the sleep of users, either acquiring data from the smartwatch when they wore it during sleep or obtaining the smartwatch usage information from the mobile phone with the log record module to know when the usage ceased at night and reactivated in the morning. Eat Well tracked the nutrition intake, asking users to take a picture of the food when they had a meal, and then analyzing the picture to understand the content and caloric information of the meal through Vision AI analysis.

Each mission was assessed daily, and there were rewards on mission completion: points to level up and coins to purchase items. Mini-games were designed for user motivation, allowing them to play a limited number of rounds per day as a reward for daily mission completion in terms of nutrition (Eat Well) and activity (Play Well). Additionally, there were one-on-one competitions and a guild system, intended to motivate the application use by developing a sense of competition and bonding with each other. A contender can randomly choose another user to participate in the one-on-one competition to see who burned more calories within a given period. A guild could be formed with classmates, and there was a ranking board for the guild and individual users so that users could compete for individual ranks and guild ranks. In an approximately three-minute tutorial, participants completed the installation and registration of the two mobile applications, WUDI! and Samsung Health, and learned the proper usage of WUDI!.

To parents, WUDI! provided the visualized statistics and the predicted body change of their adolescent. In addition, there was a weekly report summarizing the analysis of data on activity, nutrition, and sleep of the previous week. Parents could participate by updating their height and weight weekly, which was used to track body changes during the experiment and adjust the trend of change along with the official body measurement before and after the experiment. There were two additional functions to obtain more information from parents: expert columns on health, skin, food and law, and an online commerce platform with special prices offered by Samsung Electronics.

The experimental period was about 6 months: a month during summer vacation and the remaining 5 months during the semester. The experiment was initially designed to compare the difference in overall measurements between two periods. After the initial body measurements, 321 participants started using WUDI! on their own mobile phones from the 4th week of July 2022. During the experiment, participants were free to use WUDI! voluntarily. However, there were two interventions from Injewelme Co., Ltd.. First, there were two monthly draws to award gifts based on the results of the survey of the participants. Second, events were randomly scheduled during the experimental period to double the rewards awarded, such as points
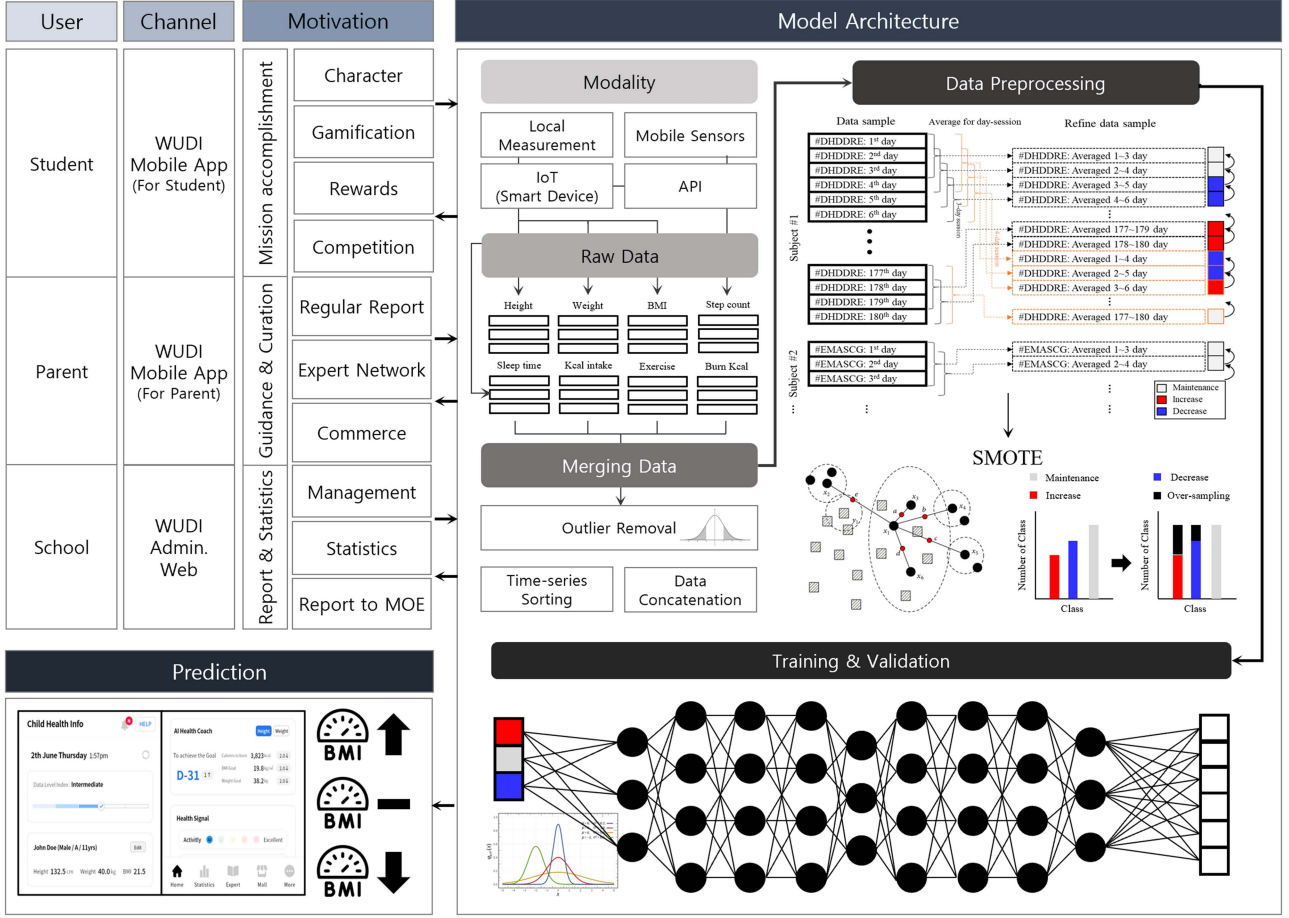
Fig. 2. Overall system framework for obesity prediction for adolescents. DeepHealthNet was proposed for training and evaluating the prediction. It comprises the outlier removal step, data preprocessing step for transforming health data, application of the SMOTE step to augment the data, training and validation step, and prediction step.

and coins, to maintain the high usage rate of WUDI! during the experiment period. Participants were free to undergo body measurements after the experiment. Consequently, 187 participants completed both body measurements, but there were some participants who continued to use WUDI! without undergoing the second body measurement.

C. Data Preprocessing

Preprocessing the raw data is essential to ensure that the data are cleaned, formatted, and transformed into a suitable format that can be effectively used for ML models (Fig. 2). The characteristics used were height, weight, step count, burned calories, calorie intake, and total sleep time. In cases where height and weight were recorded multiple times in a day, the last-day record was used. Outliers were removed if their values showed a significant deviation from the offline measurements. In cases where data was missing, the mean imputation technique was applied within each interval, utilizing the actual values adjacent to the empty data points. The data were then preprocessed into a time-series data format with multiple features, based on individuals and dates. However, this resulted in a significant number of missing values. All features were not recorded on

the same date; thus, daily records were crucial to preserve the characteristics of time-series data. Therefore, instead of deleting rows with missing values, we chose to supplement missing values with the average of the nearest one or two values (Fig. 2).

To perform labeling, we performed a preprocessing step that reduced n days' worth of data for an individual to a single row by using the average value of n days. This step involved two elements: how many days' worth of data being reduced and the length of the gap between the start dates of each bundle. For example, if we reduce 10 days of data by averaging every two days with two gaps, the data will be condensed into five data points. However, we only selected a gap length that allowed us to reduce all of the data up to the last day. For instance, if we reduce 10 days' worth of data into three bundles of three days with three gaps, the data on the 10th day cannot be condensed and hence are not included. The equation to calculate the bundle number is as follows. D is the total number of record dates, n is how many days' worth of data are being reduced, and m is the length of the gap between the start days of each bundle. Only the integer result of the function is allowed.

$$Bundles = \frac{D - (n-1)}{m} \tag{1}$$

Algorithm 1: Training Procedure of DeepHealthNet.

- **Input:** Training raw data
$X = \{x_i\}_{i=1}^{D}, \{x_i\} \in \mathbb{R}^{P \times H}$: Training data for daily health, where D is the total number of days, P is the number of participants, and H is the number of health parameters (i.e., height, weight, BMI, step count, sleep time, kcal intake, exercise, burned kcal)
$\Omega = \{O_i\}_{i=1}^{D}$: Class labels, where
$O_i \in \{Increase, Maintenance, Decrease\}$ and D is the total number of days
- **Output:** Trained DeepHealthNet

- **Step 1:** Preprocessing the input data
1 　Input X_{bin}: Merging data after the outlier removal
2 　Transform n days' worth of data for an individual to a single row by using the average value of n days
3 　Calculate the bundle features using (1) from the input data
4 　Assign the class label according to the features of each bundle
5 　Augment the data using SMOTE ((2))
6 　Output X_{bin}: Preprocessed data with class labels
　 Step 2: Training the network
7 　Input X_{bin}: a set of preprocessed data
8 　Input $\Omega = \{O_{tr}\}_{tr=1}^{D}$: multiclass labels, where $O_{tr} \in \{Increase, Maintenance, Decrease\}$, D is the total number of days
9 　The network parameters are initialized to random values for multiclass labels
10 　Calculate feature maps extracted using (3)–(4)
11 　Generate the loss value using (5)–(9)
2 　Output X_N: Weights and loss values (multiclass)
　 Step 3: Fine-tune parameters
13 　Minimizing loss values by tuning the network parameters

The labels consisted of three categories: weight maintenance, weight gain, and weight loss. The first set of data was labeled as weight maintenance by default due to the absence of a comparison group. The reason that the weight change is set as the main element of labeling is that it is a factor that changes a lot in a short period of time among the key and weight that make up the BMI. In most cases, there was no significant weight change between the previous and current groups when labeled using this method. So there was an imbalance between classes due to the large number of weight maintenance labels. Therefore, we used data augmentation using the synthetic minority oversampling technique (SMOTE) to solve the data imbalance. SMOTE is a popular data augmentation technique used to balance a class distribution by generating synthetic samples of the minority class [41], [42]. It calculates the difference vector between a minority class sample and its nearest neighbor and generate new data points by scaling the difference vector by a random ratio. x_0 represents one of the candidates for integration as a minority class through SMOTE. $I_{B(x_0,r)}$ represents the coverage of the minority class within a

range with a radius of r, centered at x_0. $pX(x)$ represents the original probability density of the minority class.

$$I_{B(x_0,r)} = \int_{B(x_0,r)} pX(x)dx \qquad (2)$$

z, the newly generated point, can be obtained by adding a uniform random variable w multiplied by the vector difference between x_k (a neighboring point) and x_0:

$$\mathrm{z} = (1 - \mathrm{w})x_0 + \mathrm{w}x_k \qquad (3)$$

The expression for the density function of point z can be represented as follows. N and K represent the number of minority class samples and neighboring samples, respectively.

$$
\begin{aligned}
p_{\mathrm{Z}}(\mathrm{z}) = {} & (N - K)\binom{N-1}{K} \int_x pX(x) \int_{r=\|\mathrm{z}-x\|}^{\infty} p_X \\
& \times \left(x + \frac{(z-x)r}{\|\mathrm{z}-x\|} \right) \\
& \times \left(\frac{r^{d-2}}{\|\mathrm{z}-x\|^{d-1}} \right) B\left(1 - I_{B(x,r)}; N - K - 1, K \right) dr dx
\end{aligned}
$$
(4)

D. Deep Learning Model Architecture

The proposed architecture of the deep learning model consists of a neural network with multiple hidden layers to effectively capture complex patterns in the input data (Algorithm 1). The input layer, which is determined by the specific dimensions of the input data, receives the data for processing by the neural network. The first hidden layer comprises 128 densely connected nodes, where each node receives input from all nodes in the previous layer. These nodes apply their individual weights and biases to the inputs received, allowing them to learn and contribute to the representations of the network.

The second hidden layer consists of 256 densely connected nodes, mirroring the connectivity pattern of the previous layer. Each node in this layer receives input from all nodes in the preceding layer and performs its own computations to extract higher-level features. This hierarchical structure enables the network to learn increasingly abstract representations as information flows through the layers.

The third hidden layer consists of 128 densely connected nodes. Similarly to the previous layers, each node in this layer receives inputs from all nodes in the preceding layer. This layer further refines the learned features and contributes to the overall understanding of the input data. The output layer, which depends on the specific task at hand, generates the final outputs of the network. In this case, the number of nodes in the output layer corresponds to the three classes under consideration, namely "Increase", "Maintenance", and "Decrease". Each node in the output layer represents the likelihood or probability of the input belonging to its corresponding class. For activation functions, the rectified linear unit (ReLU) function [43] is commonly used in dense layers, including both the hidden layer. ReLU introduces non-linearity into the network, allowing it to learn complex relationships and adapt to various data patterns. However,

depending on the requirements of the problem, other suitable activation functions can be utilized. The equation for the ReLU activation function is as follows:

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{5}$$

$$f(x) = \max(0, x) \tag{6}$$

The proposed model employs the cross-entropy loss function, which is widely used for multiclass classification problems [44]. Cross-entropy loss measures the dissimilarity between predicted probabilities and the true class labels, guiding the network to minimize this discrepancy during training. By optimizing cross-entropy loss, the model aims to improve its performance for the prediction of obesity rates. The experiment was conducted by fixing the loss function with 300 epochs, in which the loss function optimally converges. The equation for the cross-entropy loss function is expressed as follows: y represents the true one-hot encoded label vector and $\hat{y}$ represents the predicted probability distribution between classes. p is the predicted probability observation of the class.

$$CL(y, \hat{y}) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{7}$$

$$CL(y, \hat{y}) = -\Sigma(y \log(\hat{y})) \tag{8}$$

In addition, the negative log-likelihood (NLL), minimum negative log-likelihood (MNL), and maximum likelihood estimation (MLE) are calculated simultaneously below. p(y) is a scalar rather than a vector. It is the value of the single dimension where the y is the truth of the ground. Thus, it is equivalent to cross-entropy.

$$NLL(y) = -\log(p(y)) \tag{9}$$

$$MNL(y) = \min_{\theta} \sum_{y} -\log(p(y; \theta)) \tag{10}$$

$$MLE(y) = \max_{\theta} \prod_{y} p(y; \theta) \tag{11}$$

E. Comparison Models

Several ML classification methods can be used to predict obesity rates in adolescents using AI technology. We demonstrated that the proposed model outperforms traditional machine learning classification models. The following are some commonly used classification methods for this task:

· Naïve Bayes classifier (NB) [45]: The NB classifier is a probabilistic method that assumes that the features are independent of each other. It is effective for datasets with a large number of features and is computationally efficient.

· Regularized linear discriminant analysis (RLDA) [46]: RLDA is a statistical method used to find a linear combination of features that can effectively divide classes. It is effective for datasets with a small number of features and assumes that the data are normally distributed.

· Random forest (RF) [47]: This is an ensemble learning method that combines multiple decision trees to improve the accuracy of the classification. RF can handle a large number of

features and is useful for feature selection. It handles imbalanced datasets effectively.

· Decision tree (DT) [48]: A DT is a tree-like structure that represents the decision-making process. It is a popular method for classification and can handle both numerical and categorical data. Decision trees are easy to interpret and visualize.

· Support vector machine (SVM) [49]: SVM is a supervised learning algorithm that uses a hyperplane to separate data points into different classes. It is effective for handling high-dimensional data and can handle nonlinearly separable data by using kernel functions [50].

· Long short-term memory (LSTM) [51]: LSTM is a type of recurrent neural network. It can handle sequential data, such as time-series data, and is effective in capturing long-term dependencies in the data.

F. Statistical Analysis

A statistical analysis was performed to evaluate the performance of ML models in predicting obesity rates in adolescents. To ensure the validity of the analysis, normality [52] and homoskedasticity tests [53] were performed on the data, considering the small sample size. The Shapiro-Wilk test was used to verify the normality of the data, and the results showed that the null hypothesis of normality was satisfied. Additionally, we confirmed homoskedasticity using Levene's test [53] for each comparative group.

We conducted a paired t-test to compare the performance of various ML models [54]. This statistical test allowed us to determine the statistical significance between the models and identify the most effective model for predicting obesity rates in adolescents. The results of the paired t-test were analyzed to provide information on the performance of the ML models and to guide the selection of the best model to predict obesity rates.

G. Performance Measurement Metrics

Several evaluation metrics can be used to assess the performance of classification models. We use evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess the performance of the model [55], [56]. These metrics were used to measure the ability of the model to accurately predict the obesity status of adolescents. The results of the evaluation were analyzed to determine the effectiveness of SMOTE and the performance of the ML models. Accuracy, denoting the proportion of correctly classified cases, gives a broad perspective on the overall correctness of the model (12). In contrast, precision emphasizes the ratio of true positive (TP) predictions among all positive predictions, shedding light on the model's ability to minimize false positives (13). Recall, also known as sensitivity, computes the proportion of TP predictions among all actual positive instances, indicating the model's capacity to identify positive cases (14). Furthermore, the F1-score, which amalgamates precision and recall, provides a balanced metric considering both false positives (FP) and false negatives (FN) (15). TP represents true positive, which refers to cases where both the real and predicted labels are true. True negatives (TN) represent true negatives, indicating cases where both the real and predicted labels are false. FP represents

TABLE II
PERFORMANCE EVALUATION OF DEEPHEALTHNET USING PERFORMANCE METRICS, SUCH AS ACCURACY, F1-SCORE, RECALL, AND PRECISION

Models	1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold	Accuracy	F1-Score	Recall	Precision
NB	0.3700	0.3704	0.3719	0.3724	0.3711	0.3721	0.3708	0.3706	0.3719	0.3727	0.3714	0.2536	0.5546	0.3714
RLDA	0.4965	0.4969	0.4966	0.4957	0.4967	0.4978	0.4956	0.4967	0.4960	0.4953	0.4964	0.4939	0.4995	0.4964
RF	0.5470	0.5473	0.5478	0.5474	0.5475	0.5479	0.5464	0.5478	0.5486	0.5492	0.5477	0.5365	0.5530	0.5477
DT	0.6614	0.6618	0.6613	0.6614	0.6602	0.6605	0.6603	0.6596	0.6607	0.6613	0.6608	0.6616	0.6881	0.6608
SVM	0.6851	0.6839	0.6841	0.6853	0.6850	0.6851	0.6847	0.6855	0.6834	0.6836	0.6846	0.6850	0.6891	0.6846
LSTM	0.6948	0.6965	0.6984	0.7018	0.7015	0.7021	0.7021	0.7033	0.7024	0.7045	0.7008	0.7000	0.7037	0.7008
DeepHealthNet	**0.8847**	**0.8854**	**0.8820**	**0.8836**	**0.8838**	**0.8824**	**0.8833**	**0.8825**	**0.8851**	**0.8842**	**0.8837**	**0.8797**	**0.8958**	**0.8837**

false positive, indicating cases where the real label is false, but the predicted label is true. Finally, FN denotes false negative, indicating cases where the real label is true, but the predicted label is false.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (15)$$

In addition, it is crucial to evaluate the models on different subsets of the data using 10-fold cross-validation to obtain reliable performance estimates. This approach could support obtain a more robust assessment of the performance and generalization ability of the model. In this study, the models were trained on 90% of the data and tested on the remaining 10% of the data.

III. EXPERIMENTAL RESULTS

A. Predicted Performances Using Proposed and Comparison Models

The experimental results presented in Table II demonstrate the superior performance of the proposed deep learning framework. It achieved an impressive average accuracy of 0.8837, outperforming all compared models. The F1-score, which measures the balance between precision and recall, was also high at 0.8797. The recall value, which indicates the proportion of true positives identified, was 0.8958, while the precision value, representing the accuracy of positive predictions, matched the overall accuracy at 0.8837. Compared to the other models, the LSTM model exhibited a relatively high accuracy of 0.7008. Among traditional ML classifiers, SVM demonstrated a commendable performance with an accuracy of 0.6846. In contrast, the NB model exhibited the lowest accuracy of 0.3714, which is comparable to random chance accuracy. Therefore, DeepHealthNet significantly outperformed all models compared in terms of accuracy, F1-score, recall, and precision. Across different folds, the standard deviation remained consistently below 0.02 for all models. Therefore, there was no significant variance or instability during the training process. Performance metrics were consistently distributed, suggesting the robustness and reliability of the models throughout the training phase.

TABLE III
STATISTICAL ANALYSIS OF DIFFERENCES BETWEEN THE PROPOSED AND COMPARED MODELS IN TERMS OF GRAND-AVERAGE PREDICTED PERFORMANCES (ACCURACY, F1-SCORE, RECALL, PRECISION)

Comparison models	Accuracy	F1-Score	Recall	Precision
NB vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
RLDA vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
RF vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
DT vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
SVM vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
LSTM vs. DeepHealthNet	$p<0.001$	$p<0.05$	$p<0.05$	$p<0.05$

Overall, the experimental results confirmed the effectiveness of DeepHealthNet, which consistently demonstrated superior accuracy and outperformed the models compared in terms of various performance measures.

Furthermore, the results of the statistical analysis presented in Table III indicate the differences between the proposed deep learning framework and the comparison models in terms of various performance metrics. The obtained p-values indicate the level of statistical significance and help assess whether the observed differences are coincidental or truly meaningful. As the p-values were consistently less than 0.001 for accuracy, F1-score, recall, and precision, there is a highly significant statistical difference between the proposed deep learning framework and the comparison models in terms of these measures.

However, when comparing the proposed deep learning framework with the LSTM model in terms of recall and precision, the obtained p-value was less than 0.05, indicating a statistically significant difference. Although the proposed model generally outperformed the LSTM, this finding suggests that there are certain scenarios in which the LSTM might exhibit comparable performance in terms of recall and precision.

B. Performance Measurement Using Confusion Matrix

The study conducted a detailed analysis of the confusion matrix for the proposed models to analyze their performance for each class using k-fold datasets, as illustrated in Fig. 3. The average accuracy of the proposed model was determined to be 0.8837, with the "Decrease" class exhibiting the highest true positive value at 0.9905. The "Increase" and "Maintenance" classes showed a similar TP value of 0.9204 and 0.9720, respectively. Although there were some variations in performance between different classes, overall performance did not show significant differences. Therefore, the proposed model was effective in

TABLE IV
PERFORMANCE EVALUATION OF BOY AND GIRL GROUPS USING PERFORMANCE METRICS, SUCH AS ACCURACY, F1-SCORE, RECALL, AND PRECISION

|	Boy group					Girl group			
Models	Accuracy	F1-Score	Recall	Precision	Models	Accuracy	F1-Score	Recall	Precision
NB	0.3779	0.2612	0.4822	0.3779	NB	0.4076	0.3297	0.5216	0.4076
RLDA	0.5089	0.5043	0.5156	0.5089	RLDA	0.5087	0.5082	0.5099	0.5087
RF	0.6459	0.6472	0.6556	0.6459	RF	0.5840	0.5815	0.6005	0.5840
DT	0.7787	0.7787	0.7991	0.7787	DT	0.6914	0.6931	0.7302	0.6914
SVM	0.7965	0.7964	0.7993	0.7965	SVM	0.6909	0.6906	0.6987	0.6909
LSTM	0.6493	0.6491	0.6500	0.6493	LSTM	0.7675	0.7664	0.7678	0.7675
DeepHealthNet	**0.9320**	**0.9318**	**0.9346**	**0.9320**	DeepHealthNet	**0.9163**	**0.9155**	**0.9206**	**0.9163**
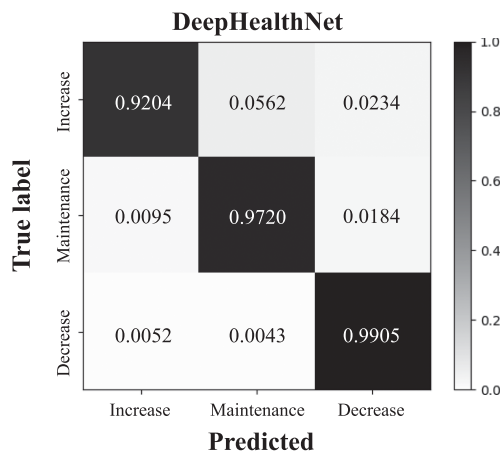
Fig. 3. Confusion matrices of each class (Increase, Maintenance, and Decrease) across all the participants using DeepHealthNet.

predicting obesity rates in adolescents, demonstrating comparable performance across all classes.

Among the models compared, the NB model achieved the highest TP value of 0.9171 for the "Increase" class. For the "Maintenance" and "Decrease" classes, the LSTM model exhibited the highest TP values of 0.7765 and 0.9203, respectively. These findings suggest that the NB model correctly identified instances of increased obesity rates, while the LSTM model recognized instances of maintaining weight or decreasing obesity levels. The high accuracy observed in this experiment can be attributed to the availability of sufficient existing data to measure obesity in adolescents over a short period and to provide valuable feedback, such as appropriate exercise recommendations, to manage obesity in this population. Consequently, the proposed model provided accurate predictions of obesity rates in adolescents, exhibiting high precision.

Analysis of the confusion matrix confirmed the effectiveness of the proposed model in predicting obesity rates in adolescents. The proposed model demonstrated the highest TP in all classes.

C. Comparison of Predicted Performances Between Boy and Girl Groups

Table IV presents the results of the obesity prediction by dividing the dataset into the boy and girl groups. The study aimed to explore whether there are differences in the manifestation of

obesity between boys and girls and whether it would be more effective to analyze the prediction performance separately for each group. The proposed deep learning framework was utilized to assess the performance in predicting obesity in both groups. The results revealed that the obesity prediction performance achieved by the proposed model was 0.9320 for the boy group and 0.9163 for the girl group. Results of boy groups, there was an improvement in performance by 0.0157 compared to results of girl groups. The F1-scores for the boy and girl groups were 0.9318 and 0.9155, respectively, further indicating a comparable performance between the genders. In particular, the prediction performance in the boy group was approximately 2% higher than in the girl group. The proposed deep learning framework demonstrated the highest predictive performance for obesity, regardless of gender. Specifically, SVM showed the highest performance (0.7965) in the boy group, while LSTM (0.7675) outperformed the other models in the girl group. The performance improved because the separation of the data reduced the sample size and made the features of each gender more visible.

Table V presents an analysis of the statistically significant differences within each gender group for the compared models and the proposed deep learning framework. The aim of this analysis was to examine whether there were differences in the obesity prediction performance between the boy and girl groups and to determine if the proposed model outperformed the compared models within each group. Statistical analysis revealed that the obesity prediction performance of the proposed deep learning framework was more statistically significant ($p < 0.001$) compared to the other general models, even when considering the boy and girl groups separately. This highlights the superiority of the proposed model in accurately predicting obesity in both genders.

Furthermore, the results presented in Table V can be used to examine whether there were significant differences in the obesity prediction performance between the boys and girls groups for each of the compared models. The results showed that for models such as DT, SVM, and LSTM, there was a statistically significant difference in performance between the boy and girl groups ($p < 0.005$ and $p < 0.001$). However, while there were no statistically significant differences for the other compared models (i.e., NB, RLDA, RF, and the proposed model), their obesity prediction performance was significantly lower. Overall, the proposed deep learning framework demonstrated a higher obesity prediction performance than the other models, without significant differences between the gender groups. This indicates

|	Within boy group				Within girl group					between boy and girl group			
Comparison models	Accuracy	F1-Score	Recall	Precision	Accuracy	F1-Score	Recall	Precision	Comparison models	Accuracy	F1-Score	Recall	Precision
NB vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	NB vs. NB	$p>0.01$	$p>0.01$	$p>0.01$	$p>0.01$
RLDA vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	RLDA vs. RLDA	$p>0.01$	$p>0.01$	$p>0.01$	$p>0.01$
RF vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	RF vs. RF	$p>0.01$	$p>0.01$	$p>0.01$	$p>0.01$
DT vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	DT vs. DT	$p<0.005$	$p<0.005$	$p>0.01$	$p<0.005$
SVM vs. DeepHealthNet	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	SVM vs. SVM	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
LSTM vs. DeepHealthNet	$p<0.001$	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.001$	$p<0.05$	$p<0.05$	$p<0.05$	LSTM vs. LSTM	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
									DeepHealthNet vs. DeepHealthNet	$p>0.01$	$p>0.01$	$p>0.01$	$p>0.01$
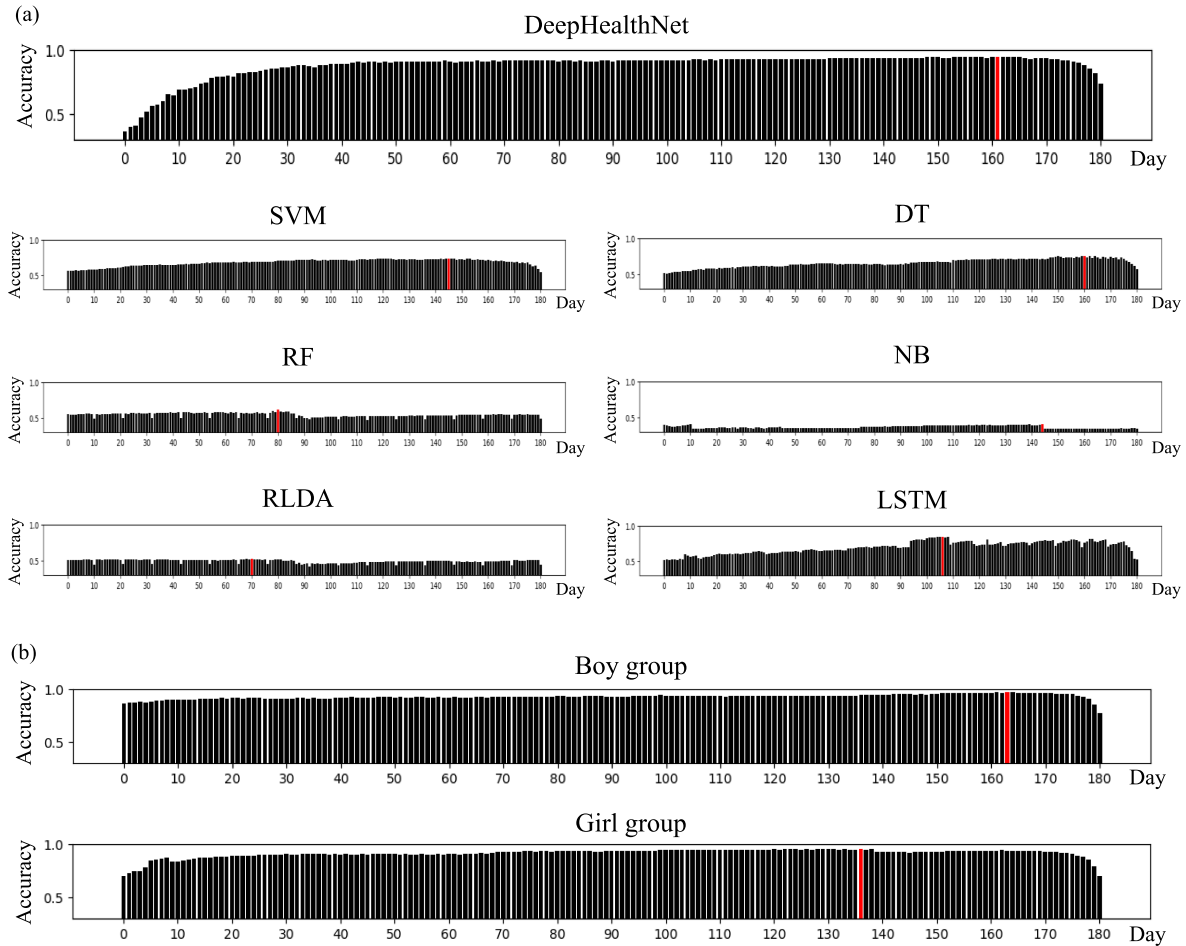
Fig. 4. Comparison of classification accuracy per nth-day sessions. (a) Classification accuracy per day session using DeepHealthNet and the compared models. The red bar indicates the highest accuracy across all day sessions. (b) Comparison of classification accuracy using DeepHealthNet by gender. In the 164th-day session, the boy group achieved the highest accuracy, and the girl group exhibited the highest accuracy in the 137th-day session.

that the proposed model can effectively address the issue of unbalanced data according to gender. The preprocessing module employed in the framework helps mitigate this issue, ensuring reliable and accurate predictions regardless of the gender distribution in the dataset.

D. Performance Comparison Using Each Day Session

Fig. 4(a) illustrates the obesity prediction performance for each day session. The graph illustrates the relationship between the day session and the corresponding predictive performance of the proposed model. As day sessions progressed, the predictive performance of obesity gradually improved. According to the proposed model, around the 60th-day session, the performance reached saturation, indicating that the model had learned and captured the underlying patterns in the data. In particular, the highest obesity prediction performance was achieved in the 162nd-day session, recorded as 0.8848. However, performance gradually decreased after the 170th-day session.

In contrast, the prediction performance of the other compared model groups showed different patterns. For example, in the case of LSTM, the highest performance was observed in the

106th-day session, but subsequently the performance exhibited fluctuations with alternating increases and decreases. These findings suggest that the proposed model can consistently predict the obesity levels of the target population for future time intervals, typically spanning approximately 5 to 6 months. However, the performance of the other compared models showed uncertainty in determining the appropriate time duration to collect new data sets to achieve accurate predictions. The performance of the compared models varied and the required period of data collection to achieve reliable predictions is unclear. These results emphasized the effectiveness and long-term predictive capabilities of the proposed model in predicting obesity levels over an extended period.

Fig. 4(b) shows the temporal evolution of obesity prediction performance using the proposed model, specifically for each gender group. The patterns depicted in Fig. 4(b) show the trajectory of obesity prediction performance for both genders. The graph shows the variations in performance over several days for boys and girls. In the 164th-day session, the boy group achieved the highest result (0.9712), similar to the 162nd-day session, which encompassed the entire data sample. In contrast, the group of girls exhibited the highest performance (0.9549) in the 137th-day session. The day session that represented the highest performance between the two gender groups had a one-month difference. This suggests that significant physical changes, including BMI, occurred over approximately four months for the girl group. In contrast, six months of daily health data were required to accurately predict BMI for the boy group, indicating a slower pace of physical change compared to the girl group. These contrasting trends in physical changes between the male and female groups were reflected in the deep learning model. The findings confirmed that the obtained data were also considered, given the faster rate of physical change in the girl group compared to the boy group. The results of this section also show the practicality of being able to quickly apply to real life with the optimized date shown in the experiment without having to experiment with all day sessions when applied in real-world.

E. Convergence Curve for Model Training

The convergence process of the proposed model was examined to validate its stability and reliability. This process was visualized through the error change curve and the accuracy values across different epochs, as depicted in Fig. 5. The proposed model achieved convergence with saturation after approximately 100 epochs in terms of training accuracy. Beyond this point, the accuracy values remained consistently high, exceeding 0.9. This indicates that the model learned effectively from the training data and reached a stable performance level. Furthermore, the loss values of the proposed model reached saturation at approximately 150 epochs, where they remained below 0.1. This means that the model minimized its training loss and obtained a low error, making it optimal for the training data.

Overall, the convergence analysis demonstrated that the proposed model became stable after approximately 100 epochs for the training accuracy and 150 for the loss values. This indicates
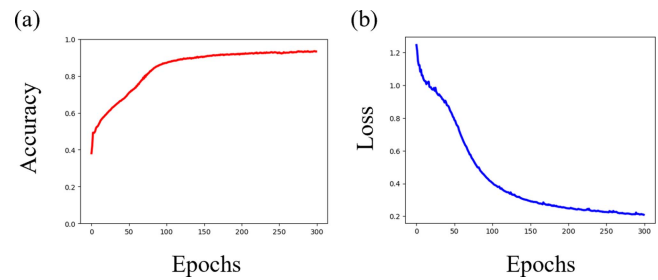

Fig. 5. Convergence curve of the training error and classification accuracy. (a) Training loss. The errors were calculated to obtain the optimized parameters during network training. (b) Predicted performance according to the number of epochs.

that the model effectively learned from the data and could provide reliable predictions with high accuracy while maintaining low levels of error.

IV. DISCUSSIONS

The results of the study demonstrated the excellent performance of the proposed deep learning framework for predicting obesity levels in adolescents. The experimental findings indicated that the proposed model outperformed all the compared models in terms of accuracy, F1-score, recall, and precision. The average accuracy achieved by the proposed model was 0.8837, exceeding the performance of all other models. The F1-score, which measures the balance between precision and recall, was also high at 0.8797. The recall value, which represents the proportion of true positives identified, was 0.8958, and the precision value, indicating the accuracy of positive predictions, matched the overall accuracy at 0.8837. Furthermore, statistical analysis indicated the significance of the differences between the proposed deep learning framework and the models compared in terms of various performance metrics. The obtained p-values were consistently below 0.001, indicating a highly significant statistical difference in favor of the proposed model.

The study also investigated the prediction performance of the model by dividing the dataset into boy and girl groups. As a result, the proposed model showed a better obesity prediction performance when boys and girls were divided than in the combined group. When training in the boy groups, it showed slightly better performance than when training in the girl groups. However, both showed higher accuracy than 0.9. This finding suggested that the proposed deep learning framework was effective in predicting obesity rates regardless of gender, demonstrating its robustness and generalizability. The ability of the model to perform well in both the boy and girl groups is significant because it demonstrates that the model is not biased toward a particular gender. This suggests that the features and patterns captured by the model are relevant and applicable to both genders to predict obesity levels. This finding has practical implications, as it indicates that the model can be used in diverse populations without the need for gender-specific modifications or adjustments. It can help prevent obesity in real-time by applying it to applications such as WUDI! to predict obesity by analyzing data in real-time.

Furthermore, the statistical analysis conducted within each gender group reinforced the superiority of the proposed model. Even when considering the boys and girls groups separately, the obesity prediction performance of the model was more statistically significant compared to those of the other models. This highlights its effectiveness in accurately predicting obesity levels in both boys and girls and provides additional evidence for its robustness. Overall, the prediction performance in separate boy and girl groups emphasized the effectiveness of the proposed deep learning framework in accurately predicting obesity levels regardless of gender. This strengthens the applicability and reliability of the model in diverse populations and contributes to its generalizability.

The proposed model's capacity to forecast obesity over time was demonstrated through the visualization of its performance, which yielded significant knowledge about its long-term predictive potential. The graph demonstrated a gradual improvement in the model's performance as the duration increased, with the highest prediction performance achieved on the 162nd day session. This indicates that the proposed model could consistently forecast obesity levels over an extended period. The improvement in performance over time suggests that the model benefits from a longer duration of data collection and learning. As more data become available, the model can capture more patterns and relationships, leading to improved predictive accuracy. The saturation point was reached after approximately 60 days, suggesting that model performance plateaus and that collecting data beyond this point may not significantly improve predictive capabilities. The compared models exhibited varying levels of performance, highlighting the uncertainty associated with determining the optimal duration to collect new data sets. In contrast, the proposed model demonstrated consistent and reliable long-term predictions, outperforming the compared models. Performance visualization of the model over time indicates that the proposed model can be utilized for long-term obesity prediction, offering reliable forecasts beyond shorter timeframes. This knowledge is valuable in decision-making processes related to interventions, public health policies, and resource allocation, as it allows for more accurate and informed predictions of obesity levels over long periods.

However, this study has some limitations. First, the study focused on a specific age group, and the generalizability of the proposed model to other age groups or populations needs further investigation. Second, the study utilized a specific dataset for training and evaluation, and the generalizability of the proposed model to other datasets should be explored. Furthermore, the study did not consider certain factors that could influence obesity, such as socioeconomic status, dietary habits, or genetic factors. Incorporating these factors into the model should enhance its predictive capabilities. Furthermore, the study employed a deep learning framework, which may require substantial computational resources and expertise for implementation. The feasibility and practicality of the proposed model in real-world settings need further consideration. Finally, the study did not perform a longitudinal analysis to assess the performance of the model over a period of time beyond the available data. In future studies, we look forward to investigating the stability and accuracy of the model over a more extended timeframe.

V. CONCLUSION AND FUTURE WORKS

This study proposed a deep learning framework to predict obesity levels in adolescents. The proposed model demonstrated superior performance compared to other models, achieving high accuracy, F1-score, recall, and precision values. Statistical analysis confirmed the significant differences in favor of the proposed model. The effectiveness of the model was consistent between gender groups, highlighting its robustness. Visualizations proved the model's ability to provide reliable long-term predictions, outperforming the compared models, and the convergence curve showed the model was optimized. Therefore, considering the limitations of this study and future research directions, we plan to further improve the generalizability of the model by using specific data such as mental factors and a residential area, incorporating additional factors and assessing its performance over extended periods.

REFERENCES

[1] L. Abarca-Gómez et al., "Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: A pooled analysis of 2416 population-based measurement studies in 128 9 million children, adolescents, and adults," *Lancet*, vol. 390, no. 10113, pp. 2627–2642, 2017.

[2] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapii, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," *Comput. Biol. Med.*, vol. 136, 2021, Art. no. 104754.

[3] A. Ferreras et al., "Systematic review of machine learning applied to the prediction of obesity and overweight," *J. Med. Syst.*, vol. 47, no. 1, pp. 1–11, 2023.

[4] K. DeGregory et al., "A review of machine learning in obesity," *Obesity Rev.*, vol. 19, no. 5, pp. 668–685, 2018.

[5] C. A. C. Montañez et al., "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 2743–2750.

[6] L. Tim, R. Jackson-Leach, J. Powis, H. Brinsden, and M. Gray, *World Obesity Atlas*. London, U.K.: World Obesity Federation, 2023.

[7] S. Caprio, N. Santoro, and R. Weiss, "Childhood obesity and the associated rise in cardiometabolic complications," *Nature Metab.*, vol. 2, no. 3, pp. 223–232, 2020.

[8] B. Singh and H. Tawfik, "Machine learning approach for the early prediction of the risk of overweight and obesity in young people," in *Proc. 20th Int. Conf. On Comput. Sci.*. Springer, 2020, pp. 523–535.

[9] M. N. LeCroy, R. S. Kim, J. Stevens, D. B. Hanna, and C. R. Isasi, "Identifying key determinants of childhood obesity: A narrative review of machine learning studies," *Childhood Obesity*, vol. 17, no. 3, pp. 153–159, 2021.

[10] M. L. Ahmed, K. K. Ong, and D. B. Dunger, "Childhood obesity and the timing of puberty," *Trends Endocrinol. Metab.*, vol. 20, no. 5, pp. 237–242, 2009.

[11] R. Hammond et al., "Predicting childhood obesity using electronic health records and publicly available data," *PLoS One*, vol. 14, no. 4, 2019, Art. no. e0215571.

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

[13] J. Torner, S. Skouras, J. L. Molinuevo, J. D. Gispert, and F. Alpiste, "Multipurpose virtual reality environment for biomedical and health applications," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1511–1520, Aug. 2019.

[14] W. Saadeh, S. A. Butt, and M. A. B. Altaf, "A patient-specific single sensor IoT-based wearable fall prediction and detection system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 995–1003, May 2019.

[15] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Internet of things: Device capabilities, architectures, protocols, and smart applications in healthcare domain," *IEEE Internet Things J.*, 2022.

[16] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[17] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[18] J.-H. Cho, J.-H. Jeong, and S.-W. Lee, "Neurograsp: Real-time EEG classification of high-level motor imagery tasks using a dual-stage deep learning framework," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13279–13292, Dec. 2022.

[19] Z. Zhao, F. Zhou, K. Xu, Z. Zeng, C. Guan, and S. K. Zhou, "LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 633–646, Mar. 2023.

[20] J.-H. Jeong, J.-H. Cho, B.-H. Lee, and S.-W. Lee, "Real-time deep neurolinguistic learning enhances noninvasive neural language decoding for brain–machine interaction," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7469–7482, Dec. 2023.

[21] Z. Zhao, Z. Zeng, K. Xu, C. Chen, and C. Guan, "DSAL: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 10, pp. 3744–3751, Oct. 2021.

[22] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018.

[23] M. Lee, J.-H. Jeong, Y.-H. Kim, and S.-W. Lee, "Decoding finger tapping with the affected hand in chronic stroke patients during motor imagery and execution," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1099–1109, 2021.

[24] M. Gupta, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti, "Obesity prediction with EHR data: A deep learning approach with interpretable elements," *ACM Trans. Comput. Healthcare*, vol. 3, no. 3, pp. 1–19, 2022.

[25] X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino, "Prediction of early childhood obesity with machine learning and electronic health record data," *Int. J. Med. Informat.*, vol. 150, 2021, Art. no. 104454.

[26] J.-H. Jeong, N.-S. Kwak, C. Guan, and S.-W. Lee, "Decoding movement-related cortical potentials based on subject-dependent and section-wise spectral filtering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 3, pp. 687–698, Mar. 2020.

[27] P. Sundaravadivel, E. Kougianos, S. P. Mohanty, and M. K. Ganapathiraju, "Everything you wanted to know about smart health care: Evaluating the different technologies and components of the Internet of Things for better health," *IEEE Consum. Electron. Mag.*, vol. 7, no. 1, pp. 18–28, Jan. 2018.

[28] L. Bastida et al., "Promoting obesity prevention and healthy habits in childhood: The OCARIoT experience," *IEEE J. Transl. Eng. Health Med.*, vol. 11, pp. 261–270, 2023.

[29] G. Yang et al., "Homecare robotic systems for healthcare 4.0: Visions and enabling technologies," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 9, pp. 2535–2549, Sep. 2020.

[30] F. G. Huffman, S. Kanikireddy, and M. Patel, "Parenthood-a contributing factor to childhood obesity," *Int. J. Environ. Res. Public Health*, vol. 7, pp. 2800–2810, 2010.

[31] M. S. Sothern, "Exercise as a modality in the treatment of childhood obesity," *Pediatr. Clin. North America*, vol. 48, no. 4, pp. 995–1015, 2001.

[32] X. Chen, M. A. Beydoun, and Y. Wang, "Is sleep duration associated with childhood obesity? A. systematic review and meta-analysis," *Obesity*, vol. 16, no. 2, 2008, Art. no. 265.

[33] W. L. Haskell et al., "Physical activity and public health: Updated recommendation for adults from the american college of sports medicine and the american heart association," *Circulation*, vol. 116, no. 9, 2007, Art. no. 1081.

[34] M. d. A. Mendes et al., "Metabolic equivalent of task (METs) thresholds as an indicator of physical activity intensity," *PLoS One*, vol. 13, no. 7, 2018, Art. no. e0200701.

[35] P. K. Mondal, K. H. Foysal, B. A. Norman, and L. S. Gittner, "Predicting childhood obesity based on single and multiple well-child visit data using machine learning classifiers," *Sensors*, vol. 23, no. 2, 2023, Art. no. 759.

[36] A. Gasmi, "Machine learning and bioinformatics for diagnosis analysis of obesity spectrum disorders," 2022, *arXiv:2208.03139*.

[37] K. Abualsaud, "Machine learning algorithms and Internet of Things for healthcare: A survey," *IEEE Internet Things Mag.*, vol. 5, no. 2, pp. 133–139, Jun. 2022.

[38] N. S. Sworna, A. M. Islam, S. Shatabda, and S. Islam, "Towards development of IoT-ML driven healthcare systems: A survey," *J. Netw. Comput. Appl.*, vol. 196, 2021, Art. no. 103244.

[39] E. R. Cheng, R. Steinhardt, and Z. Ben Miled, "Predicting childhood obesity using machine learning: Practical considerations," *BioMedInformatics*, vol. 2, no. 1, pp. 184–203, 2022.

[40] E. Lee, J. Jung, G.-M. Moon, S.-W. Lee, and J.-H. Jeong, "WUDI: A human involved self-adaptive framework to prevent childhood obesity in Internet of Things environment," 2023, *arXiv:2308.15944*.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[42] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," in *Mach. Learn.*, pp. 1–21, 2023.

[43] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 597–607.

[44] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.

[45] E. M. K. Reddy, A. Gurrala, V. B. Hasitha, and K. V. R. Kumar, "Introduction to naive bayes and a review on its subtypes with applications," in *Bayesian Reasoning Gaussian Processes Mach. Learn. Appl.*, pp. 1–14, 2022.

[46] J. Zhao, L. Shi, and J. Zhu, "Two-stage regularized linear discriminant analysis for 2D data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1669–1681, Aug. 2015.

[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[48] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *Int. J. Inf. Decis. Sci.*, vol. 12, no. 3, pp. 246–269, 2020.

[49] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Their Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[50] W. S. Noble, "What is a support vector machine," *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[52] K. R. Das and A. Imon, "A brief review of tests for normality," *Amer. J. Theor. Appl. Statist.*, vol. 5, no. 1, pp. 5–12, 2016.

[53] B. B. Schultz, "Levene's test for relative variation," *Systematic Zool.*, vol. 34, no. 4, pp. 449–456, 1985.

[54] A. Ross, V. L. Willson, A. Ross, and V. L. Willson, "Paired samples t-test," in *Basic Adv. Stat. Tests: Writing Results Sections Creating Tables Figures*, pp. 17–19, 2017.

[55] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[56] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.