IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2021, accepted September 29, 2021, date of publication October 1, 2021, date of current version October 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117120

User Stress in Artificial Intelligence: Modeling in Case of System Failure

OLGA VL. BITKINA[1], JUNGYOON KIM[2], JANGWOON PARK[3], JAEHYUN PARK[1], AND HYUN K. KIM[4]

[1]Department of Industrial and Management Engineering, Incheon National University (INU), Incheon 406772, South Korea
[2]Department of Computer Science, Kent State University, Kent, OH 44240, USA
[3]Department of Engineering, Texas A&M University—Corpus Christi, Corpus Christi, TX 78412, USA
[4]School of Information Convergence, Kwangwoon University, Seoul 01897, South Korea

Corresponding authors: Jaehyun Park (jaehpark@inu.ac.kr) and Hyun K. Kim (hyunkkim@kw.ac.kr)

This work was supported in part by the Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea (NRF), and in part by the Unmanned Vehicle Advanced Research Center (UVARC) by the Ministry of Science and ICT, Republic of Korea, under Grant 2020M3C1C1A01084900. Also, the present research has been conducted by the Research Grant of Kwangwoon University in 2021.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Texas A&M University Corpus Christi Institutional Review Board under Protocol No. 59-17.

ABSTRACT The uninterrupted operation of systems with artificial intelligence (AI) ensures high productivity and accuracy of the tasks performed. The physiological state of AI operators indicates a relationship with an AI system failure event and can be measured through electrodermal activity. This study aims to model the stress levels of system operators based on system trustworthiness and physiological responses during a correct AI operation and its failure. Two groups of 18 and 19 people participated in the experiments using two different types of software with elements of AI. The first group of participants used English proofreading software, and the second group used drawing software as the AI tool. During the tasks, the electrodermal activities of the participants as a stress level indicator were measured. Based on the results obtained, the users' stress was determined and classified using logistic regression models with an accuracy of approximately 70%. The insights obtained can serve AI product developers in increasing the level of user trust and managing the anxiety and stress levels of AI operators.

INDEX TERMS Artificial intelligence, electrodermal activity, physiological stress, stepwise regression, system failure.

I. INTRODUCTION

According to numerous official dictionaries, artificial intelligence (AI) is the capability of a machine to imitate intelligent human behavior [1]. The main modern application areas of AI are machine learning, big data, and driverless cars [2]. Widespread adoption of AI can be attributed to the positive perception of novel technologies and innovations by users and customers; however, issues of user acceptance and trust in AI technology are becoming increasingly pressing every year [3]. Positively perceived technological characteristics of AI improve technology acceptance and use. These characteristics can improve the safety and performance of AI systems. For example, human actions and movement recognition can be used in smart homes and automated office AI environments to improve user comfort and safety [4], [5]. A prior study [4] elucidated this connection based on AI environments, which could detect user actions to increase user comfort. Subsequently, the corresponding safety issues were analyzed, and an automatic crime detection method for AI environments was proposed [5]. The positive impact of this approach was supported by previously developed technology acceptance models (TAMs). Various TAMs [6]–[8] demonstrated that personal characteristics such as usefulness, ease of use, and behavioral intention are important factors that influence technology acceptance and trust. Perceived usefulness and ease of use affect users' intentions and how they accept computer technologies [6]; they can also be used for TAM development. The three-layered trust model [7] demonstrated that operators' trust and perceived characteristics

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Jiang.

differ for each AI system type. In turn, the AI trust model, which incorporates the dynamics of trust, contextual AI use, and influence of display characteristics, was proposed [8]. A connection between trust and stress existed when there were AI mistakes and unreliability. This aspect was explained through writing task performance using AI software [9], [10]. In case of AI system failures, the user's trust gradually decreases and their stress increases. In other words, AI errors lead to a higher cognitive workload, mental stress, and decreased user trust. Previous studies have shown that establishing a positive relationship between user trust and their emotional stability during the AI system operation eased the adoption of new products; people tend to distrust AI products that exhibit failures during operation. This is evident in the case of autonomous vehicles and medical equipment because failures in these are directly related to the lives of users. A car accident report [11] showed that more than 25 crashes were related to autonomous vehicles in California from 2014 to 2017. In addition, it was found that proper automated system operations built trust and increased reliance on automated technology [12]. Moreover, AI mistakes and failures increased the cognitive workload of operators and the mental stress of users [13], decreasing their work efficiency [14]. User trust in AI technology is strongly related to its reliability and accuracy [7]. However, as the study indicates, it is difficult to achieve 100% accuracy, particularly in systems with high intelligence. Moreover, in a similar trend, users have exhibited varying degrees of sensitivity to AI reliability depending on the level of automation. The above studies demonstrated the importance of AI technology acceptance by users and its connection with AI adoption and user trust. Based on this, one of the primary objectives of current research is to encourage adopting new innovations and developing human trust in AI technologies using a modeling approach.

The growth of user-perceived trust in AI is an important issue that can be implemented in two main ways. First, AI technology can be improved to prevent an AI failure. Second, the user's emotions and state of stress should be considered to protect the user from dangerous failure-related situations such as a loss of control while using medical equipment with AI elements, driverless cars, and other devices. One of the important conditions for this is the use of objective instruments for stress measurements. Previous studies [15]–[17] have reported that an accurate indicator of physiological stress and states can be human responses, such as heart rate and electrodermal activity (EDA). EDA produces continuous changes in the electrical characteristics of the skin [18]. It refers to the variation of the electrical conductance of the skin in response to sweat secretion [18]. An experimental scheme based on EDA signals, which allowed one to recognize stressful events, was proposed [15]; it was found that the correct processing of EDA signals was the base for driving stress detection. Three psychological stress levels (low, medium, and high) were detected [16] based on EDA signal metrics, Fischer projection, and linear discriminant analysis. The accuracy of the proposed methods reached the satisfactory level of 81.82% and cemented the ability of the EDA signal to characterize human emotional states. Physiological responses obtained from sensors such as changes in heart rate, skin conductance, and respiration were cemented as accurate indicators of human rest and activity states [17]. The heart rate variability metric has been proposed as the base to predict individual human severity of congestive heart failure using the Bayesian belief network algorithm [19]. Study [20] showed that an EDA signal is an accurate measure of stressful conditions. Research [21] presented methods for analyzing EDA data to detect driver stress; it was found that EDA and heart rate metrics are the most correlated with a state of stress. Studies [22], [23] also supported findings that EDA is an indicator of emotional and stressful changes in human cognitive activity. Based on previous studies [13], [14], it can be concluded that a failure in the operation of an artificial intelligence system impacts the user physiological state through user stress occurrence, and user stress, in turn, can be measured through EDA. Additionally, it was found [15]–[17] that the main metrics characterizing human stress and its levels are psychophysiological indicators such as EDA and heart rate. Machine learning methods, including regression analysis, are most commonly used to apply these metrics and separate stress levels.

Previous studies have demonstrated several standard approaches to assessing human emotional states and cognitive processes. Research [24] discussed the prospect of using different approaches to evaluate cognitive processes in AI, including machine learning. They described the possibility of using machine learning to increase the efficiency of explainable AI in decision-making for the well-being of people. Machine learning methods were discussed [25] for data storage improvement in cloud computing and big data systems. The layer-wise perturbation-based adversarial training method used to predict hard drive health degrees based on different levels was proposed. Research guidelines were proposed to assess the scope of model explanation methods [26]. During this study, the following two approaches were adopted for predicting a learned model: linear and sum pooling convolutional network models. Researchers and designers have long recognized the importance of modeling stress and trust as significant influences on the acceptance and adoption of new technologies. On the basis of the aforementioned studies [6]–[8], [16], [24]–[26], the standard approaches to evaluate cognitive processes and human emotional states can be divided into the following five main groups:

1) survey to measure qualitative characteristics of an AI system;

2) regression modeling;

3) exploratory and confirmatory factor analysis including TAM;

4) predictive modeling; and

5) advanced machine learning modeling (such as random forests and support vector machines).

In many studies, including the present research, user stress based on trust in the AI system depends on the reliability of the system and the success of the task performed. When the task is performed successfully and the system operates reliably, then the user's trust is at a low-stress level and vice versa. Research [27] reported that if a particular task is simultaneously performed by AI and humans then, the failure of AI may induct a higher level of mistrust even if the human error causes more damage. In this case, the application of AI may be further reduced. Study [28] modeled user trust in AI and found that transparency, while the AI system is in use, can have negative effects on operator trust. Contradictions occur when the user has high trust in the event of an AI system failure and vice versa. A system calibration has been proposed as an approach to improve the performance and interruptions when using AI tools. The impact of trust in the adoption of AI for financial investment services was studied [29]. A prediction regression model of the intention of AI use was developed, including user trust. Trust was found to be one of the variables with the ability to significantly predict AI technology adoption. The methodology of perceived trust evaluation in AI technology was proposed [30]. It was found that the perceived difficulty, perceived performance, success/failure of the task, and task difficulty were extracted as the important predictors of perceived trust in AI system use. Physiological signals (heart rate) were studied [31] during the modeling of perceived trust and purchase intention in the apparel business. Messages about an apparel firm's malevolent business practices caused the heart rate of the users to decelerate and the perception of the firm as untrustworthy to increase. It was found that perceived trust has a greater impact on a participant's overall purchase intention for a malevolent business. The existing literature is mainly devoted to the dependence of trust on subjective assessments of perceived characteristics. Despite the fact that previous studies have recognized the importance of combining qualitative and quantitative approaches of analysis and assessment of the AI user psychological state [7], [23], it was reported that commonly adopted modeling approaches could be related to factor analysis, development of TAMs or separation of the subjective and objective personal scales. The present research, by contrast, links an objective assessment (physiological EDA signal) with user stress and AI system trustworthiness.

The aforementioned studies demonstrated the mutual influence between users' trust, stress, physiological signals, and task success. In this regard, this study investigated the relationship between users' physiological stress and physiological signals and how a user's trust in an AI system depended on its reliability—level of success or failure for each event. This helps understand how an AI user's stress and physiological state can be affected by a reliable or unreliable AI system if the performed task fails or is successfully completed. The proposed models also demonstrate the ability of physiological signals (EDAs) to detect and classify the stress levels of AI users. The methods developed in this study can be used to define the AI operator's stress levels. This study describes the mechanisms for building operator trust in AI technology from the user's perspective. This will help to adapt the AI systems to the psychological state of the operator and reduce the stress and fatigue of the users during the interaction. The insights from this study can help AI developers improve the attractiveness of their product among users and increase trust in their technologies throughout society. Designers can introduce our findings in interactive systems with AI elements such as mobile phones and apps, wristbands, wristwatches, tablets, and laptops. The objective of this study is to understand how the perceived trust and physiological responses of users, specifically an EDA signal, are affected during tasks using reliable and unreliable automation.

The present research includes two different approaches to detect user stress when using an AI operation system based on two experiments with AI software, which are described in detail in the sections below. In this study, the uninterrupted operation of AI was understood as the correct operation of the AI system in accordance with its purpose. Correct operation of the AI software had to occur without delay in time and with the implementation of all intended functions. In the case of the performed drawing experiment, AI correct operation is recognition of the drawings and the provision of professional versions of the sketches, and in the English proofreading experiment, the provision of word verification with the correct translation. The productivity and accuracy of the tasks performed were assessed through the success and completeness of the final result obtained in accordance with the AI user expectations. In the case of a drawing experiment, this is a recognized image and correctly proposed options for sketches, and in the case of an English proofreading experiment, this is correct recognition of an error in a word and a satisfactory proposal for its replacement. A brief description of the general model development process (Figure 1) contains data collection, data preprocessing, analysis, results, and comparison of the classifiers. The data collection step describes the collected datasets and the EDA device during both experiments. Data preprocessing introduces the preliminary data processing for each experimental set. The analysis and results steps show the analysis methods used with the main results. A comparison of the classifiers provides a general comparison of the developed models. The model application shows the most applicable areas of AI for the developed methods.

II. EXPERIMENTAL FRAMEWORK

A. EXPERIMENT 1: DRAWING SOFTWARE USING AI

1) PARTICIPANTS

A total of 18 healthy students (9 males and 9 females) from the same university with an average age of 22 years (standard deviation of 2.1 years) participated in this study. The participants did not have prior experience using this software and were informed that they could discontinue the experiment at any time.

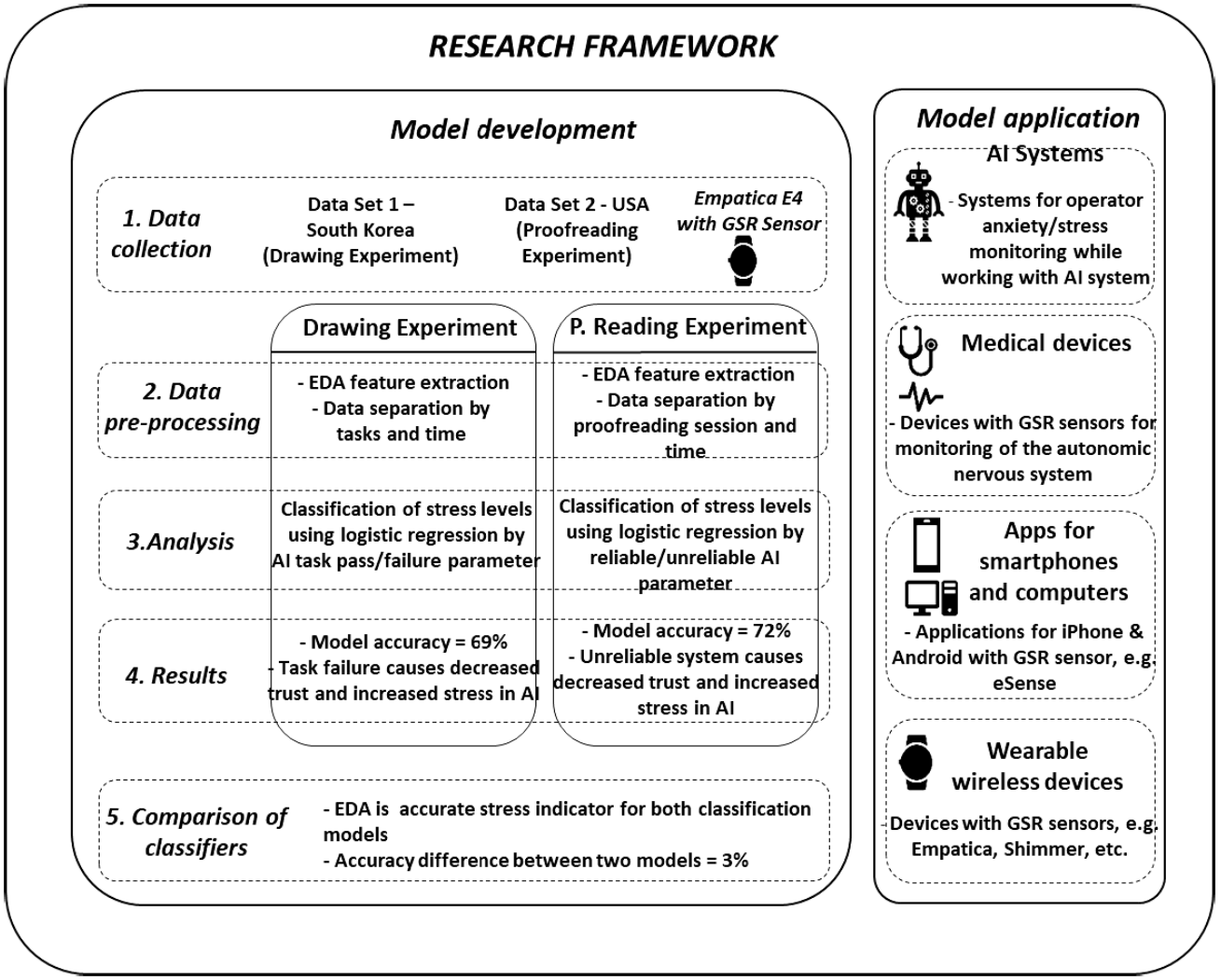

FIGURE 1. Development and application of the models.

2) EXPERIMENTAL SETUP

A Samsung Galaxy tablet (SM-T536; Samsung Group, Seoul, Korea) with a display of 10.07'' (∼255 mm), pixels resolution of 1280 × 800, and running the Android operating system was used as the experimental equipment. The participants used the stylus pen supplied with the tablet for interaction with the software. The correct operation of the devices was verified throughout the experiments. Samsung Galaxy tablet was chosen owing to its satisfactory quality that includes a thin and light structure, low power consumption, convenient surface temperature, bright display, and expandable storage system. These characteristics, combined with its reasonable price, make this tablet suitable for the experiment.

Google AutoDraw was selected as a representative AI. AutoDraw allows drawing objects based on AI principles by converting the user's inaccurate and rough input sketches into stylized drawings. Specifically, AI-based processing of the input generates candidate drawings for the users to choose and replace their original sketches.

3) DRAWING OBJECTS AND DIFFICULTY LEVEL

A preliminary experiment [30] was conducted to determine the drawing objects corresponding to words and to confirm the difficulty levels of the objects. The preliminary experiment consisted of the selection of target words by five participants who did not participate in the main experiment. The participants drew the objects corresponding to the proposed words using AutoDraw. The success of the task was determined from the correct recognition by the AI application.

A total of 50 words were selected using the Quick, Draw! game (Google LLC, Mountain View, CA, USA) from different topics to avoid biasing. The five participants then drew the objects corresponding to the words for up to 30 s. If the participant and experimenter agreed that the word was mapped onto drawings correctly, it was considered a success.

The degree of difficulty of the word was determined by the following approach. A scale from 1 to 10 was used for the assessment by the participants, with 1 representing the minimum difficulty level and 10 indicating the maximum. The success or failure of the tasks was assigned scores

of 0.5 and 1, respectively, and multiplied by the score of each participant. For example, the final score for ''blueberry'' was retrieved using the equation $0.5 \times 1 + 1 \times 10 + 0.5 \times 7 + 1 \times 7 + 0.5 \times 5 = 23.5$, where each term corresponds to one participant, with the left factor being the success/failure score and the right being the subjective score. The final scores allowed the classification of 50 words into low (score range of 2.5 to 16.5), moderate (score range of 17 to 33), and high (score range of 33.5 to 50) difficulties.

After classification, 18 words were selected from the 50 words to avoid redundancy, such as that between ''home lamp'' and ''street lamp,'' with varying interpretations according to cultural norms and conflicting, albeit correct, sketches of parts from larger objects. The remaining words, listed in Table 1, were classified according to their difficulty and used to conduct the main experiment.

TABLE 1. Experimental objects to be drawn according to their difficulty level.

Low difficulty	Moderate difficulty	High difficulty
Snail	Penguin	Dolphin
Wineglass	Fork	Unicorn
Feather	Duck	Trumpet
Windmill	Elephant	Bulldozer
Jacket	Hot air balloon	Sleeping bag
Ice-cream	Flashlight	Vacuum cleaner

4) MEASURES

An Empatica E4 wristband (EDA sensor) was used for the physiological signal collection in this experiment. The wristband [32] is a wearable and wireless device designed for comfortable, continuous, and real-time data acquisition in daily life. Data from this sensor were used as an objective measure with a sampling rate of 4 Hz throughout the tasks. In this study, for physiological EDA signals, the features proposed in [21] and the amplitude and duration calculated from signal peaks and valleys were used. The signal feature extraction process allows us to extract the following EDA characteristics of duration (OD) and amplitude (OM): the minimum (ODMin and OMMin), maximum (ODMax and OMMax), mean (ODMean and OMMean), standard deviation (ODstdev and OMstdev), summation (sum of ODsum and sum of OMsum), and the number of occurrences of duration and amplitude (ODN and OMN).

5) EXPERIMENTAL PROCEDURE

The 18 words (Table 1) were selected for the 18 participants to sketch in AutoDraw. The order of the selected words was arranged using the Latin square design. Each participant was then asked to sketch the object corresponding to the selected word. The words were not shown to the participant in advance. While drawing, the experimenter checked the success/failure, and the physiological signal of EDA was recorded using the Empatica E4 wristband.

B. EXPERIMENT 2: ENGLISH PROOFREADING SOFTWARE USING AI

1) PARTICIPANTS

A total of 19 native English speakers (10 females, 9 males) participated in the experiment, with a range of 18–82 years in age (mean = 33.6 years old; SD = 18.0). One participant's results were excluded due to an error in recording the EDA signal (data showed zero). The participants had at least two years of experience in using AutoCorrect in Microsoft Word.

2) APPARATUS

A previous program developed using Visual Studio C# (Visual Studio 2015, Microsoft Co., USA) was applied to conduct the experiment [10], [33]. The program included four different auto-proofreading sessions (i.e., sessions A, B, C, and D) [9]:

Session A: A reliable auto-proofreading condition indicating a grammatical error with an underline and without providing a suggestion (word).

Session B: A reliable auto-proofreading condition with a correct suggestion.

Session C: An unreliable auto-proofreading condition indicating a correct word with an underline and without providing a suggestion.

Session D: An unreliable auto-proofreading condition indicating a correct word with an underline and providing an incorrect suggestion.

Sentences used for the proofreading tasks were selected from online sentence completion test sets, for example, the Scholastic Aptitude Test (SAT) for the easy level and Graduate Record Examinations (GRE) for the difficulty level. A total of 34 sentences, i.e., 17 for the easy and difficult levels each, were chosen based on their readability scores, which were measured using the Readability Test Tool.

The training session contains 4 sentences, the manual proofreading session contains 10 sentences, and 20 sentences were included for each of the 4 sessions (sessions A, B, C, and D). The 13.5-inch laptop (Q524UQ, AsusTek Computer Inc., USA) was used for the experiment with a screen resolution of 1920 × 1080. The font type and size were Times New Roman and between approximately 14 and 16 points, respectively.

3) MEASURES

An Empatica E4 wristband (EDA sensor) was used for the physiological signal collection in our experiment. In this experiment, the aforementioned features proposed in the study [21] were also applied. The room temperature was controlled at approximately 22°C to block the effect of temperature on the skin conductivity. An example of data collected from this sensor throughout the tasks performed by one of the participants is shown in Figure 2. In Figure 2, the difference between reliable and unreliable sessions is indicated by a dotted line. The EDA values during the reliable experimental session were lower in comparison with those

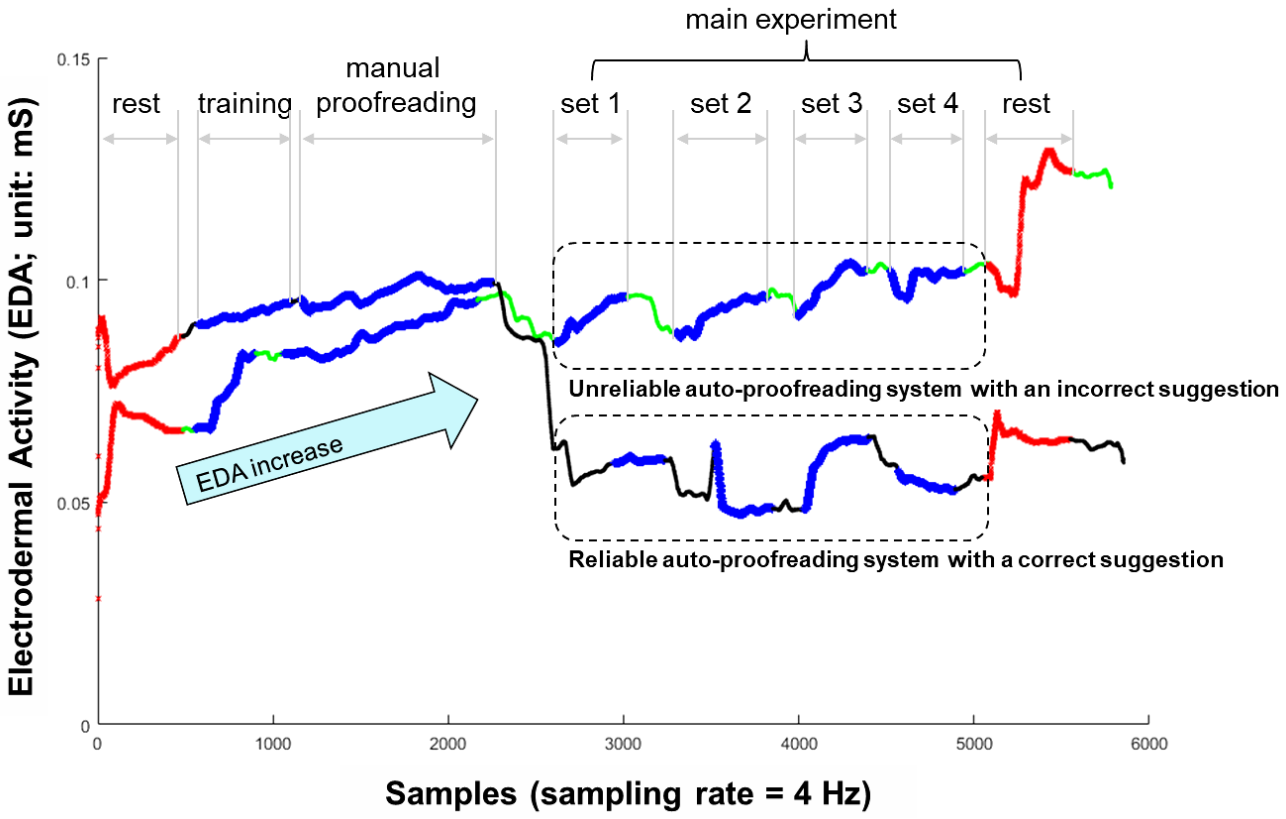

FIGURE 2. Example of collected EDA data during an experiment.

during the unreliable session. This is preliminary evidence that an unreliable system is associated with an increased stress level.

4) EXPERIMENTAL PROCEDURE

The experiment was divided into three stages: preparation, practice, and the main experiment. During the preparation stage, an Empatica E4 wristband was attached to each participant's wrist to measure the EDA signal, and the experimental procedure was described. A 2-min rest period was applied before starting the practice stage. The starting time and time for each task were recorded to synchronize with the measured EDA signals. For understanding the proofreading system, a practice stage was conducted for the participants. The users performed the proofreading tasks quickly and correctly during this stage. For increase the stress levels during the proofreading tasks, each sentence had to be corrected within 20 s. If the sentence was not completed within the time limit, the program would move automatically to the next sentence. Next, after a break, the manual proofreading session for the 10 sentences was conducted without an automated proofreading system. After the manual proofreading session, the participant had a 2-min rest period before starting the main experiment. During the main experiment stage, each participant was randomly assigned to one of the four sessions. During each session, the participant was asked to complete a set of five sentences as quickly and correctly as possible. The participants were asked to complete a total of 20 sentences, randomly separated into 4 sequential sets; perceived trust was measured at the end of each set. A short break period was included between the sets to observe a change in the physiological response.

III. ANALYSIS

A. EXPERIMENT 1: DRAWING SOFTWARE USING AI

1) PARTICIPANTS

Data analysis from the drawing AI software used in experiment 1 was based on the assumption that if the drawing task was completed successfully by the participant, then the participant has trust and a low-stress level (event "0"). A lack of trust with a high-stress level (event "1") corresponds to a failed drawing task. The analysis method was developed using a second-order polynomial logistic regression model. The dependent variable was the failure/success of the drawing AI software in the drawn word recognition. The independent variables were linear terms of extracted EDA features, products of their pairs, and squared terms of EDA features. A second-degree model was developed to find a more effective combination of predictors to increase the model performance because the first-order model showed a low accuracy of approximately 50%. At the same time, the degree of the regression model no longer increased owing to the possibility of an overload with a large number of terms in the equation. Finally, 36 variables were in the equation. The research

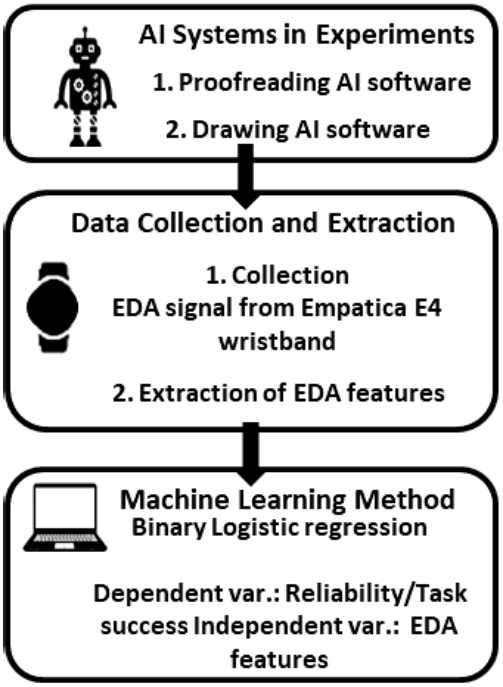

FIGURE 3. Research systems.

framework that explains the entire study system, including analysis, is illustrated in Figure 3.

Figure 3 divides the complete research framework into three systems with their respective elements. The First AI system in Experiments consists of proofreading and drawing AI software. The second system of data collection and extraction includes an EDA signal with extracted features. The third system of analysis comprises the machine learning method of binary logistic regression, wherein AI reliability and success were dependent variables, and EDA features were independent variables.

B. ANALYSIS OF EXPERIMENT 2 DATA

Data analysis from the English proofreading AI software used in experiment 2 was based on the assumption that a reliable auto-proofreading condition (with correct suggestion) corresponds to a low level of stress with trust (''0'') or a lack of trust with high-stress level (''1'') under non-reliable auto-proofreading conditions (with errors in the suggestion). The predictors were the only linear terms of the EDA features. The analysis method was developed using a first-order logistic regression model. The dependent variable was reliable/unreliable proofreading conditions. Independent variables were the only linear terms of the extracted EDA features. In this case, the linear model was sufficient to show a satisfactory result in the balance between model performance and the number of variables. Finally, seven variables were used in the equation. A schematic of the analysis process of both models is shown in Figure 4.

As shown in Figure 4, the analysis process consists of model development and cross-validation stages. The data collection step describes the collected data and information during both experiments. The model development step introduces the models obtained with dependent and independent variables (a detailed description is shown above in section 2.3). ''Extracted variables'' show the number of predictors obtained in each model. The cross-validation step provides a description of which parts of the cross-validated dataset the developed model was applied to. The ratio between the number of extracted variables from the AI experiment model and the number of cases from the cross-validated proofreading experiment makes it possible to apply the model equation only to the full cross-validated dataset. In cases of half and a quarter of the cross-validated dataset, this process was inaccessible because the number of variables obtained exceeded the number of cases in the dataset. Fewer variables in the proofreading AI model made it possible to apply this to all sections of the cross-validated AI dataset. The ''Results'' show data extracted after cross-validation. The model performance metrics obtained include the accuracy, sensitivity, specificity, and positive predictive value.

IV. RESULTS

A. EXPERIMENT 1: DRAWING SOFTWARE USING AI

During the drawing AI software experiment, stress classification was performed on the binary scale with low and high levels based on detected physiological responses from the measured EDA signal. During the performance of the task, the EDA signal was directly measured from a wristband sensor attached to the participant. In the case of successful task performance, it was assumed that trust existed along with a low-stress level (this event was coded as ''0''). Otherwise, a task failure caused a lack of trust with a high-stress level (this event was coded as ''1'').

The second-order polynomial logistic regression equation of 36 terms with the obtained coefficients can be described as follows:

$$P = 1/1 + e^{-(0.42121 - 0.00025025X1 + 0.0016284X2 + 418.96X3 + 1814.6X4 - 744.1X5 - 0.055392X6 + 0.057147X7 + 0.097449X8 - 1.4335X9 + 1.5152X10 + 1.9158X11 + 618.37X12 - 143.79X13 + 8.2973X14 - 10.881X15 + 1.6403X16 - 20.026X17 + 543.94X18 - 1661.7X19 - 2.391X20 + 3.7646X21 - 1.1173X22 + 1.2596X23 + 6.2165X24 - 361.36X25 - 1.0066X26 + 0.25811X27 + 0.47974X28 - 1.5348X29 - 0.00041707X30 + 0.0016321X31 + 0.021078X32 - 0.0051527X33 - 0.0070897X34 - 0.00038176X35)} \quad (1)$$

In equation (1), Y is a dependent variable assuming a user stress level through a failure or success of the task. All independent variables X1–X35 is squares and multiplications of extracted EDA data features (in Appendix A). Variables in the obtained model were significant with p-values not exceeding 0.04; the exceptions were only two insignificant variables, $OMMean^2$ and Constant with p-values of 0.1.

The model performance for drawing the AI software experiment is shown in Table 2. The developed model was applied and cross-validated using a full dataset of the second above

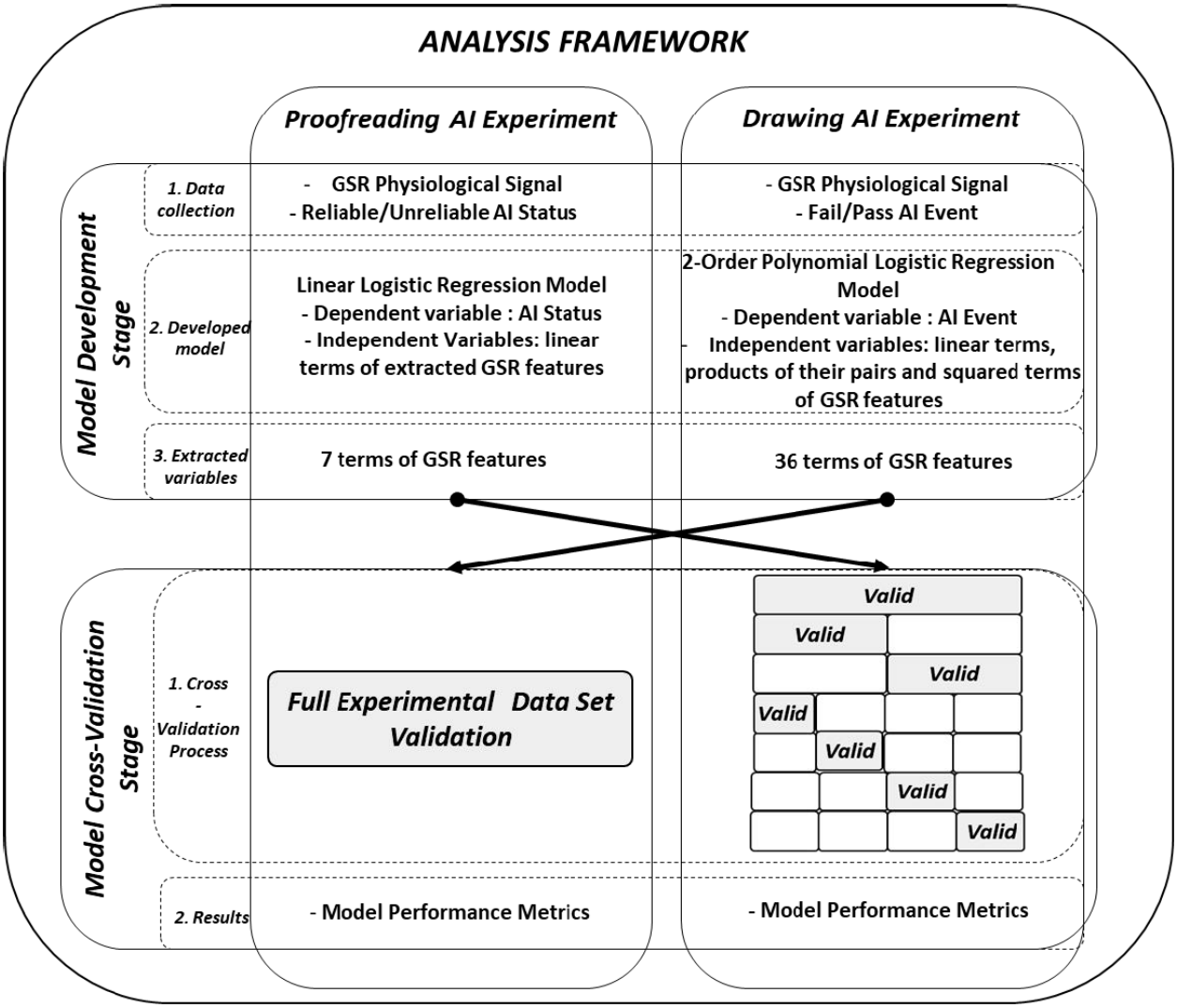

FIGURE 4. Analysis and cross-validation processes.

		Predicted: Low Stress	Predicted: High Stress	Total
Observed	Low Stress	66	76	142
	High Stress	25	157	182
		91	233	324

FIGURE 5. Confusion matrix of the originally developed model.

		Predicted: Low Stress	Predicted: High Stress	Total
Observed	Low Stress	25	7	32
	High Stress	8	32	40
		33	39	72

FIGURE 6. Cross-validated confusion matrix.

presented proofreading AI experiment. Confusion matrices for the developed model and cross-validation matrices are shown in Figures 5 and 6.

Table 2 shows that the accuracy, specificity, and sensitivity of the models ranged between 67% and 82%, whereas PPV ranged between 80% and 86% for the original model developed, i.e., "original," and the cross-validated model, "Cross-Val." The goodness of fit was evaluated using Cox and Snell pseudo-R-squares with values between 0.2086 and 0.2125. In general, a model based on an AI failure event has a satisfactory performance for both datasets.

B. ENGLISH PROOFREADING SOFTWARE WITH AI

In the English proofreading AI experiment, stress classification was also binary (low vs. high) and based on the

TABLE 2. Performance of logistic regression model based on full sets.

Model	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	Cox & Snell R^2
Original	69	67	72	86	0.2086
Cross-Val	79	82	75	80	0.2125

TABLE 3. Coefficients of logistic models based on proofreading AI experiment.

Predictor	Coefficient
ODmean	-.007
ODstdev	-.009
N	-1.168
ODsum	.002
OMmax	-3.286
OMstdev	7.284
Constant	4.535

		Predicted: Low Stress	Predicted: High Stress	Total
Observed	Low Stress	22	10	32
	High Stress	10	30	40
		32	40	72

FIGURE 7. Confusion matrix of the originally developed model.

		Predicted: Low Stress	Predicted: High Stress	Total
Observed	Low Stress	10	132	142
	High Stress	7	175	182
		17	307	324

FIGURE 8. Full set cross-validated confusion matrix.

TABLE 4. Performance of regression models.

Model	TP	FP	FN	TN	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %
Original	30	10	10	22	72	75	69	75
C/V_Full	175	7	132	10	57	57	59	96
C/V_Half1	152	25	113	34	57.5	57	58	86
C/V_Half2	150	27	108	39	59	58	59	85
C/V_Quarter1	172	10	134	8	56	56	44	94
C/V_Quarter 2	180	4	131	9	58	58	69	98
C/V_Quarter 3	168	16	122	18	57.5	58	53	91
C/V_Quarter 4	172	5	124	23	60	58	82	97

EDA signal, which was measured by the wristband sensor attached to the participant during the performance of the task. The proposed hypothesis is that a reliable auto-proofreading condition (with correct suggestion) corresponds to a low level of stress with existing trust (the event was coded as ''0'') or a lack of trust with a high-stress level (the event was coded as ''1'') under non-reliable conditions (with errors in the suggestion). The coefficients of the regression model explaining the reliability of the English proofreading AI software as a dependent variable are shown in Table 3. The model performance, along with the cross-validated results, are shown in Table 4. This model was cross-validated by applying the coefficients obtained to the dataset from the first presented AI drawing experiment. In this case, it was possible to cross-validate the model on different sections of the drawing experiment dataset (full, half, quarter) because of the balanced numbers of predictors and validating cases. The basic and validated confusion matrices obtained for the full datasets are shown in Figures 7 and 8.

In Table 4, ''Original'' is the result of the developed basic model, ''C/V Full'' indicates the results of the cross-validated full set, ''C/V_Half1'' shows the first half of the cross-validated set, ''C/V_Half2'' indicates the second half of the cross-validated set, and ''C/V_Quarter1-4'' shows the results of all cross-validated set quarters from 1–4 respectively. For the originally developed model, the accuracy is over 70%, with other characteristics of between 69%–75%. The goodness of fit was evaluated using the Cox and Snell pseudo-R-squares with a value of 0.214. For the cross-validated set, the accuracy varies between 56%–60%, with other characteristics of between 44%–97%. Based on the results obtained, the original model achieves a satisfactory performance.

V. DISCUSSION

A. PERFORMANCE OF MODELS

The present study proposed binary classification models of stress levels (high and low) of AI operators during system failure. Correct work and reliability of the AI system corresponded to low stress and the presence of trust in AI. Otherwise, if the AI system demonstrated failure or unreliability, mistrust and high-stress level occurred. The developed logistic models show a satisfactory accuracy, sensitivity/specificity, and positive predictive values (PPVs) of approximately 60%–80% on average for both models. In particular, the general PPV results show high values of approximately 90% or more. This indicates the high ability of the developed models to detect the lack of trust and high-stress level of operators while using AI systems. The goodness of model fit is assessed using various measures [34]. In our study, Cox and Snell pseudo-R-squares were used to evaluate the goodness of fit. Cox and Snell pseudo-R2 is unable to reach a value of ''1'' even for a perfect model [34]. The results obtained show that the original models developed explain between 0.20 and 0.22 of the variance at low and high-stress levels. In previous studies, there is no consensus on how

to interpret the values of the pseudo-R-squares, but some sources [35], [36] have evaluated a Cox & Snell level from 0.2 to be satisfactory and acceptable.

Previous studies used the EDA signal as the base for emotional recognition and reported the following results. A method to detect human emotions using EDA data in a word remember/recall task was proposed [22]. The authors used the positive and negative affect schedule method and support vector machine to classify the EDA response, with an accuracy of approximately 75.65%. This study [37] used various physiological responses, including EDA signals to study the cognitive and mathematical task performance. The accuracy of the models was between 75 and 95%, depending on the type of physiological signal. In this study, the authors performed the Stroop test, the Trier social stress test, and the Trier mental challenge test to evaluate emotions using EDA and speech features [23]. EDA data indicated that the best accuracy was approximately 70%. Another study [38] used the driver database and main object analysis to select the appropriate features for the identification of the stressful state. These factors resulted in an accuracy of 78.94%. An accuracy of approximately 89% was achieved [39] using the support vector machine approach to detect the stress state of participants in three types of tasks: Stroop color-word test, an arithmetic test of counting numbers, and talking about stressful experiences or events. During the comparison of previous and present methods, it was proposed that the EDA signal shows potential to recognize human emotions and stress levels. Average accuracy was between 70 and 90%, and the performance of the proposed model was within this range with scope for future improvement and development.

Based on the results obtained, the original stress models based on AI failure show satisfactory performance. This is connected with the fact that stress applied in this study is an objective measurement as well as a failure of the AI. This finding supports and expands the previous result of relations between stress and AI failure events, which were provided in studies related to autonomous driverless vehicles with AI tools. Research [40] provides the results of a survey of 1028 randomly selected Americans aged 18 and older. This study reported that 37% of men and 55% of women have anxiety about driverless car safety owing to the possibility of failure, and only 6% of people would put a child alone in a driverless car. Study [41] showed that people have a high level of anxiety when driving autonomous cars with AI systems owing to failure events. Supporters of autonomous driving have declared that AI technologies secure the driving process; however, despite this, consumers are under stress with the idea of being in a car that can break down or fail at any time without their control. AI system operators have psychological roadblocks in using automated technology because of a lack of control and understanding of how the system works, the risk of injury, and the unpredictability of failure moments [42].

Previous studies have shown that stress, anxiety, and AI failure are related to each other. In contrast to existing research, the present study proposed two validated stress models with satisfactory accuracy and performance. The proposed models are significantly different from the previously developed models. First, the combination of the trust concept with real physiological data was conducted in a single model for each individual experiment. Second, the developed models confirmed the relation between AI failure and the emotional state of the AI operator based on the objective measurements of the EDA signal. Third, a majority of the previous research was focused on building models based on subjective user assessments of the perceived characteristics of AI systems. In turn, the present study did not use subjective assessments but provided a further perspective on the combination of subjective and objective measurements of the emotional state of AI operators and users. The results obtained confirm previous research and provide new knowledge regarding sensors, AI/automation engineering, and physiological science for researchers, engineers, and designers.

B. RELATIONSHIPS AMONG EMOTIONAL STATE, PERCEPTION, AND AI OPERATION OF USERS

Both models developed in the present study have a satisfactory classification ability and demonstrate the mutual connection between user stress levels, AI failure, and system reliability. It was found that AI failure and unreliable AI systems have a positive influence on the stress of the users. The general assumption of this study is that an AI failure and unreliability cause increased stress based on a low trust in AI system operations owing to unpredictable AI reactions. A connection between trust and stress was described in previous research [9], [10], where it was found that AI mistakes and unreliability cause a higher cognitive workload and mental stress with decreased user trust. In other words, if the AI system fails or is unreliable, then the operator stress increases, and the trust level decreases. In the present study, the stress of the users was confirmed if the AI drawing software does not recognize the user's sketch or if the proofreading AI proposes an unsatisfactory suggestion, which corresponds to a low trust level. The present research expands and novelizes previous studies, which also found a general connection between user perception, trust, emotional state, and AI or automation system failures. In addition, [43] and [44] show the negative effect of automation errors on user trust. If the error occurs earlier, then the negative effect on trust is stronger. It was also found that the first impression of the system is the most important and forms the foundation of trust. Study [45] showed that if the operating system fails quickly and easily, it undermines the user's trust, and the operator's subsequent impression of the system will be ''untrustworthy.'' Based on this, one of the important problems for AI system designers is the prevention of early and easy errors by improving the feedback connection between themselves and users after a failure event. Examples were demonstrated in [46] on a collision warning system for drivers. Driver trust was significantly lower if the system gave a warning after pressing the brake pedal than before because of less benefit gained

by users. The study [47] connected human trust, stress, and physiological signals while using a computer interface with a VR tool. Electroencephalography (EEG), EDA, and heart rate variability (HRV) were used to measure trust with a virtual agent and find the connection between trust and stress. It was found that in low cognitive load tasks, EEG data reflect the trust in VR, and the cognitive load (stress or anxiety) of the user is reduced when the VR is accurate. The routine performance of automated systems causes a high level of user trust [48]. Trust becomes significant if the user does not know how to prevent the occurrence of AI system failures. This uncertainty influences the workload and further error management of the operator, particularly under time pressure conditions. A few principles were proposed to reduce human stress and increase trust during automated car driving [49]. One of the important factors in trust is the ability of the system to provide the operator with information about what the vehicle senses when a failure occurs. The interface should help predict the failure and its effects to provide the best performance. The information should be provided as quickly as possible so that users can react proactively, and in this case, the trust in the AI system increases. According to [50], stress levels and user stress responses are different and depend on the personal characteristics of the users in video gaming task performances. Users with higher experience in video gaming have lower distress and better performance. This indicates that an AI system operation causes lower stress and workload to experienced users regardless of the failure event. Trust positively affects human satisfaction and is negatively related to stress [51].

The present findings confirm mutual relationships among the user's stress, anxiety, trust, and AI operation with failure events. The developed models demonstrate the new sets of variables capable of classifying user stress during a failed AI operation.

C. LIMITATION OF THIS STUDY AND FUTURE RESEARCH

There are few limitations of this study related to the AI software experiment. First, the developed model was based on an AI success/failure event, which is not an entirely independent indicator because it is connected with the characteristics of the participants, such as their drawing skills and personal experience. These personal features cannot be predicted or controlled during the experiments. Another limitation is the short time to complete the drawing task and accordingly to precisely determine the psychophysiological signal, which could lead to mixed results of EDA detection in certain cases. The time condition was also impossible to control because the drawing time was strictly provided by AI software. Additionally, for the proofreading AI software, experimental results may vary depending on the native language and literacy of the participant. In this regard, the choice of participants for the proofreading AI experiment was limited to native English speakers.

Another limitation is the difference in the number of fully analyzed cases between the two experiments. AI drawing software experiments provide 4-times the number of cases than AI proofreading experiments. This could have influenced the results of the cross-validation, particularly in the case of the polynomial model, owing to the imbalance between the numbers of analyzed cases and predictors.

In the future, the presented research can be supplemented and expanded with a greater variety of AI tools. Future AI experiments will also be based on more versatile software that does not depend on the talents and special characteristics of the participants (e.g., talent for drawing, singing, native language, or literacy). The developed models can be improved by considering additional variables; for example, we can use physiological signal features together with another type of stress assessment tool. The stress levels presented were divided into ''low'' and ''high,'' and these categories can be extended by including a ''middle level'' as an example. The applied methods of data analysis and event prediction can be expanded by applying machine learning methods such as random forest and support vector machines.

VI. CONCLUSION

In the present paper, two cross-validated models were proposed for stress level classification (high and low) based on the physiological response (EDA), system reliability, and AI system failure. The original logistic models developed show a satisfactory performance and goodness of fit. It was found that the EDA signal features of the users can be reasonably accurate predictors for stress level classification in an AI system failure and reliable/unreliable AI system operation. The following conclusions were drawn:

1) The originally developed models achieve a satisfactory classification ability and acceptable goodness of fit and demonstrate the mutual connection among stress, AI failure, and unreliability.

2) Both stress models applied to the original experimental datasets show a satisfactory performance with an accuracy of approximately 70%.

3) Relationships among EDA signal features, stress, and AI system trustworthiness during an AI operation were found.

4) The combination of EDA features as polynomial and linear terms can predict the human stress levels during a reliable/unreliable AI system operation and successful or failed task performance.

The results obtained can be used for theoretical and practical applications. The study provides new knowledge for sensors, AI/automation engineering, and physiological science. The developed models and results obtained will help to adapt the AI systems to the psychological state of the operator and reduce the stress and fatigue of the users during interaction with the system. The insights from this study could serve AI developers to improve their product attractiveness among users and increase trust in their technologies throughout society. Designers can also introduce our findings in interacting systems with AI elements such as mobile phones and apps, wristbands, wristwatches, tablets, and laptops.

APPENDIX A

TABLE 5. Composition of polynomial Equation1 of drawing AI software.

Number of Variables in equation	Variable Composition
X1	$ODMax^2$
X2	$ODSum^2$
X3	$OMSum^2$
X4	$OMMean^2$
X5	$OMStDev^2$
X6	ODMaxN
X7	ODMinN
X8	ODStDevN
X9	ODMaxOMStDev
X10	ODMeanOMStDev
X11	ODStDevOMStDev
X12	OMMaxOMStDev
X13	OMMinOMStDev
X14	ODMaxOMMean
X15	ODMinOMMean
X16	ODSumOMMean
X17	ODStDevOMMean
X18	OMMaxOMMean
X19	OMSumOMMean
X20	ODMaxOMSum
X21	ODMinOMSum
X22	ODSumOMSum
X23	ODMeanOMSum
X24	ODStDevOMSum
X25	OMMaxOMSum
X26	ODMinOMMin
X27	ODSumOMMin
X28	ODSumOMMax
X29	ODMeanOMMax
X30	ODSumODStDev
X31	ODMaxODMean
X32	ODMinODMean
X33	ODSumODMean
X34	ODMinODSum
X35	ODMaxODMin
Abbreviation of EDA signal duration (OD) and amplitude (OM) are used: minimum (ODMin and OMMin), maximum (ODMax and OMMax), mean (ODMean and OMMean), standard deviation (ODstdev and OMstdev), summation (sum of ODsum and sum of OMsum), and the number of occurrences of duration and amplitude (N).	

APPENDIX B

TABLE 6. Table of nomenclature.

Symbol	Definition	Units
AI	Artificial Intelligence	-
EDA	Electrodermal Activity	mS
$ODMax^2$	Maximal Signal Duration	ms
$ODSum^2$	Summ. Signal Duration	ms
$ODMean^2$	Mean Signal Duration	ms
ODMin	Minimal Signal Duration	ms
ODStDev	Standard deviation of Signal Duration	ms
$OMSum^2$	Summ. Signal Amplitude	mS
$OMMean^2$	Mean Signal Amplitude	mS
$OMStDev^2$	Standard deviation of Signal Amplitude	mS
$OMMax^2$	Maximal Signal Duration	mS
OMMin	Minimal Signal Amplitude	mS
N	Number of Signal Peak Occurrences	-

REFERENCES

[1] "Artificial intelligence," in *Merriam-Webster Dictionary*, 11th ed. Springfield, MA, USA: Encyclopedia Britannica, 2003. [Online]. Available: https://www.merriamwebster.com/dictionary/artificial%20intelligence

[2] R. Ayers. (2016). *The Future of Artificial Intelligence: 6 Ways it Will Impact Everyday Life*. Accessed: Dec. 28, 2020. [Online]. Available: https://bigdata-madesimple.com/the-future-of-artificial-intelligence-6-ways-it-will-impact-everyday-life/

[3] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manage. Ann.*, vol. 14, no. 2, p. 627, Jul. 2020, doi: 10.5465/annals.2018.0057.

[4] L. M. Gladence, H. H. Sivakumar, G. Venkatesan, and S. S. Priya, "Home and office automation system using human activity recognition," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 0758–0762.

[5] L. M. Gladence, V. M. Anu, S. Revathy, and P. Jeyanthi, "Security management in smart home environment," *Soft Comput.*, 2021.

[6] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Manage. Sci.*, vol. 35, pp. 982–1003, Aug. 1989.
[7] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.
[8] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.
[9] H. Jeong, J. Park, and B. C. Lee, "Effects of automation type on human performance in proofreading tasks," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 62, no. 1, 2018, p. 1140.
[10] C. Branstrom, H. Jeong, J. Park, B.-C. Lee, and J. Park, "Relationships between physiological signals and stress levels in the case of automated technology failure," *Hum.-Intell. Syst. Integr.*, vol. 1, no. 1, pp. 43–51, Mar. 2019.
[11] F. M. Favaro, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, "Examining accident reports involving autonomous vehicles in California," *PLoS ONE*, vol. 12, no. 9, Sep. 2017, Art. no. e0184952.
[12] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, Jun. 2003.
[13] H. Jeong, J. Park, J. Park, T. Pham, and B. Lee, "Analysis of trust in automation survey instruments using semantic network analysis," in *Advances in Human Factors and Systems Interaction*, E. L. Nunes, Ed. London, U.K.: Springer, 2019, pp. 9–18.
[14] H. Jeong, J. Park, J. Park, and B. B. Lee, "Inconsistent work performance in automation, can we measure trust in automation?" *Int. Robot. Autom. J.*, vol. 3, no. 6, pp. 378–379, Dec. 2017.
[15] A. Affanni, R. Bernardini, A. Piras, R. Rinaldo, and P. Zontone, "Driver's stress detection using skin potential response signals," *Measurement*, vol. 122, pp. 264–274, Jul. 2018.
[16] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behavioural Brain Res.*, vol. 341, pp. 50–53, Apr. 2018.
[17] S. Ollander, C. Godin, S. Charbonnier, and A. Campagne, "Feature and sensor selection for detection of driver stress," in *Proc. 3rd Int. Conf. Physiolog. Comput. Syst.*, 2016, pp. 115–122.
[18] W. Boucsein, "Principles of electrodermal phenomena," in *Electrodermal Activity*, 2nd ed. New York, NY, USA: Springer, 2012, p. 2.
[19] L. M. Gladence, T. Ravi, and M. Karthi, "An enhanced method for detecting congestive heart failure—Automatic classifier," in *Proc. IEEE Int. Conf. Adv. Commun., Control Comput. Technol.*, May 2014, pp. 586–590.
[20] R. Zangróniz, A. Martínez-Rodrigo, J. Pastor, M. López, and A. Fernández-Caballero, "Electrodermal activity sensor for classification of calm/distress condition," *Sensors*, vol. 17, no. 10, p. 2324, Oct. 2017.
[21] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.
[22] D. B. Setyohadi, S. Kusrohmaniah, S. B. Gunawan, P, Pranowo, and A. Prabuwono, "Galvanic skin response data classification for emotion detection," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 31–41, 2018.
[23] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 209–214.
[24] J. E. T. Taylor and G. W. Taylor, "Artificial cognition: How experimental psychology can help generate explainable artificial intelligence," *Psychonomic Bull. Rev.*, vol. 28, no. 2, pp. 454–474, 2021.
[25] J. Zhang, J. Wang, L. He, Z. Li, and P. S. Yu, "Layerwise perturbation-based adversarial training for hard drive health degree prediction," 2018, *arXiv:1809.04188*. [Online]. Available: http://arxiv.org/abs/1809.04188
[26] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2018, *arXiv:1810.03292*. [Online]. Available: http://arxiv.org/abs/1810.03292
[27] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, 2015.
[28] P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *J. Decis. Syst.*, vol. 29, no. 4, pp. 260–278, Oct. 2020.
[29] W. Noonpakdee, "The adoption of artificial intelligence for financial investment service," in *Proc. 22nd Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2020, pp. 396–400.
[30] O. V. Bitkina, H. Jeong, B. C. Lee, J. Park, J. Park, and H. K. Kim, "Perceived trust in artificial intelligence technologies: A preliminary study," *Hum. Factors Ergonom. Manuf. Service Industries*, vol. 30, no. 4, pp. 282–290, Jul. 2020.
[31] J. Ha-Brookshire and G. Bhaduri, "Disheartened consumers: Impact of malevolent apparel business practices on consumer's heart rates, perceived trust, and purchase intention," *Fashion Textiles*, vol. 1, no. 1, pp. 1–12, Dec. 2014.
[32] (2018). *Empatica, User's Manual*. [Online]. Available: https://empatica.app.box.com/v/E4-User-Manual
[33] B. C. Lee, J. Park, H. Jeong, and J. Park, "Validation of trade-off in human—Automation interaction: An empirical study of contrasting office automation effects on task performance and workload," *Appl. Sci.*, vol. 10, no. 4, p. 1288, Feb. 2020.
[34] P. Allison. (2013). *What's the Best R-Squared for Logistic Regression*. [Online]. Available: https://statisticalhorizons.com/r2logistic
[35] B. Erichson, W. Plinke, and R. Weiber, *Multivariate Analysemethoden*, vol. 11, K. Backhaus, Ed. Berlin, Germany: Springer, 2006.
[36] L. Kraus, I. Wechsung, and S. Müller, *A Comparison of Privacy and Security Knowledge and Privacy Concern as Influencing Factors for Mobile Protection Behavior. Logistics Management Contributions of the Section Logistics of the German Academic Association for Business Research*, D. C. Mattfeld, T. S. Spengler, J. Brinkmann, and M. Grunewald, Eds. Braunschweig, Germany, 2015.
[37] K. Palanisamy, M. Murugappan, and S. Yaacob, "Multiple physiological signal-based human stress identification using non-linear classifiers," *Elektronika Elektrotechnika*, vol. 19, no. 7, pp. 80–85, 2013.
[38] Y. Deng, C.-H. Chu, H. Si, Q. Zhang, and Z. Wu, "An investigation of decision analytic methodologies for stress identification," *Int. J. Smart Sens. Intell. Syst.*, vol. 6, no. 4, pp. 1675–1699, 2013.
[39] E. Lutin, R. Hashimoto, W. De Raedt, and C. Van Hoof, "Feature extraction for stress detection in electrodermal activity," in *Proc. 14th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2021, pp. 177–185.
[40] A. Langfield. (2015). *Who is Most Anxious About Self-Driving Cars*. CBS Interactive. [Online]. Available: https://www.cbsnews.com/news/who-is-most-anxious-about-self-driving-cars/
[41] R. Ferris. (2017). *There is a Ton of People Who Still Don't Want to Ride in Self-Driving Cars, Says Survey*. [Online]. Available: https://www.cnbc.com/2017/08/24/consumers-still-anxious-about-autonomous-cars-says-gartner.html
[42] A. Shariff, J.-F. Bonnefon, and I. Rahwan, "Psychological roadblocks to the adoption of self-driving vehicles," *Nature Human Behaviour*, vol. 1, no. 10, pp. 694–696, Oct. 2017.
[43] D. Manzey, J. Reichenbach, and L. Onnasch, "Human performance consequences of automated decision aids: The impact of degree of automation and system experience," *J. Cognit. Eng. Decis. Making*, vol. 6, no. 1, pp. 57–87, Mar. 2012.
[44] J. Sanchez. (2006). *Factors That Affect Trust and Reliance on an Automated Aid*. [Online]. Available: https://smartech.gatech.edu/handle/1853/10485?show=full
[45] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, "Automation failures on tasks easily performed by operators undermine trust in automated aids," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 48, no. 2, pp. 241–256, Jun. 2006.
[46] G. Abe and J. Richardson, "Alarm timing, trust and driver expectation for forward collision warning systems," *Appl. Ergon.*, vol. 37, no. 5, pp. 577–586, 2006.
[47] K. Gupta, R. Hajika, Y. S. Pai, A. Duenser, M. Lochner, and M. Billinghurst, "In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in VR," in *Proc. 25th ACM Symp. Virtual Reality Softw. Technol.*, Nov. 2019, pp. 1–10.
[48] S. E. McBride, W. A. Rogers, and A. D. Fisk, "Understanding human management of automation errors," *Theor. Issues Ergonom. Sci.*, vol. 15, no. 6, pp. 545–577, Nov. 2014.
[49] O. Carsten and M. H. Martens, "How can humans understand their automated cars? HMI principles, problems and solutions," *Cognition, Technol. Work*, vol. 21, no. 1, pp. 3–20, Feb. 2019.
[50] J. Lin. (2017). *The Impact of Automation and Stress on Human Performance in UAV Operation. Electronic Theses and Dissertations, 2004–2019. 5710*. [Online]. Available: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=6710&context=etd
[51] J. Guinot, R. Chiva, and V. Roca-Puig, "Interpersonal trust, stress and satisfaction at work: An empirical study," *Personnel Rev.*, vol. 43, no. 1, pp. 96–115, Jan. 2014.

OLGA VL. BITKINA is currently with the User Value Laboratory, Incheon National University, Incheon, South Korea. Her research interests include data-driven human factors engineering, artificial intelligence, usability, and user experience in medicine.

JUNGYOON KIM received the B.S. degree in electronics and the M.S. degree in electrical and computer engineering from the University of Ulsan, Ulsan, South Korea, in 2004 and 2006, respectively, and the Ph.D. degree in information sciences and technology from Pennsylvania State University, University Park, PA, USA, in 2014. He is currently an Assistant Professor with the Department of Computer Science, Kent State University, USA. His areas of experience and expertise are smart health and wellbeing, especially in real-time cardiovascular disease and stress monitoring, physiological sensor design, and infrastructure and intelligent decision supports, sleep-related breathing disorder prediction, and environmental monitoring and assessment.

JANGWOON PARK received the B.S. degree in industrial engineering from Ajou University, in 2007, and the Ph.D. degree in industrial engineering from Pohang University of Science and Technology (POSTECH), in 2013. He is currently an Assistant Professor with the Department of Engineering, Texas A&M University—Corpus Christi, TX, USA. His research interests include human factors and ergonomic product designs.

JAEHYUN PARK received the B.S. and Ph.D. degrees in industrial and management engineering from Pohang University of Science and Technology (POSTECH). He worked at the User Experience Team, Samsung Electronics, as a Senior Designer, from 2013 to 2015. He is currently an Associate Professor with the Department of Industrial and Management Engineering, Incheon National University (INU). His research interests include semantic network analysis, machine/deep learning on physical behavior, and computational cognitive engineering.

HYUN K. KIM received the B.S. and Ph.D. degrees in industrial and management engineering from Pohang University of Science and Technology (POSTECH). She worked at Samsung Electronics, as a Senior Designer, from 2015 to 2018. She is currently an Associate Professor with the School of Information Convergence, Kwangwoon University.