A Data-Driven Approach Utilizing Body Motion Data for Trust Evaluation in Industrial Human-Robot Collaboration

Campagna, Giulio; Dadgostar, Mahed; Chrysostomou, Dimitrios; Rehm, Matthias

Published in:
33rd IEEE International Conference on Robot and Human Interactive Communication (IEEE RO-MAN 2024)

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

A Data-Driven Approach Utilizing Body Motion Data for Trust Evaluation in Industrial Human-Robot Collaboration*

Giulio Campagna[1], Mahed Dadgostar, Dimitrios Chrysostomou[2], Matthias Rehm[1]

Abstract— **Industry 5.0 signifies a transformative era where humans and robots collaborate closely, leading to advancements in manufacturing efficiency and personalization. In light of this, it becomes essential to assess the robot's trustworthiness to ensure a secure environment and equitable workload distribution. The majority of trust assessments hinge on post-hoc questionnaires for the extent of trust experienced during the interaction. A data-driven approach is required to promptly assess trust levels in real-time, allowing for the adjustment of robot behavior to align with human needs. The paper proposes a chemical industry scenario where a robot assisted a human in the process of mixing chemicals. Several machine learning models, including deep learning, were developed using body motion data to categorize the level of trust exhibited by the human operator. The models achieve an accuracy exceeding 90%. The results clearly show the feasibility of data-driven trust assessment.**

I. INTRODUCTION

Trust is a critical component of successful Human-Robot Collaboration (HRC), enabling humans to rely on robotic teammates to competently perform assigned tasks [1]. As industrial environments increasingly adopt collaborative robots, ensuring seamless HRC and maintaining appropriate trust levels is imperative. While high levels of trust facilitate fluent teamwork, placing blind trust in robotic capabilities or insufficiently monitoring robotic actions can endanger the safety of human operators. Conversely, lacking trust impedes collaboration, overburdening the human operators with decision fatigue [2], [3]. As Lu et al. [4] discussed, safety, cognitive health, psychological wellness and wellbeing form the base of the Industrial Human Needs Pyramid. Thus, appropriately calibrating trust between humans and robots is required for productive and safe HRC while enabling self-actualization.

Trust can be described as the operator's confidence in the machine's competence, emphasizing the necessity for the operator to believe that the system effectively fulfills its tasks [1]. Hancock et al. [5] analyzed factors influencing trust in human-robot interaction across three categories: human-related (e.g. abilities, personality), robot-related (e.g. reliability, proximity), and environment-related (e.g. group dynamics, task complexity). Experimental evidence demonstrates certain factors impact trust, including robot transparency [6], robot appearance [7], humanized dialogue systems [8] and task criticality [9].

Most prior research has evaluated trust post-hoc via questionnaires [10]–[13]. While validated, these provide only summative evaluations, lacking real-time assessment while retrospective rationalization in surveys may not accurately capture actual behaviors [14]. Categorizing trust dynamics during interactions is essential for effectively managing trust when controlling robots. Nevertheless, dynamically adjusting robot behavior requires recognizing trust fluctuations as they occur. To address these gaps, recent research explores using data-driven approaches for implicit, continuous trust evaluation. For instance, body-worn inertial sensors have been applied for detecting distrust through increased limb movements [15]. In similar fashion, vision systems can estimate trust levels by tracking facial expressions and body language [16]. However, research using wearable sensors to robustly infer trust remains limited.

Recent studies demonstrate proxemics and risk-taking impact user trust, with closer proximity and unexpected robotic motions diminishing trust [17]. Building on this prior work [17], this study investigates using on-body sensors and machine learning to correlate motion data with trust levels. The chemical industry context offers a representative testbed, requiring close collaboration on handling hazardous materials. The tasks involve the human waiting while the robot pours liquid, providing a scenario where sudden robotic arm movements may indicate declining trust in the robot arm. The main contributions of this work are twofold:

- Devising a data-driven framework to categorize trust in real-time based on human motion data from on-body IMU sensors.
- Training of the framework with state-of-art machine learning models to map motion cues to trust ratings. This will allow us to adapt the robot controller in real-time to maintain appropriate trust, ensuring safety and balanced workloads.

II. METHODOLOGY

In a prior study, we demonstrated that trust in a human operator is affected by both low and high performance of the robot, resulting in correspondingly diminished or elevated trust ratings [17]. The current data collection process was built upon these findings, aiming to replicate the scenario for automatically labeling behavioral data. A chemical industry scenario was devised, where the task unfolded in two distinct stages. Initially, the robot handed a beaker containing a chemical to the human, who held it. Following this, the

* The work described in this paper was funded by the Independent Research Fund Denmark, grant number 1032-00311B.

[1] are with the Technical Faculty of IT and Design, Aalborg University, 9000 Aalborg, Denmark. {gica, matthias}@create.aau.dk.

[2] is with the Faculty of Engineering and Natural Sciences, Aalborg University, 9220 Aalborg, Denmark. dimi@mp.aau.dk.

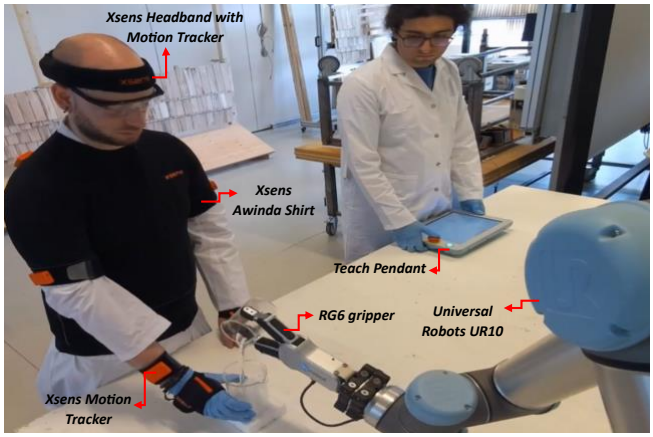


Fig. 1: The chemical industry scenario with the body motion capture system.

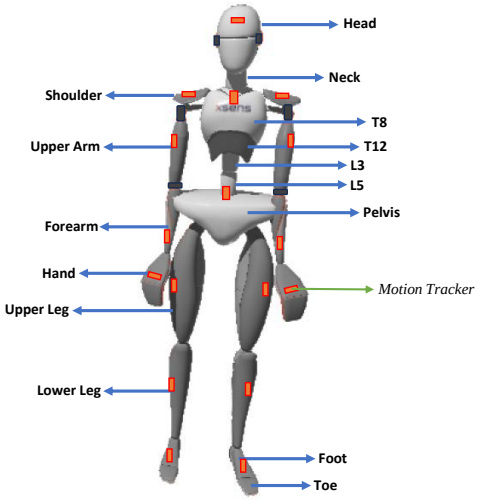


Fig. 2: The various anatomical sections of the body (indicated with blue arrows) and the placement of the motion trackers.

robot grabbed another beaker filled with a different chemical and poured it into the beaker held by the human. The robot exhibited two distinct operational modalities in terms of its performance capabilities: *low performance* eliciting *low trust* and *high performance* eliciting *high trust*. For each operational modality, there were two distinct trials conducted. In each trial, the robot executed distinct trajectories for handing the beaker to the user and for pouring the chemical.

In the *low-performance* condition, during the handing phase, the robot's approach to the human was overly close resulting in an uncomfortable and non-ergonomic situation for the human. In the pouring phase, the robot's actions gave the impression that it might pour the chemical onto the human's hand, thereby creating a potentially hazardous situation for the human operator.

In the *high-performance* case, the robot was efficiently handing and position the beaker as required. These actions were executed smoothly so that they didn't cause any stress or anxiety for the user. Likewise, during the pouring stage the robot was successfully pouring the chemical into the beaker without encountering any issues or unexpected behaviors.

The task itself did not involve any movement on the part of the human operator during any of its stages, unless for safety reasons. Consequently, as also affirmed by the participants, any noticeable movements on the operator's part were a result of diminished confidence in the robot's performance due to unexpected behaviors and potentially hazardous situations (e.g., the robot approaching the human too closely, posing a risk of collision).

A. Experimental Setup

Figure 1 provides an overview of the experimental setup. Participants interacted with the 6-axis *UR10-CB3-Series Robot*, which was equipped with the flexible 2-fingered gripper *RG6*. The robot's trajectory was pre-programmed. However, the participants were informed that the robot exhibited dynamic and autonomous behavior, which made it susceptible to potential malfunctions. To capture body motion data, participants wore a *Xsens MVN Awinda* motion tracking suit with 17 wireless motion trackers (MTws) that were positioned to specific anatomical locations (refer to Fig 2). These sensor modules come equipped with inertial and magnetic measurement units, housing 3D gyroscopes, 3D accelerometers, and 3D magnetometers. The data were collected with a frequency of 60 Hz.

Lastly, both the participants and the assistant were outfitted with personal protective gear, which included lab coats, gloves, and safety glasses. Regarding the chemicals used, the human-held beaker contained baking powder, while the robot-held beaker contained water. Consequently, during the reaction phase, only carbon dioxide was generated, ensuring the safety of all involved. The true composition of the chemicals was disclosed at the end of the experiment, as participants were initially informed that the substances could be potentially hazardous. As an additional safety measure, the assistant was equipped with an emergency button to stop the robot in the event of potential collisions between the human operator and the robot.

B. Experimental Protocol

The study included 20 participants, 10 males and 10 females with different age (M=29.1, SD=7.54). The participants were recruited from the students of Aalborg University.

The experiment protocols adhered to the *Declaration of Helsinki*. Involving human participants, the study underwent ethical review and received approval from the institutional review board. Additionally, at the beginning of the experiment, each individual was given a printed consent form and provided with information regarding the study's purpose, the tasks involved, associated risks, research methods, and potential benefits of the analysis.

The participant received support from the assistant in wearing the protective equipment and ensuring that the Xsens suit was tightened as much as possible to minimize motion artifacts, such as the shifting of MTws. Then, the 17 MTw

sensors were affixed to the body straps. When positioning sensors on the body, the initial alignment between the sensors and body segments is often uncertain [18]. To address this, a calibration procedure becomes essential for establishing both the alignment of the sensors with the body and the body's dimensions. The *N-pose calibration* was chosen (refer to [19] for further details) and body measurements were obtained for achieving precise calibration results. After the calibration, the participant performed the task four times with the operational modalities mentioned in the beginning of this section. Taking into account all the stages of the experiment, each participant required a total of 30 minutes.

C. Data Collection

To ensure the creation of a compact and meaningful dataset, data collection was strategically timed to capture specific moments within the overall task, particularly focusing on human reactions during the handing and pouring phases. Therefore, recording commenced upon receipt of control signals from the robot, and data acquisition occurred at a frequency of 60 Hz, encompassing all 23 body segments[1]. Kinematic quantities, specifically *linear* and *angular velocities*, as well as *linear* and *angular accelerations*, were recorded for each body segment, as they provide direct insights into human reactions.

In conclusion, the dataset for the *handing phase* comprised 20277 samples and 276 columns, representing 23 body segments, each with data on four kinematic quantities (linear and angular velocities, linear and angular accelerations), and three components (x, y, z). As for the *pouring dataset*, it consisted of 16412 samples.

D. Data Pre-Processing

For each participant, each data sample was automatically labeled as *high* or *low trust* based on the trial's characteristics. Specifically, if the robot operated with a low-performance, it was assigned a *low-trust* label; otherwise, it was assigned a *high-trust* label. Exploratory data analysis was conducted as an essential initial step, involving the examination of data distributions, pattern identification, and the detection of potential outliers. Therefore, the dataset underwent a refinement process where outliers were identified and subsequently removed. This outlier removal was executed utilizing the Z-score method, thus ensuring a more robust and reliable dataset for further analysis and modeling. A z-score of 3 was used as threshold. Consequently, the *Handing dataset* comprised 19702 samples (reduction of 2.8%), whereas the *Pouring dataset* contained 15835 samples (reduction of 3.5%). Afterwards, the magnitudes of linear and angular velocities (likewise for accelerations) for each body segment i were determined at each time instant t_k. As exemplification, it is provided the calculation with linear velocities (1):

$$v^i(t_k) = \sqrt{v_x^i(t_k)^2 + v_y^i(t_k)^2 + v_z^i(t_k)^2} \tag{1}$$

[1]https://base.movella.com/s/article/MVN-Biomechanical-Model?language=en_US

where $v^i(t_k)$ represents the magnitude of the velocity, while $v_x^i(t_k)$, $v_y^i(t_k)$ and $v_z^i(t_k)$ correspond to the x,y, and z components of the linear velocity at time t_k for each body segment i.

In conclusion, the *Handing dataset* contained 19702 samples and 92 columns, representing the magnitudes of the four kinematic quantities for each of the 23 body segments. Similarly, the *Pouring dataset* also consisted of 15835 samples with 92 columns of data. In the following, it is described the data processing for both machine learning and deep learning analysis, as the authors examined both approaches to understand the correlation between body motion data and trust levels of the human operator. Referring to *deep learning* analysis, the last preprocessing steps involved employing label encoding and standardization of the dataset.

In the context of *machine learning*, the next phase consisted in the extraction of the features. For every kinematic quantity related to each specific body segment, it was calculated the following features within a 10-sample window (downsampling applied for noise reduction): mean, median, standard deviation, minimum, and maximum. The aforementioned features were determined for every trial undertaken by each participant. Consequently, the *Handing dataset* comprised 1960 samples and 460 features columns, while the *Pouring dataset* contained 1557 samples and 460 features. Following that, label encoding and dataset standardization were performed. Due to the extensive range of features available, a two-step approach was employed: first, a tree-based algorithm was utilized to select the most significant features, followed by Principal Component Analysis (PCA). This sequential approach was chosen to provide a new feature-space using a set of representative features. Given the utilization of ensemble methods such as Random Forest, XGBoost, and LightGBM in the machine learning analysis, it was reasonable to employ a tree-based algorithm for feature selection. To this end, XGBoost was employed due to its exceptional efficiency and optimized speed in handling high-dimensional data. The gain metric was employed for feature importance assessment. Gain quantifies the average performance enhancement attributed to each feature during the model's training process. To conduct feature selection, we initially applied a threshold of 5% of the maximum importance score to select the most crucial features. This resulting number was then fine-tuned through a trial-and-error approach. Ultimately, 31 features were chosen for the *Handing dataset*, while 29 features were retained for the *Pouring dataset*. In the concluding phase, PCA was applied to retain essential feature information while reducing the dataset's complexity. PCA accomplishes this by mathematically deriving a new set of coordinate axes in the feature space, referred to as principal components. These principal components are calculated such that the first one explains the maximum variance in the original data. The second principal component, orthogonal to the first, explains the second most variance, and so on. In PCA, a crucial element is determining how many principal components to keep. The authors followed the common approach that involves

selecting components that capture 95% of the total variance. As a result, 11 principal components were chosen for the *Handing dataset*, and 13 were selected for the *Pouring dataset*. After applying PCA to reduce the dimensionality of the dataset, the resulting principal components were used as input for the machine learning modeling phase.

III. EXPERIMENTAL RESULTS

The proposed analysis is centered on the investigation of a potential correlation between the user's trust level, which has been categorized as either *high* or *low* trust (i.e. binary classification problem), and the observable behavioral changes detected within body motion data. This examination involved the utilization of a combined approach, incorporating both machine learning and deep learning techniques for both handing and pouring sections of the experiment.

A. Machine Learning Analysis

To examine the binary classification of trust (i.e., distinguishing between high and low trust), a selection of machine learning models was made. These models included **Random Forest**, **XGBoost**, and **LightGBM**. They all employ an ensemble learning approach, combining multiple models (specifically, decision trees) to enhance predictive accuracy. Furthermore, these algorithms demonstrate robustness when dealing with noisy data and outliers, making them suitable for real-world datasets. In addition to employing these classifiers, the authors opted to incorporate a **Voting Classifier** into their methodology. A Voting Classifier is an ensemble technique that combines the predictions of multiple individual classifiers. This combination of diverse models allows the Voting Classifier to leverage the unique strengths and characteristics of each constituent classifier. By considering the input from multiple models, the ensemble strategy aims to improve the robustness of predictions, reduce overfitting, and ultimately achieve higher classification accuracy.

For both *Handing* and *Pouring* datasets, the training set consisted of 70% of the data (14 participants), while the test set included the remaining 30% (6 participants). This participant-based split ensured that the model was evaluated using unseen data. In the following, each model's hyperparameters tuning and the corresponding classification accuracy are documented. The model's hyperparameters were fine-tuned using *Grid Search Cross-Validation* with 5-fold cross-validation approach.

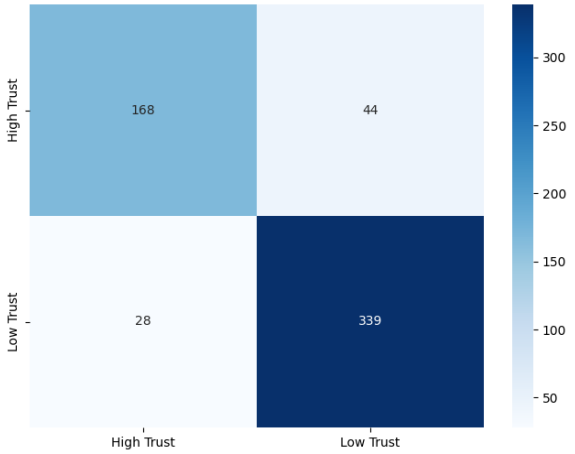
Concerning **Random Forest**, the hyperparameters subjected to tuning included the maximum depth of each decision tree within the ensemble (*max depth*), the minimum number of samples required in a leaf node (*min samples leaf*), the minimum number of samples necessary to split an internal tree node (*min samples split*), and the number of individual decision trees in the ensemble (*n estimators*). For the *Handing* analysis, the best hyperparameters were determined to be *max depth* 10, *min samples leaf* 4, *min samples split* 2, and *n estimators* 100. By comparison, for the *Pouring* analysis, the optimal hyperparameters were found to be *max depth* 20, *min samples leaf* 1, *min samples split* 5, and *n estimators* 50. The corresponding classification accuracy rates were 87.22% for *Handing* and 91.94% for *Pouring*.

With reference to **XGBoost**, the considered hyperparameters included *max depth*, *n estimators*, and *learning rate*. The learning rate regulates the step size at each iteration when approaching the loss function's minimum. For both *Handing* and *Pouring* analysis, the optimal hyperparameters were identified as *max depth* 3, *n estimators* 50, and *learning rate* 0.05. The resulting accuracy was 87.56% for *Handing* and 92.16% for *Pouring*.

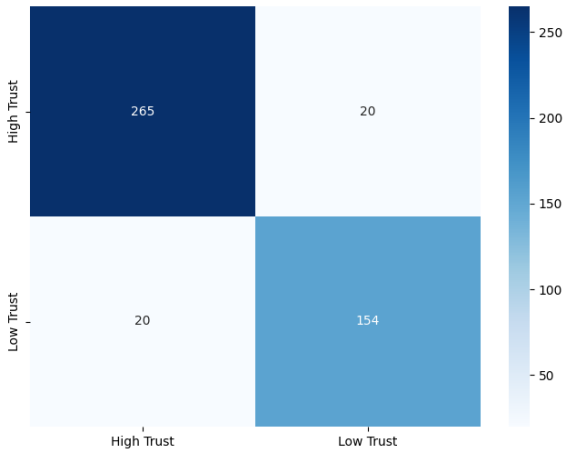
The third ensemble algorithm utilized was **LightGBM**. The hyperparameters considered for optimization encompassed *boosting type*, *learning rate*, *max depth*, *n estimators*, and the number of leaves in the decision tree (*num leaves*). Boosting type refers to the strategy used to combine the outputs of numerous weak learners, frequently represented as decision trees, in order to create a resilient predictive model. Concerning the *Handing* study, the optimal hyperparameters were found to be as follows: *boosting type* 'dart', *learning rate* 0.2, *max depth* set to None, *n estimators* 50, *num leaves* 31. Similarly, in the *Pouring* study, the optimal hyperparameters were determined to be: *boosting type* 'goss', *learning rate* 0.05, *max depth* set to None, *n estimators* 100, *num leaves* 63. The classification accuracy was 87.05% for *Handing* and 87.80% for *Pouring*.

Lastly, the analysis included the use of the **Voting Classifier**. The Voting Classifier employs two primary voting techniques: hard voting and soft voting. In the former, each individual model within the ensemble contributes a prediction, and the final prediction is determined by selecting the class that garners the majority of votes. In the latter, the Voting Classifier considers the class probabilities predicted by each individual model rather than counting class labels. It computes the average probability for each class and selects the class with the highest average probability as the final prediction. Therefore, the hyperparameter to tune was the voting method. For *Handing* analysis, it was selected 'hard' voting while for *Pouring* 'soft' voting. The level of agreement among models can differ between datasets, leading to variations in the suitability of hard or soft voting. In cases of high model agreement, hard voting may be effective, while in datasets with more diverse model predictions, soft voting could be a better option. Concerning each base learn model (Random Forest, XGBoost, LightGBM), it was utilized the previously discovered optimal hyperparameters. The classification accuracy was found to be 87.56% for *Handing* and 91.07% for *Pouring*.

To conclude, Table I presents the machine learning models and their associated performance indicators for *Handing*, while Table II provides the corresponding information for *Pouring*. Additionally, the confusion matrices relative to the Voting Classifier are reported for both *Handing* (Fig. 3a) and *Pouring* (Fig. 3b).



(a) Handing.



(b) Pouring.

Fig. 3: Confusion Matrices relative to the Voting Classifier.

TABLE I: Machine Learning models and performance indicators for Handing analysis.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	87.22%	0.86	0.86	0.86
XGBoost	87.56%	0.87	0.86	0.86
LightGBM	87.05%	0.86	0.85	0.86
Voting	87.56%	0.87	0.86	0.86

TABLE II: Machine learning models and performance indicators for Pouring analysis.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	91.94%	0.92	0.91	0.91
XGBoost	92.16%	0.92	0.91	0.92
LightGBM	87.80%	0.87	0.87	0.87
Voting	91.07%	0.90	0.91	0.91

B. Deep Learning Analysis

To supplement the analysis, deep learning algorithms were incorporated to explore intricate global sequential patterns within the body motion data. To this end, it was utilized the following architectures: **Long Short-Term Memory (LSTM)**, **Gated Recurrent Unit (GRU)**, and **1D Convolutional Neural Network (1D-CNN)**.

In the case of both the *Handing* and *Pouring* datasets, the division of data was as follows: the training set comprised 60% of the data (12 participants), the test set encompassed 20% (4 participants), and the validation set also accounted for 20% (4 participants). Similarly to the machine leaning analysis, this partitioning, based on participants, guaranteed that the model underwent evaluation with entirely new, unseen data.

In the following, the model architectures will be presented, along with the corresponding classification accuracy results. For both *Handing* and *Pouring*, it is noteworthy that two

TABLE III: Deep Learning models and performance indicators for Handing analysis.

Optimizer	Model	Accuracy	Precision	Recall	F1-score
Adam	LSTM	94.95%	0.94	0.96	0.95
	GRU	96.30%	0.96	0.97	0.96
	1D-CNN	95.98%	0.95	0.97	0.96
SGD	LSTM	93.05%	0.92	0.94	0.93
	GRU	93.53%	0.93	0.95	0.93
	1D-CNN	95.24%	0.94	0.96	0.95

TABLE IV: Deep Learning models and performance indicators for Pouring analysis.

Optimizer	Model	Accuracy	Precision	Recall	F1-score
Adam	LSTM	83.71%	0.87	0.74	0.77
	GRU	84.11%	0.87	0.75	0.78
	1D-CNN	84.34%	0.87	0.75	0.78
SGD	LSTM	83.40%	0.87	0.74	0.77
	GRU	84.62%	0.88	0.76	0.79
	1D-CNN	84.76%	0.88	0.76	0.79

optimization algorithms were applied to the models: *Adaptive Moment Estimation (Adam)* and *Stochastic Gradient Descent (SGD)*. Concerning *Adam*, learning rate was 0.001 to regulate the weight update step size. Additionally, the parameters *beta 1* and *beta 2* were set at 0.9 and 0.999, respectively, governing the exponential decay rates for the first and second moments of gradients, contributing to adaptive learning rates. Regarding *SGD*, a learning rate of 0.001 was applied, and a momentum factor of 0.9 was incorporated to utilize past gradients for faster convergence. Lastly, the models were compiled using the loss function binary cross-entropy and were trained for 20 epochs with a batch size of 32.

The first algorithm utilized was **LSTM**. The two LSTM layers were configured as follows: the first layer had 32

units with *tanh* activation and featured L2 regularization on kernel, bias, and activity terms, along with a dropout rate of 0.2. The second layer had identical settings. Subsequently, the data was flattened. A dense layer with 64 units (*'relu'* activation) was used, which incorporated L2 regularization (0.5 dropout). The final dense layer 1 unit, *'sigmoid'* activation) handled binary trust level classification with L2 regularization. Referring to *Handing*, when employing the Adam optimizer, an accuracy level of 94.95% was achieved, compared to 93.05% with SGD. Regarding *Pouring*, the utilization of the Adam optimizer yielded an accuracy rate of 83.71%, in contrast to 83.40% when using SGD.

Subsequently, the implementation of the **GRU** model was carried out. The model implemented two GRU layers with 64 units and *'tanh'* activation, while L2 regularization and dropout (*rate 0.2*) prevent overfitting. After flattening the data, two dense layers with 128 units and *'relu'* activation followed, each with L2 regularization and dropout (*rate 0.5*) to enhance generalization. The final layer, with 1 unit and *'sigmoid'* activation, performed binary classification while also using L2 regularization. In the context of *Handing*, the classification accuracy reached 96.30% when utilizing the Adam optimizer, whereas SGD yielded an accuracy of 93.53%. By comparison, for *Pouring*, the utilization of the Adam optimizer resulted in a classification accuracy of 84.11%, while using SGD yielded a slightly higher accuracy of 84.62%.

The last algorithm concerned **1D-CNN**. The model comprised a 1D convolutional layer with 32 filters and a kernel size of 3 using *ReLU* activation and L2 regularization. This layer captured local patterns. It was followed by a max-pooling layer with a pool size of 2 for dimensionality reduction while retaining essential information. The flattened layer prepared the 3D feature maps for fully connected layers. A dense layer with 64 units and *ReLU* activation captured complex relationships along with applying L2 regularization. To prevent overfitting, a dropout layer with a rate of 0.5 was introduced. Finally, the output layer, suitable for binary classification, had a single unit with *sigmoid* activation and L2 regularization. Concerning *Handing*, using Adam optimizer yielded an accuracy of 95.98%, whereas SGD produced a slightly lower accuracy of 95.24%. With reference to *Pouring*, Adam optimizer provided an accuracy rate of 84.34%, while SGD resulted in a slightly reduced accuracy of 84.76%.

To conclude, Table III presents the deep learning models and their associated performance indicators for *Handing*, while Table IV provides the corresponding information for *Pouring*.

IV. DISCUSSION

In this study, the purpose was to examine how trust impacts the behavioral changes of human operators in industrial settings, as manifested through body motion data analysis. As discussed previously, achieving real-time trust response is essential for tailoring a robot's actions to match the trust level of the human operator. To address this, a data-driven methodology was adopted, harnessing the power of machine learning and deep learning algorithms.

Concerning *machine learning* analysis, the exploration involved the utilization of ensemble models, including Random Forest, XGBoost, and LightGBM. Significantly, these algorithms demonstrated robust performance during the experiments. For comprehensive performance metrics in both the *Handing* and *Pouring* tasks, refer to Table I and Table II, respectively. XGBoost emerged as the top-performing algorithm, achieving the highest accuracy scores in both tasks, notably recording 87.56% accuracy for *Handing* and 92.16% for *Pouring*. XGBoost's superior performance can be attributed to its operation as a gradient boosting algorithm. It systematically builds a sequence of decision trees in a sequential fashion, with each subsequent tree dedicated to correcting the errors made by its predecessors. This iterative approach often leads to improved predictive accuracy, setting it apart from Random Forest, which constructs trees independently. Although LightGBM also harnesses gradient boosting, XGBoost offers distinct advantages, especially in its ability to handle overfitting and optimize the learning process. Subsequently, a Voting Classifier was utilized to combine the predictions generated by the aforementioned models, enhancing the overall classification accuracy and improving the robustness of the results. This ensemble method yielded an accuracy of 87.56% for *Handing* and 91.07% for *Pouring*. Notably, for *Pouring* it demonstrated slightly lower accuracy compared to XGBoost, likely attributed to a minor presence of overfitting.

Within the domain of *deep learning*, the implementation encompassed the utilization of the subsequent models: LSTM, GRU, and 1D-CNN. These algorithms delivered outstanding accuracy results, which are detailed in Table III and Table IV for *Handing* and *Pouring*, respectively. With reference to *Handing*, the highest level of accuracy was achieved through the utilization of the GRU model with the Adam optimizer, specifically 96.30%. In the case of *Pouring*, the highest accuracy, specifically 84.76%, was delivered by the 1D-CNN model using the SGD optimizer. GRU and 1D-CNN demonstrate superior performance compared to LSTM due to their streamlined architectures, computational efficiency, and ability to capture short-term dependencies effectively. These models share the benefit of reduced overfitting risk, faster training times, and adept handling of immediate contextual information. Moreover, the proficiency of 1D-CNN in feature extraction empowers it to excel in identifying intricate local patterns within sequential data, further underscoring its utility in specific applications.

The results underscore the empirical foundation for trust evaluation via data-driven methods. Nevertheless, there remains a need for enhancing trust categorization through sensor fusion techniques. Additionally, to gather valuable feedback, participants were queried about their impressions and potential ways to enhance the spectrum of trust levels. Their responses indicated discomfort when the robot experienced malfunctions, suggesting that simulated noise of robot malfunctions could potentially elicit a wider range of trust

levels. Lastly, participants also underscored the impact of researchers presence and the controlled laboratory environment on their trust responses, potentially veiling more authentic reactions that would occur in real-world settings, such as within an industry environment. The study's limitations primarily revolved around the utilization of predetermined trajectories. Addressing these limitations could involve the implementation of dynamic and unpredictable trajectories, a step that holds the potential to introduce a wider array of challenging and high-risk scenarios for trust examination. In conclusion, as further improvement, adjusting the transparency of robot actions could unveil further nuances in trust responses.

V. CONCLUSION

In this study, a data-driven approach for trust assessment was developed. Body motion data were analyzed as an indicator of how human behavior changes in relation to their trust levels in the robot's performance. The scenario unfolded in a chemical industry context where the robot's responsibilities included tasks such as handing a beaker and mixing chemicals. To uncover potential correlations between body motion data and trust levels, machine learning and deep learning algorithms were utilized. The findings were noteworthy, as machine learning algorithms achieved an accuracy of 87.56% for the Handing task and 92.16% for Pouring when utilizing XGBoost. By comparison, deep learning surpassed expectations, yielding exceptionally impressive results. Specifically, the Handing task reached an outstanding 96.30% accuracy when employing GRU with the Adam optimizer, while Pouring delivered a commendable 84.76% accuracy with 1D-CNN using SGD optimizer. Overall, these findings hold great promise, demonstrating that body motion data is a valuable sensor input for assessing trust levels. Nevertheless, for the purpose of proficiently monitoring and adjusting to trust levels, the strategy will incorporate sensor fusion in upcoming endeavors. This integration will draw data from multiple sensors to offer a more objective measurement, enabling us to fine-tune the robot's behavior to align with the human's trust level, thus fostering a safer environment and a more balanced workload.

REFERENCES

[1] B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[2] K. Hald, M. Rehm, and T. B. Moeslund, "Human-robot trust assessment using top-down visual tracking after robot task execution mistakes," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 892–898.

[3] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[4] Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, L. Wang, Z. Qin, and J. Bao, "Outlook on human-centric manufacturing towards industry 5.0," *Journal of Manufacturing Systems*, vol. 62, pp. 612–627, 2022.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[6] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

[7] A. S. Ghazali, J. Ham, E. I. Barakova, and P. Markopoulos, "Effects of robot facial characteristics and gender in persuasive human-robot interaction," *Frontiers in Robotics and AI*, vol. 5, p. 73, 2018.

[8] C. Li, A. K. Hansen, D. Chrysostomou, S. Bøgh, and O. Madsen, "Bringing a natural language-enabled virtual assistant to industrial mobile robots for learning, training and assistance of manufacturing tasks," in *2022 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2022, pp. 238–243.

[9] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 73–80.

[10] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[11] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, 2016.

[12] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in human-robot interaction*. Elsevier, 2021, pp. 3–25.

[13] M. Rueben, S. A. Elprama, D. Chrysostomou, and A. Jacobs, "Introduction to (re) using questionnaires in human-robot interaction research," *Human-Robot Interaction: Evaluation Methods and Their Standardization*, pp. 125–144, 2020.

[14] B. Leichtmann, V. Nitsch, and M. Mara, "Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement," *Frontiers in Robotics and AI*, vol. 9, p. 838116, 2022.

[15] J. Male and U. Martinez-Hernandez, "Recognition of human activity and the state of an assembly task using vision and inertial sensor fusion methods," in *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, vol. 1. IEEE, 2021, pp. 919–924.

[16] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–6.

[17] G. Campagna and M. Rehm, "Analysis of proximity and risk for trust evaluation in human-robot collaboration," in *32nd IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2023.

[18] D. Roetenberg, H. Luinge, P. Slycke, *et al.*, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," *Xsens Motion Technologies BV, Tech. Rep*, vol. 1, pp. 1–7, 2009.

[19] M. Schepers, M. Giuberti, G. Bellusci, *et al.*, "Xsens mvn: Consistent tracking of human motion using inertial sensing," *Xsens Technol*, vol. 1, no. 8, pp. 1–8, 2018.