Detecting Change in Engineering Interest in Children through Machine Learning using Biometric Signals

1st Gustavo Olivas
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
gustavo.olivas@tec.mx

2nd Milton Candela
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
milton.candela@tec.mx

3rd Juan Ocampo
School of Medicine and Health Sciences
Tecnologico de Monterrey
Monterrey, Mexico
juancarlos.ocampo@tec.mx

4th Luis Acosta
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
luis.acosta@tec.mx

5th Aimé Aguilar
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
a00828691@itesm.mx

6th Esther Delgado
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
a00827948@itesm.mx

7th Myriam Alanis
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
myriam.alanis@tec.mx

8th Mauricio Ramírez
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
mauricio.ramirezm@tec.mx

9th Jorge Lozoya
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
jorge.lozoya@tec.mx

10th Jorge Murrieta
MachineCare Education
Monterrey, Mexico
jorge.murrieta@machinecare.com.mx

11th Ricardo Ramírez
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, Mexico
ricardo.ramirez@tec.mx

***Abstract*—Promoting interest in Science, Technology, Engineering, and Mathematics (STEM) is a worldwide priority. Thus, accurate measures of professional interest, an aspect of personality according to vocational theories, are necessary. There is empirical evidence supporting personality prediction using biometric signals. Therefore, professional interest may be estimated similarly. The objective of this study was to generate a machine learning algorithm based on physiological data of children performing engineering-related activities that estimated their professional interest in engineering subfields. Thirteen children between 6 and 15 years old participated. Using eight electroencephalographic channels, we measured electrodermal activity, heart rate variability, facial gestures, and body temperature in four 2-hour sessions as the children performed engineering-related educational activities. Psychometric tests evaluated their interest in specific engineering subfields per the activities. We processed the generated data to design a machine learning algorithm, which resulted in 80% precision in detecting the change in interest. The results indicate that a pattern for change in engineering interest exists and can be measured.**

Keywords — professional interest, children, machine learning, biometric signals, educational innovation

I. Introduction

Promoting Science, Technology, Engineering, and Mathematics (STEM) interest among youth is a priority in many nations worldwide [1], evidenced by specific public and private policies and initiatives [2]. Psychometric instruments are the habitual, most utilized way to measure STEM interest. Unfortunately, these measurements are vulnerable to cultural differences and response biases [3]. Research shows a close theoretical and empirical relationship between professional interest and personality [4]. Moreover, in the past three decades, physiological data has successfully estimated psychological variables, including personality traits [5] [6] [7] [8]. Thus, we argue that it is possible to assess professional interest through biometric means.

This endeavor follows current worldwide interest in creating optimal opportunities for individuals to flourish, yielding positive results for our societies [9]. Passion is defined as a strong inclination towards specific, personally gratifying, motivating, and meaningful activities. Evidence has shown its connection to professional success, self-realization, life satisfaction, and other positive outcomes [10]. "Interest," sometimes used as an interchangeable term with "passion," is better understood as an early stage before it matures into a passion [11]. In this work, we propose our study of interest as a simplified window into the concept of passion. Moreover, while children develop their passions through exploration, their interests are an ideal approximation that is readily measurable. Suppose passion in its germinal form can be estimated from a young age. In that case, it can be appropriately nurtured and directed to achieve long-term professional, academic, and personal goals.

II. Objectives

The general objective of this study was to generate a machine learning algorithm based on physiological data of children doing STEM-related activities that estimates their professional interest in engineering subfields. The specific objectives were:

1. To evaluate the participants' interest in engineering through psychometric evaluation.

2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop | 978-1-6654-2763-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53024.2021.9733772

2. To record the participants' physiological responses while engaged in engineering-related tasks through electroencephalography (EEG), electrodermal activity (EDA), heart rate, heart rate variability (HRV), body temperature, and emotion detection.
3. To prepare data using preprocessing techniques (filtering, artifact removal) and analyze the data using a Machine Learning framework incorporating Python libraries. For more information, refer to the "Methodology section."
4. To contrast the participants' change in engineering interest with their physiological response when engaged in engineering-related tasks.

III. Literature Review

Literature reviews to date concur that Holland's vocational model is the most researched theoretical foundation for professional interest [12]. This model proposes occupational themes based on personality typologies; it presupposes that professional interest is an expression or aspect of personality [5]. Initially conceived for adult research, the model has been widely used to study children's interest development [13]. Additionally, increasing evidence showing significant personality change throughout the lifespan contests the hypothesis that personality consolidates during adulthood, which may be a concern when researching interests during childhood [14].

Furthermore, considerable advances have been made in the last three decades on personality prediction using biomedical methods such as EEG [6, 7]. Although without a decisive conclusion, one review states that some studies successfully differentiated extroverts from introverts using EEG [8]. Another review concludes that event-related potentials (ERP) during cognitive tasks are a functional technique for estimating personality traits [15]. In [16], the authors report a connection between EEG characteristics and the elements of the 16 personality traits theory. Finally, EEG measurements can be helpful when predicting extraversion, agreeableness, openness, conscientiousness, and neuroticism [17].

Neuroscientific research methodologies can provide new measurement paradigms for professional interest. Despite this, we only found one article using these techniques for the stated purpose [18]. This study evaluated EEG signals as predictors of vocational interests in a psychometric instrument applied to university students, obtaining initial results, and concluding that further research is needed. So, there is a knowledge gap and inconclusive evidence regarding the potential of biometric signals to detect professional interest.

IV. Method

A. Participants

Thirteen children from 6 to 15 years old participated; the mean age was 11 with a standard deviation of 2 years. The sample was composed of 11 males and 2 females. All participants were enrolled in three sequential courses on 3D Design, Programming, and Robotics offered by MachineCare Education, an extracurricular educational company based in Monterrey, Mexico.

B. Ethical considerations

Participants and their parents or legal guardians signed an informed assent and consent form, respectively. Both children and adults were clearly informed about the study, their rights, risks, and benefits and remained in close contact with the research team.

V. Data Collection

A. Methodology

Several devices obtained biometric signals. Skin temperature, EDA, and blood volume pulse (BVP) were acquired using the Empatica E4 bracelet [19]. EEG signals were received using the LiveAmp 8 device [20], which has eight active gel electrodes arranged per the international 10-20 system (FP1, FP2, FC3, FC4, C3, C4, Pz). Facial gestures were recorded using a stereo RGB Intel d435i camera [21]. Finally, a psychometric scale was administered to measure interest in specific engineering topics. Due to the data quantity available, the project's methodology contained six steps, summarized in Figure 1.

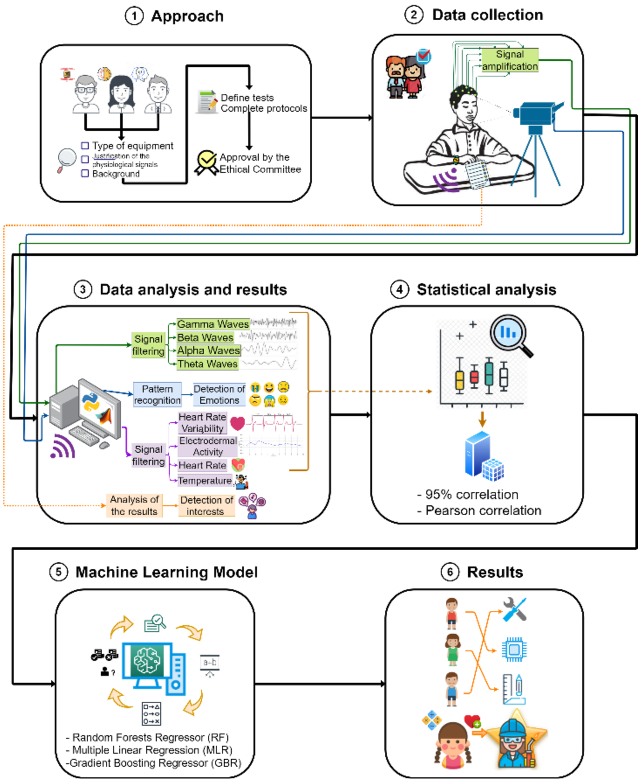

Figure 1. Project's Methodology

1. Approach: We formed a team with members from multiple engineering and social areas, such as mechatronics, biomedicine, data science, and psychology, to design an acquisition system, implement the psychometric tests, and analyze all received data.

2. Data collection: After receiving the legal tutor's authorization, we obtained the following physiological signals from each subject: skin temperature, brain waves (8 EEG channels), EDA, and BVP at 4, 250, 4, and 64 Hz, respectively.

    a. Participants completed an adapted version of the engineering subscale of the STEM-CIS before and after each class. STEM-CIS is a psychometric questionnaire consisting of 11 items measuring interest in STEM subjects [1], and it has been administered to middle school and elementary school children [22] [23]. Two adaptations were made. First, the words "engineering" and "engineer" in each item were substituted for the subfields being studied (e.g., "I can do well in activities that involve engineering" modified as "I can do well in activities that involve 3D design/programming/robotics"). Second, the questionnaire was translated into Spanish as there is no Spanish version yet. Reliability and validity analyses indicated adequate results except for the last item ("I know someone in my family who is a 3D designer/programmer/robotics professional"), which was excluded from the subsequent evaluation.

3. Data analysis and results: the signals were analyzed as indicated below.

    a. Brain waves: Filtered by an initial Butterworth bandpass filter of order 7 between 1 and 100 Hz; a notch filter was also used, centered at 60 Hz. These filters removed powerline-related noise and drift. Then an independent component analysis (ICA) was performed to remove muscular and ocular movement artifacts. Finally, an adaptive filter was applied to obtain the different frequency bands sought: theta (4-8 Hz), alpha (8-13 Hz), low beta (13-21 Hz), high beta (21-30 Hz), and gamma (30-80 Hz). The adaptive filter had two fixed values: a bandpass ripple of 7 dB and a stopband attenuation of 10 dB.

    b. Electrodermal activity: Phasic peaks where the signal increased its value over a pre-defined threshold were marked, denoting the presence of phasic activation, which is triggered by external stimuli. These peaks are identified by the zeroes coming from the second derivative of the original EDA signal [24] used by Python's library Neurokit2 [25]. The threshold, or baseline, is determined by the average tonic level registered while the user is under rest conditions in the absence of external stimuli [26].

    c. Blood volume pulse: HRV metrics in time and frequency were extracted using HeartPy [27] and Neurokit2 [25] Python libraries.

    d. Skin temperature: Used to confirm that the user had the Empatica smartband firmly placed. If so, the temperature remains constant; otherwise, it indicates that the user removed the equipment.

    e. Video: Python's DeepFace library [28] analyzed each subject's facial gestures to determine an emotion (angry, disgust, fear, happy, sad, surprise) or the lack of a particular feeling (neutral). These emotions were determined every 10 seconds using the lightweight dlib C++ backend detector with a wrapper for Python. This backend detector has a reported 99.38% of accuracy for face recognition using the commonly known "faces in the wild (LFW)" test [29].

4. Statistical analysis: All features having a 95% of correlation were removed; a Pearson correlation parameter was also performed. These steps and the Machine Learning analysis are discussed further in the next section.

5. Machine Learning Model: The data's distribution guided the selection of the best model for the project. Details are discussed in further sections of this paper.

6. Results: The generated results were critically analyzed to discard any bias for a determined result. They are also discussed in the following section.

B. Statistical and Machine Learning analysis

Afterwards, the data was processed for Machine Learning (ML) purposes. Variables were classified into source and target.

- Source variables: Data used to predict target variables. In this study, biometric data is provided by EEG, Empatica, and the Artificial Vision (AV) algorithm that estimated emotions.
- Target variables: Data estimated using source variables. This study predicted one target variable: change in engineering subfield interest (provided by psychometric testing.)

Regarding biometric data, EEG and Empatica outputs had continuous numerical variables. So, their features' domain is diverse, while the AV algorithm produced a categorical nominal variable regarding the emotion detected. Each feature was transformed into a continuous variable with a domain between 0 and 1, thus $0 \leq x \leq 1$.

We applied a standard scaler to transform the values of EEG and Empatica continuous numerical data; so, we had a mean equal to 0 and standard deviation equal to 1, based on the following equation:

$$x(t)_{(norm),F,S} = \frac{x(t)_{F,S} - \mu_{F,S}}{\sigma_{F,S}} \tag{1}$$

where $x(t)$ is the non-normalized value, μ its mean, and σ the standard deviation of each feature F and student S. The sigmoid function, which ranges between 0 and 1, was further applied to each normalized value. Thus, for each $x(t)_{norm}$ value, a new $x(t)_{sigmoid}$ value between 0 and 1 is returned. Therefore, the transformation is shown in the next equation:

$$x(t)_{sigmoid} = \frac{1}{1 + e^{-x(t)_{norm}}} \tag{2}$$

The previous pipeline was followed for Empatica and EEG features. However, a set of combined features were first created before normalization by iterating over each feature (not combining features from different devices). The combined features were calculated as follows:

- F_i -I: Inverse on i^th feature ($\frac{1}{x(t)_{Fi}}$)
- F_i -L: Natural logarithm on i^th feature ($ln\, ln\, (x(t)_{Fi} + \varepsilon)$)
- F_i -M- Fj: Multiplication of i^th feature with j^th feature ($x(t)_{Fi} \times x(t)_{Fj}$)
- F_i -D- Fj: Division of i^th feature with j^th feature ($\frac{x(t)_{Fi}}{x(t)_{Fj+\varepsilon}}$)

It can be noted that a small number, epsilon $\varepsilon = 1 \times 10^{-6}$, was used to avoid errors in 0-sensitive functions: natural logarithm and division by 0. Moreover, a separate list of combinations was used to track the combinations between the i^th and j^th features and not generate duplicate features when multiplying i^th feature with j^th feature and vice versa (as they produce the same number).

Moving on to the AV algorithm, we created an emotions' probability of distribution based on a 5-minute window. Additionally, we used an adapted version of Laplace's rule of succession transformation for probabilities, which removes probability 0 on all events. The adapted version of this transformation is shown below:

$$p_{e,w} = \frac{f_e + 1}{n_w + m} \tag{3}$$

The previous equation is derived from a fundamental equation of probability, where the likelihood of an event p is given by the frequency f of an event in a specific window of events, although the frequency is summed by 1 on the numerator and by the possible events m in the denominator. f_e is the total number of occurrences of a given emotion in the 5-minute windows, and n_w is the total number of emotions detected in the window, while m is the number of emotions available (6). This transformation always returns a probability different from 0, removing sparsity when the AV algorithm does not detect a wide range of emotions in a given window, mainly due to error rather than the person not displaying any emotion.

VI. Results

A. *Numerical results*

The mean results of the psychometric test on Programming, Robotics, and 3D Design before the course were 3.04, 3.17, and 3.23, respectively, where 0 means no interest and 4 denotes a very high interest. After the course, the new means were 3.22, 3.46, and 3.11 in the same order. The results show a high to very high interest in all STEM subdimensions before and after the intervention. Furthermore, the tests taken after the intervention report a higher interest in all subdomains except 3D Design. The high interest may relate to sampling bias, as the examined sample is intentionally enrolled in extracurricular STEM training. Thus, we may infer there was prior interest. Additionally, if we were to isolate the stimuli, we may infer that the measured before and after a change is partly due to the educational activities.

Further analysis showed that children aged 6 to 11 had a higher mean interest before the intervention in all subdomains than children aged 12 to 15. This finding holds after the intervention except in 3D Design. A probable cause may be that younger children are prone to extreme response bias, meaning some participants report extreme answers. For example, younger children might answer with extremity responses in the Likert scale much often than their older counterparts. Another conjecture concerns the changes in surprise as an emotion throughout the age range [30]. The mean interest also increases when segmented by age group on all occasions, except 3D Design, supporting the possibility that the intervention may increase interest in Programming and Robotics.

We removed features that correlated by more than 95% on the Machine Learning side, which reduced the number of features from 10799 to 898. Then, a feature selection criterion known as the Mean Decrease Impurity (MDI), computed by a Random Forest Regression algorithm via the GINI index, was used for further feature reduction. The algorithm is randomized; thus, a total of 20 iterations were performed to generalize on multiple seeds and obtain unbiased results. The top 20 features with the greatest MDI were chosen.

The previous process was followed to extract the best features for the target variable (the change in STEM interest). "Best features" had a final feature selection process using the Pearson Correlation Coefficient (r) with the target variable, where the p-value determined if the correlation between a feature and the target variable was statistically significant. We used a significance level of 5%, so accepted features must have a p-value < 0.05. This criterion removed the last 3 non-significant features for change in STEM interest, leaving 17 significant features (EEG = 15, AV = 1, Empatica = 1) shown in Table I, highlighted with bold letters.

The feature with the highest r was *surprise* (0.3491), which positively correlated with the change of STEM interest, meaning that evoking surprise in a child during a STEM lecture might increase their STEM interest. On the other hand, *fatigue* (-0.2322) negatively correlated with the target variable, meaning that high fatigue during a STEM lecture might decrease a child's STEM interest. The examples above are two valuable insights that could be easily applied to a future investigation to improve education based on biometric data.

TABLE I. Correlation between change in STEM interest and the 20 features with highest MDI

Feature	r	p-value	Device
Surprise	0.3491	**0×10^{-4}**	AV
Alpha_F4-D-Alpha_FP1	0.3206	**0×10^{-4}**	EEG
LowBeta_F4-M-High-Beta_PZ	0.3180	**0×10^{-4}**	EEG
Fatigue-D-Alpha_F4	-0.2692	**0×10^{-4}**	EEG
Fatigue	-0.2322	**0×10^{-4}**	EEG
LowBeta_PZ-D-Fatigue	-0.2304	**0×10^{-4}**	EEG
Alpha_C4-D-Alpha_PZ	-0.2196	**0×10^{-4}**	EEG
Fatigue-D-Alpha_C4	-0.2092	**0×10^{-4}**	EEG
Alpha_F4-D-Fatigue	0.2066	**0×10^{-4}**	EEG
Alpha_PZ-M-Alpha_F4	0.1929	**0×10^{-4}**	EEG
Gamma_F3-M-Fatigue	0.1714	**0×10^{-4}**	EEG

Alpha_PZ-I	-0.1706	**0 x 10^{-4}**	EEG
Alpha_C4-M-Load	-0.1231	**0.0021**	EEG
HRV-MedianNN	-0.1077	**0.0072**	Empatica
Alpha_PZ-M-LowBeta_FP1	-0.1038	**0.0096**	EEG
Alpha_PZ-M-LowBeta_C3	-0.0867	**0.0306**	EEG
Fatigue-D-Load	-0.0856	**0.0328**	EEG
Theta_C4-D-Theta_C3	0.0359	0.3717	EEG
Alpha_PZ-D-Alpha_C4	0.0230	0.5666	EEG
HRV_MeanNN-D-HRV_MedianNN	0.0054	0.8920	Empatica

B. *Graphical results*

Figure 2 displays the statistically significant features in a bar plot, where the height shows the linear correlation between the feature and the change in STEM interest. It should be noted that EEG is the biometric device from which the most significant features are drawn (15); however, this device had fewer features than the Empatica device (38 and 64 features before generation). On the other hand, the *probability of surprise* and *HRV_MeanNN* are two non-EEG features present in both target variables.

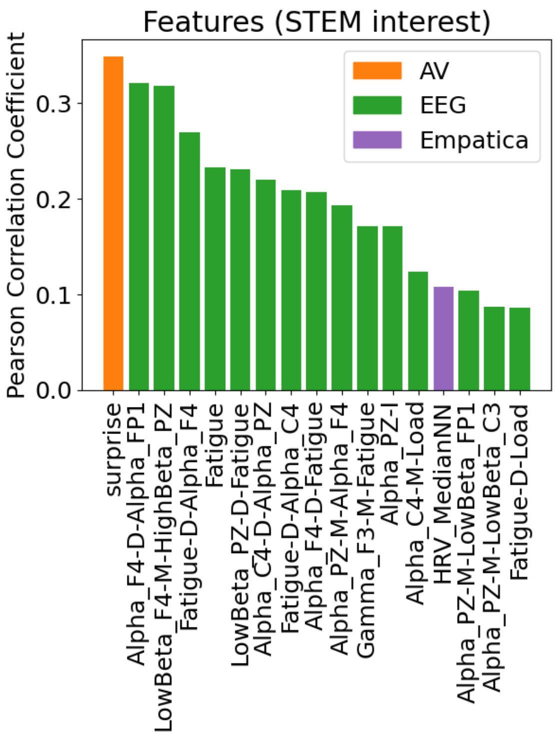

Figure 2. Statistically accepted features ($p < 0.05$) and their correlation with the target variable.

C. *Proposed improvements*

In addition to the best-accepted features for the target variable, the model employed additional one-hot encoded categorical nominal features that indicated the lecture's subject (Robotics, Programming, 3D Design). Given that four lectures were used, this generated 4 new features (17 previous + 4 information features = 21 total features). However, it lowered the coefficient of determination (R^2) by 0.0859, using Multiple Linear Regression (MLR) as an algorithm. This may suggest that the lecture's subject does not significantly predict the change in STEM interest, which might be related to the fact that a standardized test was applied to determine this metric.

The relatively small sample size (13) could be considered a limitation in this study. Additional improvements that could be done to increase the model's performance include adding participants to the sample to increase the data available for model training. However, this presented work proposes the methodology as a proof-of-concept tool rather than a study generalizable to the population. In this sense, the sample size is adequate, considering the similar (and smaller) sample size in works that present similar neuro-engineering tools to identify mental fatigue (8) [31] and drowsiness (10) [32]. Moreover, this study used only the median value in a 5-minute window (EEG and Empatica features). However, more features may be generated using additional statistical and mathematical techniques, such as mean, mode, standard deviation, maximum, minimum, interquartile range, skewness, kurtosis, and others.

D. *Validation*

Using the most significant features, we fitted three regression models:

- Random Forest Regressor (RF)
- Multiple Linear Regression (MLR)
- Gradient Boosting Regressor (GBR)

The validation methodology used to measure a model's performance was Leave-One-Out (LOO), using lectures. This methodology uses all but one lecture to train a Machine Learning model, while the removed lecture is used for testing. Thus, for each student and lecture, models were trained to predict the score on the removed lecture samples. Afterwards, the predicted scores mean was calculated to get the overall lecture score, saved in an array for later comparison. A student could have multiple lectures, so when a lecture was used for testing, the others were removed from the training dataset, which reduced bias caused by using lectures from the same student in the training and testing datasets.

The array with predicted scores was used to evaluate each model's performance via the coefficient of determination (R^2), using the ground truth. Afterwards, the numerical prediction was encoded into an ordinal categorical feature. The encoding was employed to calculate an additional performance metric: accuracy, defined as the number of correctly predicted labels divided by the total number of samples. The best model was selected based on the combination of both numerical and categorical performance metrics.

Results based on the evaluation metrics are displayed in Table II. Using the three models (MLR, RF, GBR), *best model* has the highest coefficient of determination and accuracy when predicting the change in STEM interest, which was the RF model ($R^2 = 0.4504$, accuracy = 80.56 %). The MLR model performed worst ($R^2 = 0.0457$, accuracy = 66.67 %), while GBR performed similarly to the RF model ($R^2 = 0.4264$, accuracy = 75.00 %). Ensemble learning methods that combine multiple decision trees (GBR, RF) performed best, and linear-based methods, such as MLR, performed quite poorly in explaining the target variable, with a coefficient of determination of 0.0457.

TABLE II. PERFORMANCE METRICS FOR CHANGE IN STEM INTEREST

Model	R^2	***Accuracy (%)***
MLR	0.0457	66.67
GBR	0.4264	75.00
RF	0.4504	80.56

The detailed predicted scores and ground truth are further represented in Figure 3. We can observe that the LOO validation does not perform well in predicting extreme scores or the maximum and minimum values from the current population. This was a limitation of the data due to class imbalance, as some participants performed excellently while others quite poorly. However, the validation scheme accurately predicted when the score of a given lecture was near to the population's mean, as more similar data could be used to generate the given prediction.

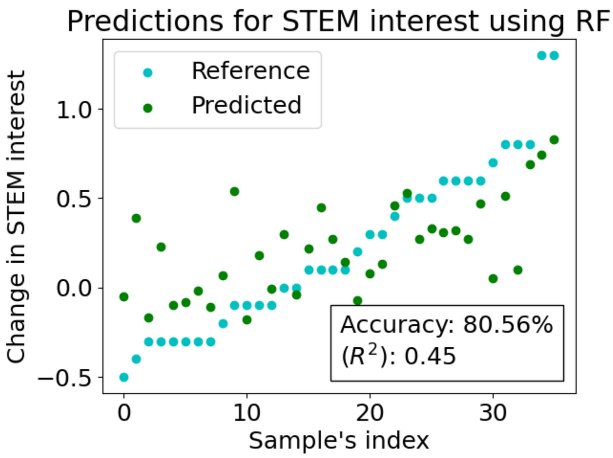

Figure 3. Predictions and reference using the best model (RF) when predicting change in STEM interest.

VII. CONCLUSIONS

In summary, the present investigation focuses on proposing a new methodology in which engineering interest change can be detected and analyzed using machine learning algorithms. This proposal builds upon previous research that demonstrates the relationship between physiological parameters and cognitive variables that include engagement indices, cognitive load, and the stress that each subject may experience. The goal is to refine the present detection system and create a path for new investigation frontiers which can be focused on an engineering context or expanded to broader professional interest detection.

Furthermore, this investigation highlights the relevance and importance of analyzing emotions while learning. As shown by the ML findings, surprise, as an emotion, is a crucial factor in detecting a change in interest in the subjects studied. This finding is supported by psychological studies such as Vogl et al., that surprise can lead the subjects to develop interest and internal motivation to continue exploring their areas of interest [33].

Ultimately, this instrument can be used for individual purposes and for private and public educational institutions and companies. The significant potential in these novel measurement methodologies lies in the promptness of assessing interest without the response biases inherently present in psychometric tests. This may profoundly impact a country's economic development: it could focus its resources to prepare well-trained professionals to work in sectors where they will thrive due to their capabilities. Moreover, neuroscience as a tool for education makes intelligent data-based decision-making possible. However, because neuroscience in this context lacks the adequate instruments to impact education [34], the tool we propose can create a bridge between neuroscience and education.

VIII. ACKNOWLEDGMENT

The authors would like to acknowledge the technical and financial support of Writing Lab, Institute for the Future of Education, Tecnologico de Monterrey, Mexico, in the production of this work.