



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85100>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

AI Fitness OS: Verifiable, Multi-Metric Fitness Guidance by Constraining a Large Language Model with Deterministically Computed Targets

Selvin Lopes¹, Perseus Bhavnagri²

Student, Department of Information Technology, Wilson College (Autonomous), University of Mumbai, Mumbai, India

Abstract: Fitness applications usually record one behaviour in isolation, and large-language-model (LLM) coaches produce fluent advice whose numbers are rarely checked. This paper presents AI Fitness OS, a web application that keeps training, nutrition, hydration, sleep, weight, steps and goals in one per-user record, computes every numerical target deterministically from published equations and stated heuristics, and passes those targets to Google Gemini as binding constraints; each generated meal plan is re-checked in code against a $\pm 5\%$ tolerance. The premise was examined through secondary research on a public dataset of 2,111 records, and the artefact through tests, simulation and live measurement. All eight lifestyle metrics that the system tracks or addresses were significantly associated with obesity level (Bonferroni-adjusted $\alpha = 0.00625$), although only three remained so on 515 records resembling the original survey. A Random Forest that combined the eight metrics reached a macro-F1 of 0.720 against 0.263 for the best single metric (Wilcoxon $W = 55$, $p = 0.00098$). Across 1,000 simulated user profiles, every constrained meal plan fell within $\pm 5\%$ of its calorie target, against 75 of 1,000 unconstrained plans (exact McNemar $p < 10^{-200}$). The calculator agreed with the dataset's weight-category labels for 98.82 % of records, and live responses took 2.5–6.1 s at no cost. The findings support computing arithmetic, generating language and checking the output against the computed numbers.

Keywords: Fitness management; large language models; Gemini; constrained generation; health calculator; Mifflin–St Jeor; obesity levels; secondary research; Flutter Web; Firebase.

I. INTRODUCTION

Regular physical activity, a balanced diet, adequate hydration and sufficient sleep are interdependent foundations of health, yet the digital tools that people use to manage them are usually designed around one behaviour at a time. A user may count steps in one application, log meals in another and track sleep on a wearable, with no single place in which these signals meet. This study approaches the problem by building and evaluating an integrated, AI-assisted fitness management system and by testing, on public data, the assumptions on which its design rests.

A. Fragmented Self-tracking

Evidence on standalone applications is sobering. A systematic review by He et al. [1] screened 8,403 records, included seven studies (four of them high-quality randomised controlled trials) and found no significant effect of smartphone application interventions on physical activity in healthy young adults.

Self-monitoring itself is also fragile: Lin et al. [2], modelling 21 days of dietary logging by 97 participants in a digital weight-loss programme, found that goal pursuit dominated adherence while the contribution of habit diminished, and that tailored feedback helped to sustain logging. Both findings suggest that recording data is not enough; the data must be connected to feedback that is personal and timely.

B. Large Language Models as Coaches

Systems such as PlanFitting [3] for exercise planning and ChatDiet [4] for nutrition-oriented food recommendation show that an LLM can turn a user's context into readable, personalised advice. However, an LLM produces text by prediction rather than by calculation, so a calorie total or a protein figure inside a generated plan is not guaranteed to be correct, and a user without training cannot easily tell when it is wrong.

C. The AI Fitness OS Approach

AI Fitness OS separates the two tasks. A deterministic calculator computes body mass index (BMI) with the World Health Organization categories [5], basal metabolic rate (BMR) with the Mifflin–St Jeor equation [6], total daily energy expenditure (TDEE), a goal-adjusted calorie target with a safety floor, protein, fat, carbohydrate and water targets. These numbers are embedded in every prompt as binding constraints for Google Gemini (model gemini-3.5-flash-lite, accessed through Firebase AI Logic on the free Gemini Developer API), which writes workout plans, four-meal Indian-style meal plans, post-workout meal suggestions, healthier swaps, a daily tip and coaching answers.

Every generated meal plan is then re-checked in code: if its total misses the calorie target by more than 5 %, the portions are rescaled proportionally and the user is shown whether the plan meets the target. The application is built with Flutter Web, Riverpod, Firebase Authentication and Cloud Firestore on Firebase's no-cost Spark plan with no billing account, so the system costs nothing to operate.

II. PROBLEM STATEMENT AND NEED FOR THE STUDY

Fitness information that is spread across separate applications cannot support decisions that depend on more than one behaviour, such as lightening a workout after a short night's sleep or adjusting a meal plan to a calorie deficit. At the same time, LLM-based coaches can personalise advice in natural language, but their numerical content is not verified, and evaluations in this field are still immature.

There is a need for a fitness management system that brings the relevant metrics into one record, derives its numbers from published equations rather than from the language model, and demonstrates — with evidence rather than assertion — that the advice generated stays within the computed targets, at no cost to developer or user.

A. Need for the Study

This study is undertaken to —

- 1) examine, on public data, whether the lifestyle metrics that a multi-metric fitness system tracks are actually associated with weight status;
- 2) test whether combining those metrics carries more information than any single metric, which is the empirical case for an integrated record;
- 3) design and build a system in which an LLM is constrained by deterministically computed targets and checked in code afterwards;
- 4) measure whether that constraint mechanism makes generated meal plans verifiably accurate, and at what latency and cost;
- 5) report honestly the limitations of the dataset, the simulation and the small live-LLM sample.

III. LITERATURE REVIEW

The relevant research falls into four strands: LLM coaching for exercise and health, LLM nutrition planning, evaluation of such systems, and evidence on self-tracking and target equations.

- 1) LLM-based exercise and health coaching. Shin et al. [3] built PlanFitting, an LLM-driven conversational agent that helps users create and refine weekly exercise plans. It was evaluated through a user study with 18 participants, an intrinsic evaluation, which found that the generated plans reliably followed established exercise guidelines, and an expert evaluation in which three planners rated their frequency, intensity and time composition above average while also identifying room to improve the combination of exercise types. Jörke et al. [7] designed GPTCoach, a GPT-4 coaching chatbot that implements the onboarding conversation of an evidence-based programme, uses motivational interviewing and queries wearable data through tool use; it was informed by interviews with 12 health professionals and 10 potential coaching recipients and evaluated in a laboratory study with 16 participants. The authors note that the probabilistic nature of LLM outputs requires a fundamentally different system design from rule-based chatbots. At larger scale, Cosentino et al. [8] fine-tuned Gemini into a personal health LLM, evaluated it on 857 expert-curated case studies, found its fitness responses not statistically different from those of experts, and reported scores of 88 % and 79 % on fitness and sleep examinations respectively. Heydari et al. [9] proposed a multi-agent Personal Health Agent with data-science, domain-expert and health-coach sub-agents, evaluated across 10 benchmark tasks with more than 7,000 annotations; they observe that LLM reasoning over complex numerical data remains a significant challenge.

- 2) LLM-based nutrition and meal planning. Yang et al. [4] introduced ChatDiet, which combines a causal personal model and a population model through an orchestrator that supplies retrieved information to the LLM, improving the personalisation and explainability of food recommendations. Khamesian et al. [10] built NutriGen, which uses prompt engineering and a personalised nutrition database to generate meal plans aligned with user-defined calorie targets; the best models, Llama 3.1 8B and GPT-3.5 Turbo, showed percentage errors of 1.55 % and 3.68 %, while others deviated more, up to 24.67 % for GPT-4o Mini. Gao et al. [11] combined a knowledge graph with an LLM in HealthGenie to provide explainable dietary guidance, evaluated in a within-subject study with 12 participants. Azimi et al. [12] tested GPT-4o, Claude 3.5 Sonnet and Gemini 1.5 Pro on 1,050 registered-dietitian examination questions under four prompting strategies and found that both accuracy and consistency varied with the prompt and the domain. In these works, numerical accuracy is measured after the fact rather than guaranteed.
- 3) Evaluation of LLM coaches. In a scoping review of 20 studies on LLM-based exercise and health coaching, Lai et al. [13] found that 75 % used proprietary models, the median evaluation rigour was 2.5 out of 5, 55 % of studies were of low rigour, and 45 % reported reliability inconsistently. The field therefore lacks reproducible, quantitative checks of whether generated advice is correct.
- 4) Self-tracking and target equations. Beyond the null result of He et al. [1] and the adherence dynamics described by Lin et al. [2], Sun et al. [14] reported an S-shaped (curvilinear) effect of gamification feature richness on exercise-adherence intention: more features are not always better. On the computation side, the Mifflin–St Jeor equation [6] predicts resting energy expenditure from weight, height, age and sex; Navalta et al. [15], in a systematic-review protocol, point out that resting-metabolic-rate prediction equations often rely on binary sex categories and limited derivation populations, so their outputs are estimates rather than measurements. For hydration, Stookey et al. [16], studying 5,817 adults aged 40–70, associated better hydration with a median total water intake above 40 mL/kg including at least 1 L/day of plain water.

IV. GAP ANALYSIS

The reviewed systems are strong in one domain — exercise [3], [7] or nutrition [4], [10], [11] — but none keeps workouts, food, water, sleep, weight and steps in one record from which every recommendation is generated. Where LLMs produce numbers, accuracy is reported as an average error [10], [12] rather than enforced: a plan that misses its target is still shown to the user. Large personal-health models [8], [9] depend on fine-tuning, wearables or research infrastructure that a student or a small team cannot reproduce at zero cost, and evaluations in the area are often of low rigour [13].

The gap addressed here is a reproducible, zero-cost system that (a) integrates multiple tracked metrics in one record, (b) computes every target deterministically from published equations and stated heuristics, (c) constrains the LLM with those targets, and (d) verifies each generated meal plan in code — together with a quantitative, statistical evaluation of both the premise and the mechanism.

V. OBJECTIVES, VARIABLES AND HYPOTHESIS

A. Objectives

This study aims to —

- 1) build a single per-user record for workouts, food, water, sleep, weight, steps and goals, with trackers, a dashboard and progress analytics;
- 2) implement a deterministic health calculator traceable to published equations and validate it against an external dataset;
- 3) generate workout plans, meal plans and coaching answers with an LLM constrained by the computed targets, and verify meal plans in code;
- 4) test, through secondary research, whether the tracked lifestyle metrics are associated with obesity level and whether combining them adds predictive value;
- 5) Deliver the system as a responsive, zero-cost web application with owner-only data security.

B. Variables

Table I lists the variables of the study.

TABLE I
VARIABLES OF THE STUDY

Type	Variable	Measurement
Independent	Tracked lifestyle metrics: FAF, CH2O, FCVC, NCP, TUE, SCC, FAVC, CAEC	Ordinal (1–3, 1–4, 0–3, 0–2) or categorical survey answers
Independent	Feature set (single metric versus eight combined metrics)	Model configuration
Independent	Constraint mechanism (computed targets plus code-side check versus profile only)	Planning condition
Dependent	Obesity level	Seven ordered classes, Insufficient weight to Obesity III
Dependent	Predictive performance	Macro-F1 under cross-validation
Dependent	Plan compliance	Proportion of meal plans within ± 5 % of the calorie target

C. Hypotheses

- 1) H1 (Relevance of the tracked metrics). H0: the lifestyle metrics that AI Fitness OS tracks or addresses — physical activity frequency (FAF), daily water intake (CH2O), vegetable consumption (FCVC), number of main meals (NCP), technology-use time (TUE), calorie monitoring (SCC), frequent high-calorie food (FAVC) and eating between meals (CAEC) — are not associated with obesity level. H1: they are significantly associated.
- 2) H2 (Integration). H0: a model that combines the eight tracked metrics predicts obesity level no better than the best single tracked metric. H1: the combined model predicts significantly better.
- 3) H3 (Verifiability). H0: supplying deterministically computed targets as binding constraints, with a code-side compliance check, does not change the proportion of meal plans that land within ± 5 % of the user's calorie target, compared with planning from the profile alone. H1: it changes (increases) that proportion.

VI. METHODOLOGY

A. Research Design

The study combines descriptive and analytical secondary research with the design-science approach of Hevner et al. [17], in which knowledge is produced by building an artefact and evaluating it rigorously. The secondary research was conducted in October 2026 on existing public records; the unit of analysis is one respondent record. The artefact was built with Flutter 3.47.6, Riverpod 3, Firebase Authentication and Cloud Firestore (region asia-south1, Mumbai), with security rules allowing each user to access only their own data. The artefact was evaluated through an automated test suite (82 tests passing, 3 live tests skipped by design), a 1,000-profile simulation, validation of the calculator against the dataset, and live measurements of Gemini responses. No user study, survey or interview was conducted.

B. Data Collection

The dataset "Estimation of obesity levels based on eating habits and physical condition" by Palechor and de la Hoz Manotas [18] was downloaded from the UCI Machine Learning Repository (id 544, licence CC BY 4.0; it is also mirrored on Kaggle). It contains 2,111 records and 17 attributes describing respondents from Colombia, Peru and Mexico, with no missing values and 24 exact

duplicate rows, which were retained as published. About 23 % of the records came from the authors' web survey; about 77 % were generated synthetically by the authors with SMOTE [19] to balance the seven classes. Because SMOTE produces fractional values, the 515 records whose ordinal answers and age are all whole numbers were extracted as an "original-like" subset for robustness checks. The analysis used Python 3 with pandas, SciPy and scikit-learn [20].

C. Sample Size

The minimum sample size for estimating a proportion was calculated with Cochran's formula [21]:

$$n_0 = Z^2 \times p \times (1 - p) / e^2 = (1.96^2 \times 0.5 \times 0.5) / 0.05^2 = 384.16 \approx 384$$

where $Z = 1.96$ (95 % confidence), $p = 0.5$ (maximum variability) and $e = 0.05$ (5 % margin of error). The dataset's 2,111 records exceed this minimum by a wide margin, and even the 515-record original-like subset exceeds 384. The sample is, however, non-probabilistic: it reflects the dataset authors' web survey and their synthetic augmentation, which limits generalisation.

VII. DATA ANALYSIS AND FINDINGS

A. Description of the Data Set

The target attribute (NObyesdad) has seven classes that are nearly balanced in the full dataset: Insufficient weight 272, Normal 287, Overweight I 290, Overweight II 290, Obesity I 351, Obesity II 297 and Obesity III 324. The sample is evenly split by gender (1,068 male, 1,043 female), and 1,726 records report a family history of overweight. Table II summarises the numerical attributes.

TABLE II
NUMERICAL SUMMARY OF THE DATASET (N = 2,111)

Attribute	Mean	SD	Min	Q1	Median	Q3	Max
Age (years)	24.31	6.35	14.00	19.95	22.78	26.00	61.00
Height (m)	1.70	0.09	1.45	1.63	1.70	1.77	1.98
Weight (kg)	86.59	26.19	39.00	65.47	83.00	107.43	173.00
BMI	29.70	8.01	13.00	24.33	28.72	36.02	50.81
FCVC (1–3)	2.42	0.53	1.00	2.00	2.39	3.00	3.00
NCP (1–4)	2.69	0.78	1.00	2.66	3.00	3.00	4.00
CH2O (1–3)	2.01	0.61	1.00	1.58	2.00	2.48	3.00
FAF (0–3)	1.01	0.85	0.00	0.12	1.00	1.67	3.00
TUE (0–2)	0.66	0.61	0.00	0.00	0.63	1.00	2.00

The records describe young adults (median age 22.78 years) whose mean BMI (29.70) lies at the boundary between overweight and obese. Most records indicate frequent consumption of high-calorie food (1,866 of 2,111) and eating between meals "sometimes" (1,765), while only 96 indicate monitoring of calorie intake. Before testing the hypotheses, the system's calculator was validated against the dataset: applying the app's BMI formula and WHO cut-offs [5] to each record and collapsing the dataset's seven labels into four categories, 2,086 of 2,111 records agreed (98.82 %; 99.42 % on the original-like subset). All 25 disagreements had a BMI between 18.49 and 30.36, that is, at category boundaries.

B. Hypothesis 1

H0: the eight tracked lifestyle metrics are not associated with obesity level. H1: they are significantly associated.

The five ordinal metrics were tested with the Kruskal–Wallis H test [22] (effect size ϵ^2), and the three categorical metrics with Pearson's chi-square test (effect size Cramér's V). Because eight tests were run, a Bonferroni-adjusted significance level of $0.05 / 8 = 0.00625$ was applied. Table III reports the results on the full dataset.

TABLE III
HYPOTHESIS 1 — ASSOCIATION OF EACH TRACKED METRIC WITH OBESITY LEVEL (N = 2,111)

Hypothesis	Test Used	Statistic	p-value	Decision
H1 – FAF	Kruskal–Wallis	H = 88.42 (df 6), $\epsilon^2 = 0.039$	6.45×10^{-17}	Reject H0
H1 – CH2O	Kruskal–Wallis	H = 94.75 (df 6), $\epsilon^2 = 0.042$	3.12×10^{-18}	Reject H0
H1 – FCVC	Kruskal–Wallis	H = 650.06 (df 6), $\epsilon^2 = 0.306$	3.69×10^{-137}	Reject H0
H1 – NCP	Kruskal–Wallis	H = 230.95 (df 6), $\epsilon^2 = 0.107$	4.80×10^{-47}	Reject H0
H1 – TUE	Kruskal–Wallis	H = 45.04 (df 6), $\epsilon^2 = 0.019$	4.60×10^{-8}	Reject H0
H1 – SCC	Chi-square	$\chi^2 = 123.02$ (df 6), V = 0.241	3.77×10^{-24}	Reject H0
H1 – FAVC	Chi-square	$\chi^2 = 233.34$ (df 6), V = 0.332	1.48×10^{-47}	Reject H0
H1 – CAEC	Chi-square	$\chi^2 = 802.98$ (df 18), V = 0.356	7.38×10^{-159}	Reject H0

All eight metrics are significantly associated with obesity level at the Bonferroni-adjusted level; for example, more frequent physical activity goes with a lower obesity level (Spearman $\rho = -0.180$).

As a robustness check, the same tests were repeated on the 515 original-like records. Only CH2O ($p = 0.0022$), NCP ($p = 0.0025$) and CAEC ($p = 1.45 \times 10^{-6}$) remained significant at the adjusted level; FCVC ($p = 0.0153$) was significant only at 0.05, and FAF ($p = 0.111$), TUE ($p = 0.233$), SCC ($p = 0.290$) and FAVC ($p = 0.540$) were not significant. Effect sizes were much smaller on these records (FCVC $\epsilon^2 = 0.019$ compared with 0.306 on the full data), which indicates that synthetic augmentation inflates the associations.

C. Hypothesis 2

H0: a model combining the eight tracked metrics predicts obesity level no better than the best single tracked metric. H1: the combined model predicts significantly better.

A Random Forest [23] was evaluated with stratified 10-fold cross-validation using macro-F1, first on each metric alone and then on all eight together, and the paired fold scores were compared with a one-tailed Wilcoxon signed-rank test [24]. As Fig. 1 shows, the best single metric was vegetable consumption (FCVC, macro-F1 0.263), followed by FAF (0.237) and CH2O (0.225), while SCC (0.069) was the weakest; the combined model reached 0.720 ± 0.039 .

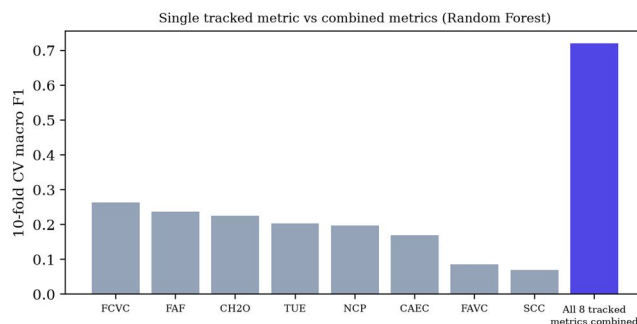


Fig. 1 Single tracked metric versus the combined metrics (H2)

Table IV reports the test.

TABLE IV
HYPOTHESIS 2 — COMBINED METRICS VERSUS THE BEST SINGLE METRIC

Hypothesis	Test Used	Statistic	p-value	Decision
H2	Wilcoxon signed-rank (one-tailed, 10 paired folds)	W = 55; mean gain in macro-F1 = +0.456	0.00098	Reject H0

The combined model was better in every fold; $p = 0.00098$ is the smallest value attainable with ten folds, and a paired t-test agreed ($t = 29.98$, $p = 1.25 \times 10^{-10}$).

D. Hypothesis 3

H0: supplying computed targets as binding constraints, with a code-side check, does not change the proportion of meal plans within $\pm 5\%$ of the calorie target, compared with planning from the profile alone. H1: it changes that proportion.

One thousand user profiles were simulated with a fixed seed (2026), with random age (18–65 years), sex, height (150–195 cm), weight (45–120 kg), activity level, goal and diet preference. For each profile, two meal plans were produced from the same dish library used by the system's offline engine: a constrained plan, following the application's path (calorie target computed, slots allocated as 25 % breakfast, 35 % lunch, 10 % evening snack and 30 % dinner, then checked and rescaled by the $\pm 5\%$ compliance check), and an unconstrained plan using standard portions with no target and no check. The paired outcomes were compared with the exact McNemar test, and Table V reports the result.

TABLE V
HYPOTHESIS 3 — MEAL PLANS WITHIN $\pm 5\%$ OF THE CALORIE TARGET (1,000 PROFILES)

Hypothesis	Test Used	Statistic	p-value	Decision
H3	Exact McNemar (paired)	Within $\pm 5\%$: constrained 1,000 / 1,000, unconstrained 75 / 1,000; discordant pairs $b = 925$, $c = 0$	$< 10^{-200}$	Reject H0

Constrained plans deviated from target by a mean absolute 0.05 % (maximum 3.19 %), against 29.32 % (maximum 60.48 %) for unconstrained plans.

The mechanism was corroborated with the live model. For a test profile (male, 25 years, 70 kg, 175 cm, moderately active, weight loss, vegetarian), Gemini returned a four-meal plan totalling 2,094 of the 2,094 kcal target before the code check; for a second, eggetarian profile it returned 2,151 of 2,151 kcal, with protein, carbohydrate and fat also exactly on target. Only these two live meal plans were measured, because AI usage was deliberately kept minimal within the free tier.

E. Supporting Evaluation of the Artefact

Two further analyses complement the hypotheses. First, four classifiers were compared on the lifestyle attributes only. Height and weight were excluded because the target is defined from BMI; including them would be target leakage (a Random Forest with height and weight reaches 0.920 accuracy). The data were split 70/30 with stratification (1,477 training and 634 test records) and also assessed with stratified 5-fold cross-validation. Table VI and Fig. 2 present the results.

TABLE VI
COMPARISON OF FOUR CLASSIFIERS ON LIFESTYLE ATTRIBUTES (TEST SET AND 5-FOLD CROSS-VALIDATION)

Model	Accuracy	Macro-F1	ROC-AUC	CV accuracy	CV macro-F1
Logistic Regression	0.621	0.587	0.876	0.607 ± 0.025	0.580 ± 0.027
Decision Tree	0.757	0.756	0.857	0.756 ± 0.016	0.750 ± 0.015
Random Forest (300 trees)	0.820	0.820	0.970	0.849 ± 0.011	0.847 ± 0.011
SVM (RBF kernel)	0.732	0.724	0.944	0.768 ± 0.014	0.760 ± 0.015

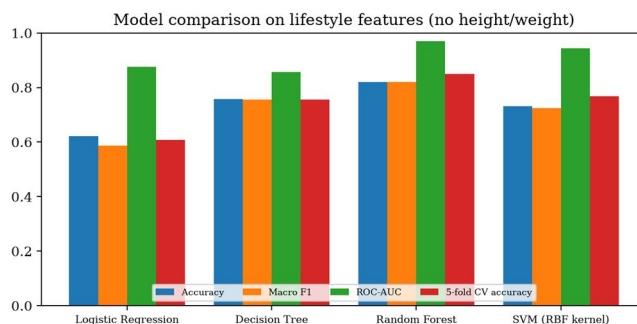


Fig. 2 Comparative performance of the four classifiers

The Random Forest performed best, far above the majority-class baseline accuracy of 0.167; age, FCVC, NCP, FAF and CH2O were its most important attributes. On the 515 original-like records, however, the same model reached a macro-F1 of only 0.278, reflecting severe class imbalance and the absence of synthetic smoothing.

Second, the operational behaviour of the built system was measured. With the earlier configuration (gemini-3.5-flash with low thinking), a workout plan took 31.5 s and a meal-plan request failed because the service reported high demand; that model's free tier also allowed only 20 requests per day. With the adopted configuration (gemini-3.5-flash-lite with minimal thinking), responses took 2.5–6.1 s: chat and daily tip 2.5 s, healthy swaps 3.4 s, meal plan 4.1 s and a seven-day workout plan 6.1 s. The Firebase project runs on the Spark plan with no billing account, so no charge is possible; when the free daily AI allowance is used up, the application shows a plain-language message and saves nothing.

VIII. RESULTS

- 1) H1 is supported on the published dataset, and only partially robust on original-like records. All eight tracked lifestyle metrics are significantly associated with obesity level after Bonferroni adjustment, which justifies tracking them in one record. On the 515 records closest to the real survey, three of the eight remain significant at the adjusted level (four of eight at 0.05), and effect sizes are much smaller, so the strength of the associations on the full data is partly an artefact of synthetic augmentation.
- 2) H2 is supported. Combining the eight metrics raised macro-F1 from 0.263 to 0.720, a gain of 0.456 that held in every fold. No single behaviour explains weight status well; the information lies in the combination, which supports, but does not prove, the value of a unified per-user record.
- 3) H3 is supported, strongly for the mechanism. Constraining generation with computed targets and checking the result in code moved compliance from 75 of 1,000 plans to 1,000 of 1,000. The two live Gemini plans met their targets exactly even before the check, although two cases are too few to estimate how often this happens; the code-side check provides the guarantee. The live sample is small, and the check guarantees the calorie total, not micronutrient or wider nutritional quality.

The calculator matches an external labelling for 98.82 % of records, with disagreements only at category boundaries, and the system responds in 2.5–6.1 s at zero cost.

IX. RECOMMENDATIONS

- 1) Compute, generate, check. Developers of AI health applications should compute every numerical target in code from published equations or explicitly stated rules, give those numbers to the language model as constraints, and verify the output in code before it is shown or saved; this converts an average error [10] into a bounded one.
- 2) Integrate the metrics. Since the combined metrics predicted weight status far better than any single one, fitness tools should hold activity, diet, hydration, sleep and weight in one record and let recommendations draw on all of them.
- 3) Prefer real survey data. Researchers using this dataset should report results on the original-like records alongside the full data, since synthetic augmentation inflates effect sizes.
- 4) Choose fast, free model settings deliberately. For interactive coaching, a lighter model with minimal reasoning was several times faster than the earlier configuration in this study (6.1 s against 31.5 s for a workout plan) and stayed within a separate free allowance.
- 5) Extend the evaluation. Future work should include a user study of adherence and usefulness, a larger sample of live LLM plans, micronutrient checks, revision of the water heuristic (its 35 mL/kg base rate is a plain-water heuristic, not directly comparable with the 40–45 mL/kg median total water intake reported by Stookey et al. [16]), and features not yet built, such as wearable synchronisation, push reminders and camera-based posture feedback.

X. CONCLUSION

This study presented AI Fitness OS, a zero-cost web application that unifies a user's training, nutrition, hydration, sleep, weight and step data, computes all targets deterministically from published equations and stated heuristics, and constrains Google Gemini with those targets while verifying each meal plan in code. Secondary research on 2,111 records showed that the metrics the system tracks or addresses are associated with obesity level and that, combined, they predict it far better than any single metric; a 1,000-profile simulation showed that the constraint-and-check mechanism places every meal plan within ± 5 % of its calorie target, against 7.5 % without it. The conclusions are bounded by honest limitations: about 77 % of the dataset is synthetic, the respondents come from three Latin-American countries and answered by self-report, only two live LLM meal plans were measured, and no study with users was conducted. Within those bounds, the work shows that a large language model can be made verifiable in a health setting by a simple rule — arithmetic is computed, language is generated, and the generated output is checked against the computed numbers.

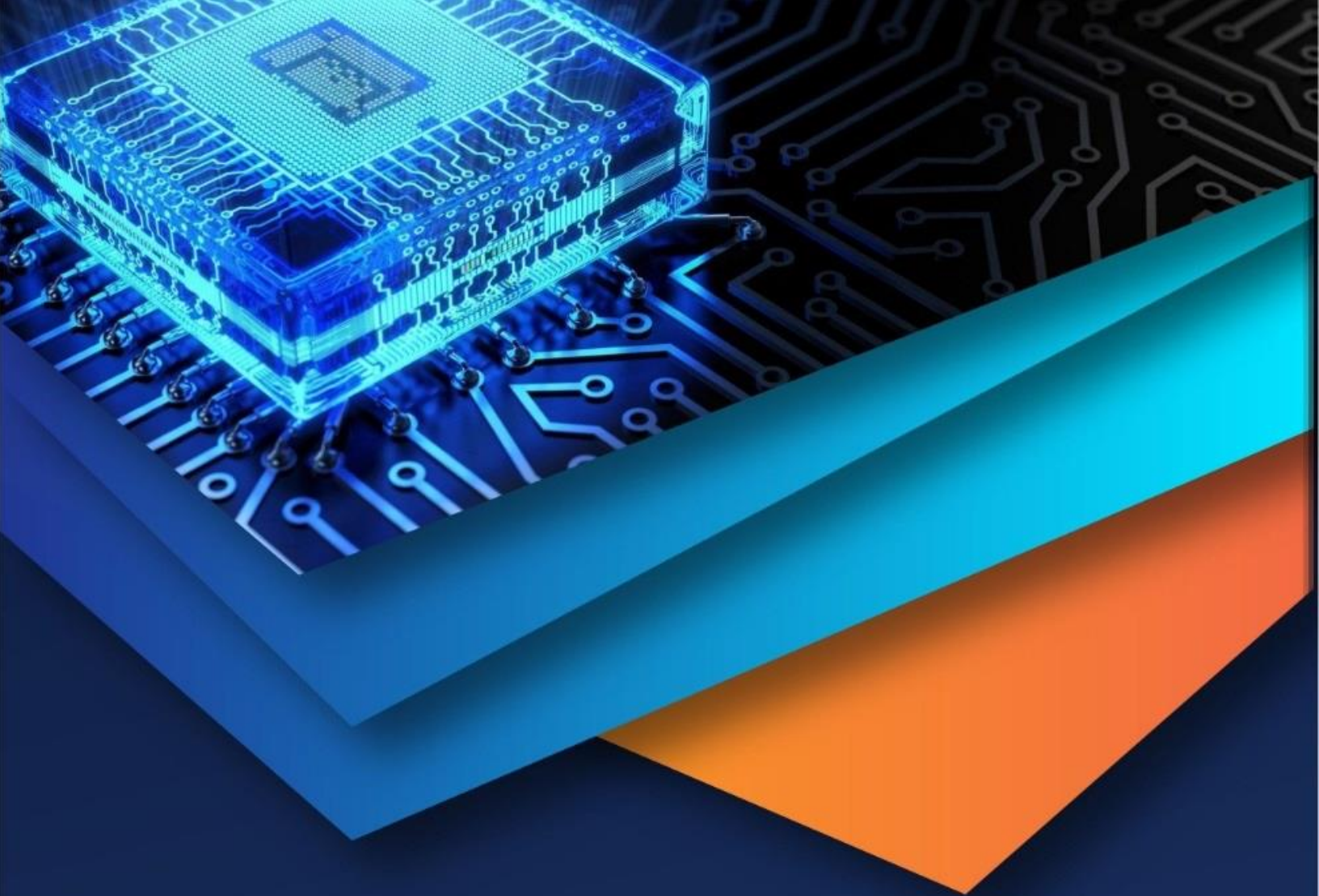
XI. ACKNOWLEDGMENT

The authors thank [Project Guide Name] and the Department of Information Technology, Wilson College (Autonomous), Mumbai, for their guidance and support.

REFERENCES

- [1] Z. He, M. A. Hassan, P. Saiz-González, S. Ryu, R. Wang, and Z. Gao, "Smartphone app-based interventions on physical activity behaviors and psychological correlates in healthy young adults: A systematic review," *PLOS ONE*, vol. 19, no. 4, p. e0301088, 2024.
- [2] H. Lin, M. Yang, Z. Zhou, Y. Zhang, and N. Deng, "Exploring the dynamics of dietary self-monitoring adherence among participants in a digital behavioral weight loss program: Model development study," *J. Med. Internet Res.*, vol. 27, p. e65431, 2025.

- [3] D. Shin, G. Hsieh, and Y.-H. Kim, "PlanFitting: Personalized exercise planning with large language model-driven conversational agent," arXiv:2309.12555, 2023.
- [4] Z. Yang et al., "ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework," arXiv:2403.00781, 2024.
- [5] World Health Organization, "Obesity: Preventing and managing the global epidemic," WHO Technical Report Series 894, Geneva, 2000.
- [6] M. D. Mifflin, S. T. St Jeor, L. A. Hill, B. J. Scott, S. A. Daugherty, and Y. O. Koh, "A new predictive equation for resting energy expenditure in healthy individuals," *Am. J. Clin. Nutr.*, vol. 51, no. 2, pp. 241–247, 1990.
- [7] M. Jörke et al., "GPTCoach: Towards LLM-based physical activity coaching," arXiv:2405.06061, 2024.
- [8] J. Cosentino, A. Belyaeva, X. Liu, and N. A. Furlotte, "Towards a personal health large language model," arXiv:2406.06474, 2024.
- [9] A. A. Heydari, K. Gu, V. Srinivas, and H. Yu, "The anatomy of a personal health agent," arXiv:2508.20148, 2025.
- [10] S. Khamesian, A. Arefeen, S. M. Carpenter, and H. Ghasemzadeh, "NutriGen: Personalized meal plan generator leveraging large language models to enhance dietary and nutritional adherence," arXiv:2502.20601, 2025.
- [11] F. Gao et al., "HealthGenie: Empowering users with healthy dietary guidance through knowledge graph and large language models," arXiv:2504.14594, 2025.
- [12] I. Azimi, M. Qi, L. Wang, A. M. Rahmani, and Y. Li, "Accuracy and consistency of LLMs in the registered dietitian exam: The impact of prompt engineering and knowledge retrieval," arXiv:2408.02964, 2024.
- [13] X. Lai, Y. Lai, J. Chen, S. Huang, Q. Gao, and C. Huang, "Evaluation strategies for large language model-based models in exercise and health coaching: Scoping review," *J. Med. Internet Res.*, vol. 27, p. e79217, 2025.
- [14] Y. Sun, H. Dong, and W. Jiang, "Is more always better? An S-shaped impact of gamification feature richness on exercise adherence intention," *Front. Psychol.*, vol. 16, art. 1671543, 2025.
- [15] J. W. Navalta, J. D. Thomas, and W. J. Stone, "Demographic profiles and methodologies used in the generation and validation of resting metabolic rate prediction equations: Protocol for a systematic review," *JMIR Res. Protoc.*, vol. 15, p. e82482, 2026.
- [16] J. D. Stookey et al., "Hydration and health at ages 40–70 years in Salzburg Austria is associated with a median total water intake over 40 mL/kg including at least 1 L/d plain drinking water," *Front. Public Health*, vol. 13, art. 1668981, 2025.
- [17] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
- [18] F. M. Palechor and A. de la Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data in Brief*, vol. 25, art. 104344, 2019.
- [19] 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.
- [20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [21] W. G. Cochran, *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.
- [22] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Statist. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.
- [23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [24] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)