IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 June 2022, accepted 14 July 2022, date of publication 25 July 2022, date of current version 29 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193694

RESEARCH ARTICLE

Real-Time Team Performance and Workload Prediction From Voice Communications

CATHERINE SANDOVAL[1], (Member, IEEE), MELISSA N. STOLAR[2], SIMON G. HOSKING[2], DAWEI JIA[2], AND MARGARET LECH[1], (Member, IEEE)

[1]School of Engineering, RMIT University, Melbourne, VIC 3000, Australia
[2]Defence Science and Technology Group, Fishermans Bend, Port Melbourne VIC 3207, Australia

Corresponding author: Catherine Sandoval (catherine.sandoval.rodriguez@rmit.edu.au)

This research was supported by a Defence Science and Technology Group Research Agreement and Defence Science Institute Postgraduate Scholarship.

ABSTRACT Automatic prediction of team performance and workload plays a crucial role in team selection, training, evaluation, and re-training processes. This study investigated the potential of using voice analysis of team-based communication for predicting team workload (TW) and team performance (TP). Both the TW and TP categories were labeled objectively. Ten teams of three participants were tasked with completing a computer-based command-and-control simulation that required communication of task-specific information to each team member. Recordings of each participant's voice communications were used to train Convolution Neural Network (CNN) models for each team separately. It was hypothesized that integrating TW and TP information into the prediction process would support the prediction of both TW and TP categories. Two experiments were conducted. In the first experiment, the TP prediction networks were fine-tuned to predict TW, and conversely, the TW prediction networks were fine-tuned to predict TP. In the second experiment, the TP or TW prediction based on the assembly of interconnected TP and TW classifiers was tested. Both experiments confirmed the hypothesis. It was shown that task-related pre-requisite knowledge embedded into the neural network reduced neural network model training time and improved performance without increasing the training data size. Predictions based on combined TW and TP classification outcomes (using either separate or interconnected TW or TP classifiers) outperformed the baseline method using a single CNN model trained to predict either TW or TP alone. The classification accuracy was consistent with previously reported cognitive load prediction based on objective measures.

INDEX TERMS Speech classification, voice classification, cognitive load prediction, workload prediction, performance prediction, team performance monitoring, team training, transfer learning, deep learning.

I. INTRODUCTION

The level of mental effort required by an individual when performing a given task indicates cognitive workload [1]. The excessive workload can significantly impact cognitive and attentional resources, leading to reduced performance and learning ability, and decreases in information processing and decision-making ability. The link between Cognitive Load (CL) and task performance has been extensively investigated, leading to significant theoretical developments [2], [3]. Due to the recent development of efficient numerical techniques, there is renewed interest in modeling and predicting CL, such that the relationship between workload and performance can be validated through numerical simulations of theoretical models. For example, computational analysis of objective indicators, such as bio-signals and audio-visual recordings, can be used to predict human physiological and mental states [4]–[6]. This approach provides new empirical insights into the dependencies and mechanisms underlying the relationship between CL and human performance. A steadily growing body of work proposes different experimental conditions where CL can be controlled

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

and measured using objective or subjective criteria. Similarly, links between CL and related factors such as interpersonal trust and familiarity [7] can be investigated.

The current study provides twofold contributions to this line of research. Firstly, two new computationally efficient machine learning approaches (a pre-trained Convolution Neural Network (CNN) model and a two-channel decision-making system) to the prediction of the team workload (TW) and team performance (TP) from the speech are proposed and experimentally validated. Secondly, the research hypothesis assuming learning dependencies between TW and TP prediction models is tested.

The remaining parts of this paper are organized as follows: Section II presents a summary of related works. Section III explains the proposed methodology. The validation dataset is described in Section IV. Section V includes experimental results and discussion. Section VI describes the demonstration video showing an example of real-time TW prediction, and Section VII concludes the paper.

II. RELATED WORK

Psychology and management theory have extensively examined the relationship between workload and performance [8]–[11]. A wide range of predictor factors such as age and gender [2], type of personality [3], number of working hours [12], employment status [13], or other working conditions influencing job performance [11] have been investigated. In [8], the effect of workload on the performance of individuals, and the relationship between workload and the quantity and quality of output, were examined using statistical modeling. It was observed that as the workload increased, the output quantity of employees increased only up to a certain point, after which it decreased. Factors affecting the prediction of mental workload in both single- and multitask environments were investigated in [14] using a conceptual model. It was found that both task-related factors and individual-related factors could significantly affect mental workload. A theoretical model was proposed in [15] to predict air traffic controller workload. The model included various airspace factors and operational constraints. Although task demand showed a strong relationship with the workload, the primary factor affecting workload was the capacity of the controller to select priorities, manage their cognitive resources, and regulate their performance. It was concluded that effective modeling of workload must be supported by modeling controllers' strategies for regulating the cognitive impact of task demand. The concept of information entropy was used in [16] to predict the mental workload of an urban railway dispatcher. The study proposed a comprehensive mental workload prediction model combining various information factors (information display duration and the amount of information) that could influence mental workload. Experimental validation indicated a strong correlation between the model's theoretical predictions and experimental results.

Significant research has also been carried out in the area of performance prediction. A review of existing theoretical and computational models of human performance can be found in [17]. The study identified three general factors that contribute to performance: tasks, cognition, and vision models. In [9], [10], performance models under stress conditions were discussed. Mathematical models for predicting human performance from the discrete wavelet transform (DWT) of event-related potentials (ERP) elicited in the EEG signals by task-relevant stimuli were analyzed in [18]. Neural network models trained on DWT features provided evidence that a pattern of low-frequency activity (1 to 3.5 Hz) occurring at specific times and scalp locations is a reliable correlate of human performance. In [19], a deep Recurrent Neural Network (RNN) was applied to model and predict human performance in target selection from a vertical list or a menu displayed on a computer screen. Various model architectures were analyzed, and potential extensions were proposed.

Over the last decade, significant progress has been achieved through the development of machine learning (ML). The ML models are given as large sets of equations with the parameters optimized to provide the closest approximation of the experimental training data. The prediction outcomes can be given either as continuous numerical values or vectors of categorical probabilities. This type of modeling is particularly suitable for continuous, real-time monitoring of mental workload and performance prediction. The information can be supplied to the model in the form of streaming biometric signals such as EEG, ECG, pulse, speech, or video images. EEG signals have been explored particularly extensively in mental workload prediction. For example, [20] examined the workload prediction from EEG in Human-Computer Interactions (HCI). Shallow and deep CNN models were compared. A shallow CNN was shown to provide the best performance. Functional Near-Infrared Spectroscopy (fNIRS) was applied in [21] to classify mental workload during HCI tasks. Three machine-learning models were compared: Logistic Regression (LR), Support Vector Machine (SVM), and CNN. When identifying two workload categories, CNN outperformed LR and SVM. It was also reported that the person-dependent modeling outperformed a person-independent approach.

The increasing ubiquity of speech-based communication is a rich source of physiological and mental state information that affords an opportunity for automatic real-time detection of a speaker's workload. Speech classification has been used to predict workload [22], [23] and inter-personal trust and familiarity of team members [7]. In [23], a person-dependent approach based on two subjectively labeled workload categories was investigated using an SVM algorithm and a conventional set of low-level acoustic speech parameters. The outcomes of this study were improved in [22] by replacing the traditional acoustic speech features with amplitude spectrograms.

The current study continues the speech-based line of research presented in [22], [23]. However, objective rather than subjective labeling is investigated to perform workload and performance prediction. The SVM was replaced by a pre-trained CNN model or a two-channel decision-making

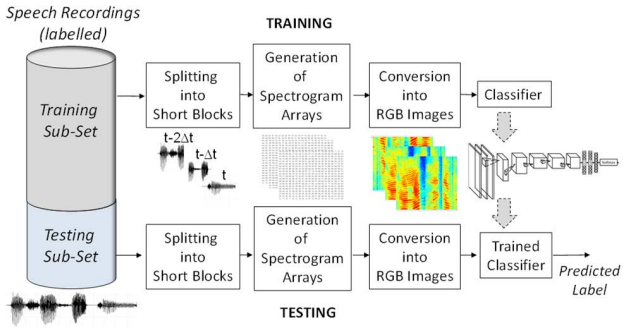

FIGURE 1. Prediction framework for a single classifier.

system. In addition, the concept of reinforcing the prediction outcomes by combining mental workload information with performance information was examined.

III. METHODOLOGY

A. RESEARCH HYPOTHESIS

Based on the known link between team performance and team workload [2], [3], [8]–[13], it was hypothesized that combining the TW and TP information either within a single classifier model or a distributed multichannel classification system could improve the prediction of different TP and TW states. Given the speech dataset described in Section IV, an experimental framework for the TP and TW prediction was designed to test the hypothesis.

B. PREDICTION FRAMEWORK

In the baseline approach shown in Fig. 1, a single classifier was trained either to categorize team performance (TP classifier) or team workload (TW classifier). In both cases, a convolutional network model (CNN) pre-trained to categorize RGB images converted from speech recordings was applied.

In the multichannel decision-making approach shown in Fig. 2, two parallel channels consisting of baseline classifiers were employed to make the respective TP or TW classification based directly on the speech data. The outcomes of both channels, given as probability vectors for TP and TW, were concatenated and fed into two final decision-making classifiers, one for TP and one for TW.

The baseline classifiers made their predictions based on speech analysis using only single-domain information (either TP or TW). At the same time, the final decision-making classifiers provided the same type of classification as the baseline classifiers; however, their decision was based on metadata (probabilities) combining two different information domains (TP and TW). Therefore, it was expected that the outcomes from the decision-making classifiers should be, on average, more accurate than from the baseline classifiers. The baseline classifiers were given as pre-trained CNN structures. In contrast, the decision-making classifiers were given as simple feed-forward perceptron neural networks, each with two hidden layers and 100 nodes per layer.

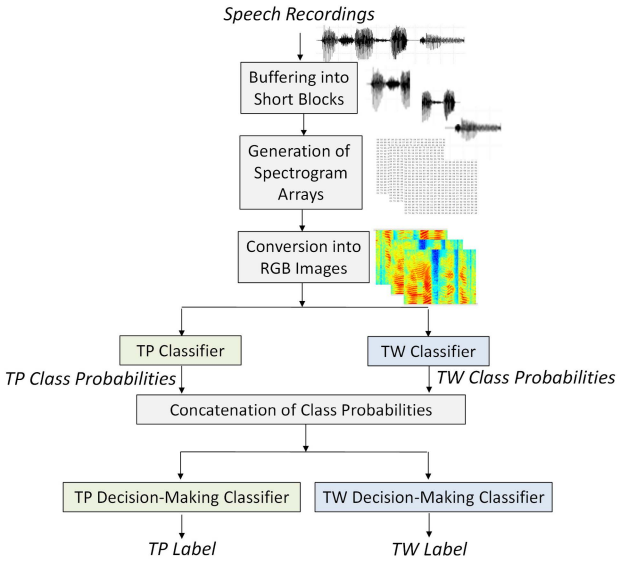

FIGURE 2. Prediction framework for a multichannel decision-making system.

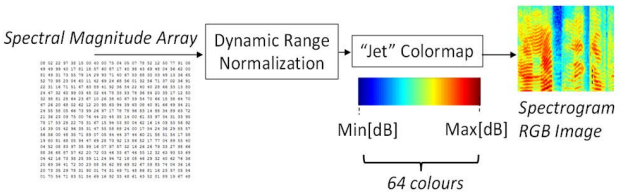

FIGURE 3. Conversion of the speech samples into RGB images of speech spectrograms.

C. CONVERSION FROM TIME WAVEFORMS TO SPECTROGRAM IMAGES

Since the baseline speech classification was performed using CNNs pre-trained to classify images, the speech waveforms were first transformed into images and then passed to CNNs. The TP and TW classifier inputs in the baseline system (Fig. 1) and the multichannel system (Fig. 2) were given as RGB images representing speech amplitude spectrograms. As shown in Fig. 1, speech time waveforms were divided into 1-second blocks ($\Delta t = 1$ second). The time shift between subsequent blocks was 0.1 seconds. For each block, an amplitude spectrogram array was calculated, normalized, and transformed into an RGB image format using the MATLAB "jet" 64-colour map (Fig. 3). The normalization was based on the average min and max values of the spectrogram arrays calculated over the whole database.

While the time axis of the spectrogram arrays was represented on the linear scale, the vertical frequency axis used the logarithmic scale. Two examples of the same speech spectrogram, one with linear and one with a logarithmic frequency scale, are illustrated in Fig. 4. It can be observed that the logarithmic scale reveals more details of the low-frequency range where the fundamental frequency and the first formant information are located. The logarithmic scale

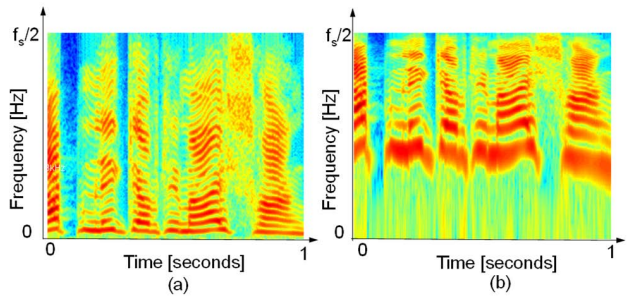

FIGURE 4. Examples of RGB images of speech spectrograms for the same speech sample: (a) linear frequency scale, (b) logarithmic frequency scale.

was shown to be most effective in related speech classification tasks [22], [24], [25].

D. CONCATENATION OF PROBABILITY VECTORS

The parallel classification channels in Fig. 2 generated soft probability vectors indicating the probability that the input image $i, (i = 1, \ldots, N)$ represents a given data category, where N denotes the total number of input images. The TP channel was trained to identify N_{TP} team performance categories, whereas the TW channel was trained to identify N_{TW} team workload categories. Assigning the index 1 to the TP channel and 2 to the TW channel, the class probability vectors for the *ith* image generated by the TP and the TW channels were given by (1) and (2), respectively:

$$P_{i,1} = \left[p_{i,1,1}, p_{i,1,2}, \ldots, p_{i,1,N_{TP}}\right] \quad (1)$$

$$\mathrm{P}_{i,2} = \left[p_{i,2,1}, p_{i,2,2}, \ldots, p_{i,2,N_{TW}}\right] \quad (2)$$

These vectors were then concatenated to generate combined probability vectors C_i given as

$$C_i = \left[p_{i,1,1}, p_{i,1,2,\ldots,}p_{i,1,N_{TP}}, p_{i,2,1}, p_{i,2,2}, \ldots, p_{i,2,N_{TW}}\right] \quad (3)$$

As shown in Fig. 2, the C_i vectors were applied as features in the process of training two decision-making classifiers, one to identify team performance and the other to identify team workload.

E. PERFORMANCE MEASURES

Given N data categories, the classification accuracy Acc_i for a given category $i\,(i = 1, \ldots, N)$ was calculated as,

$$Acc_i = \frac{t_{pi} + t_{ni}}{t_{pi} + t_{ni} + f_{pi} + f_{ni}} \quad (4)$$

where, t_{pi} and t_{ni} denoted numbers of true-positive and true-negative classification outcomes for category i, respectively, f_{pi} and f_{ni} were the numbers of false-positive and false-negative classification outcomes for category i, respectively. The average classification outcome was estimated as an average across all categories using (5).

$$Acc = \frac{1}{N}\sum_{i=1}^{N} Acc_i \quad (5)$$

Due to the unbalanced training data, we used the *F1* score to determine if the model gives a uniform accuracy distribution across categories. It was estimated using (6).

$$F1 = \frac{2 * recall * prcision}{recall + precision} \quad (6)$$

The recall and precision values were calculated using (7) and (8), respectively.

$$recall = \frac{TP}{TP + FN} \quad (7)$$

$$precision = \frac{TP}{TP + FP} \quad (8)$$

The ultimate goal was to achieve high values of both accuracy and *F1* scores.

IV. VALIDATION DATASET

The validation dataset included recordings of communications transmitted between team members while completing a low-fidelity simulated command-and-control simulation. The recordings were made within a controlled training environment. Ten teams consisted of three participants (with no military experience), with each team member randomly allocated to one of three roles. Each role had sole access to specific information sources necessary to successfully complete the task (see below). Team members communicated via a push-to-talk two-way radio (for voice reception and transmission). Each team completed three sessions per week for four consecutive weeks, with each session held on a different day during the week. Trials consisted of 24 waves of aircraft entering the airspace every 60 seconds. Team workload was determined by the number of aircraft in each wave. The low workload condition consisted of six aircraft per wave; medium workload had seven aircraft per wave, and high workload had eight aircraft per wave. Each session consisted of two trials in the same workload condition. Hence, each team completed a total of 24 trials over the four weeks of testing.

In each wave of aircraft, there was a single target aircraft. The task for each team was to identify and destroy all target aircraft within the trial time limit of about 25 minutes. Each of the team members was allocated to a different role that determined their view of the airspace and the sources of information available for identifying target aircraft as follows:

(i) Participants assigned to the role of Surveillance Operator (SO) had visual access to the entire air space and could interrogate point of origin and route information for each aircraft for comparing aircraft movements with approved routes.

(ii) Participants assigned to the role of Tactical Director (TD) had visual access to a controlled airspace only (simulating a reduced sensor capability) and could only interrogate aircraft within the controlled airspace to determine speed and altitude for prioritizing the threat level of each aircraft.

(iii) Participants assigned to the role of Fighter Controller (FC) had visual access to an engagement zone

TABLE 1. Distribution of spectrogram images across teams for two TP categories, LTP and HTP.

No. Imgs	Team									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Total	102476	143393	110362	146914	112337	113479	117948	146466	158355	93631
LTP	65304	72539	55821	77993	65907	62035	74813	90584	89961	53985
HTP	37172	70854	54541	68921	46430	51444	43127	55882	68394	39646

TABLE 2. Distribution of spectrogram images across teams for three TW categories LTW, MTW, and HTW.

No. Imgs	Team									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Total	102476	143393	110362	146914	112337	113479	117948	146466	158355	93631
LTW	31392	39825	33755	42297	35097	28993	33122	42982	46734	24533
MTW	35584	47541	40166	50904	37123	38443	40656	48657	60077	31286
HTW	35500	56027	36441	53713	40117	46043	44162	54827	51544	37812

within the controlled airspace only, consisting of two small-diameter circles (simulating the sensor capability of fighter jets) that could be directed towards aircraft to interrogate the ''identification, friend, or foe'' (IFF) and aircraft maneuver information.

If the information integrated across the three team members indicated that the ''rules of engagement'' were met for a particular aircraft, then it would be classified as a target aircraft, and the fighter aircraft would engage a simulated missile launch to destroy it.

Team performance during each session was monitored by recording the time needed to identify and destroy each target correctly (trial-time). A median trial-time value of 104 seconds was estimated from a histogram of trial times across all teams and training sessions. Trials that did not eventuate with a correct target engagement were labeled ''not hit.'' Spectrogram images corresponding to speech recorded during waves that destroyed targets after a time shorter than the median were labeled as high TP (HTP). Spectrogram images corresponding to speech recorded during waves that either destroyed the target aircraft after a time longer than 104 seconds or those that were marked as ''not hit'' were labeled as low TP (LTP). The numbers of generated images varied slightly across teams and TP categories (Table 1). The speech spectrograms were labeled into three TW categories: low team workload (LTW), medium team workload (MTW), and high team workload (HTW). Table 2 shows the distribution of spectrogram images representing each of the three TW categories.

Generally, Tables 1 and 2 indicate that the numbers of images representing TP and TW categories were only slightly imbalanced. It should be noted that the TP and TW labels used in this study were determined objectively, either by the experimental outcomes (TP) or by the experimental conditions (TW). No subjective assessment labels for the TP or TW were used.

TABLE 3. Average accuracy (%) of TP and TW prediction using a single classifier and a single transfer learning (baseline approach).

	Team									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
TP: Mean 63.02 Ste 1.11 Median 63.29 Std 3.7 Min 57.47 Max 68.16										
Acc	66.85	58.50	57.47	61.42	65.16	60.29	65.88	68.16	65.64	60.85
F1	65.47	58.65	57.63	61.09	64.78	59.81	64.75	66.91	65.20	60.47
TW: Mean 58.23 Ste 1.35 Median 56.21 Std 4.2 Min 54.54 Max 66.47										
Acc	66.47	60.02	55.78	54.54	54.80	55.71	56.20	65.12	57.53	56.22
F1	66.05	59.68	55.58	54.08	54.58	55.41	55.88	64.68	57.31	55.94

V. EXPERIMENTAL RESULTS AND DISCUSSION

Classification experiments were conducted to test the research hypothesis and create a comparison baseline. A 3-fold cross-validation technique was adopted for each experiment, with 80% of the data distribution used for training and 20% used for testing the classifier. The testing data were never used during the training procedures. The *Acc* values defined in (4) were calculated during the testing procedure using only ''unseen'' data and averaged over three tests.

A. TP AND TW PREDICTION USING SINGLE TRANSFER LEARNING (BASELINE)

A baseline approach was tested in this experiment to provide a reference point for comparison. Single transfer learning, as shown in Fig. 1, was applied to fine-tune a CNN given as a pre-trained ResNet-18 network [26]. The ResNet-18 was pre-trained on a vast number of images to perform a general image object classification task. In this experiment, it was fine-tuned to complete the task of either TP or TW classification. Due to the extensive pre-training contained within the ResNet-18 network, the fine-tuning required only a relatively small set of training data labeled with TP or TW states compared to what would be required if the model was trained ''from scratch.'' This type of training is hereafter referred to as single transfer learning. Separate ResNet-18 models were trained for each team to predict either TP or TW.

The average prediction accuracies (*Acc*) for each team are presented in Table 3 alongside statistical functions of the *Acc* calculated across all teams. The results varied from team to team; however, the standard deviation was only 3.7% for the TP and 4.2% for the TW prediction. In general, the TP prediction outcomes were slightly higher (57.47% - 68.16%) than the TW outcomes (54.54% - 66.47%). However, this was expected, as there were only two TP categories as opposed to three TW categories.

Since the achieved level of accuracy was, in both cases, too low to be used in practical applications, there was a clear need for improvement. Therefore, the following two experiments (Section V-B and Section V-C) investigated if providing additional information to the network could improve the prediction of TP and TW without increasing the training data set or the complexity of the CNN models.

TABLE 4. Average accuracy (%) of TP and TW prediction using a single classifier and a double transfer learning.

	Team									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
TP: Mean 64.40 Ste 1.17 Median 64.53 Std 3.7 Min 58.79 Max 69.44										
Acc	68.27	59.87	58.80	62.66	66.41	62.01	67.01	69.44	67.21	62.40
F1	67.09	59.95	58.98	62.30	66.01	61.54	65.93	68.37	66.75	62.03
TW: Mean 59.59 Ste 1.35 Median 57.68 Std 4.28 Min 55.96 Max 67.79										
Acc	67.79	61.45	56.91	55.96	56.16	57.03	57.90	66.52	58.75	57.47
F1	67.12	61.09	56.33	55.32	55.95	56.75	57.62	66.04	58.12	57.01

TABLE 5. Average accuracy (%) of TP and TW prediction using the multichannel decision-making system.

	Team									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
TP: Mean 64.69 Ste 1.2 Median 64.98 Std 3.82 Min 59.23 Max 69.95										
Acc	68.58	59.89	59.23	63.05	66.92	62.07	67.55	69.95	67.45	62.28
F1	67.57	59.98	59.33	62.79	66.60	61.69	66.56	68.93	67.03	61.91
TW: Mean 60.07 Ste 1.35 Median 58.12 Std 4.29 Min 56.46 Max 68.29										
Acc	68.29	61.78	57.61	56.46	56.66	57.43	58.18	67.09	59.15	58.07
F1	67.83	61.16	57.13	56.05	56.47	57.17	57.92	66.67	58.96	57.83

B. TP AND TW PREDICTION USING DOUBLE TRANSFER LEARNING

In this experiment, the research hypothesis formulated in Section III-A was tested by integrating the TP and TW information within a single CNN using the process of double transfer learning. The ResNet-18 model pre-trained to recognize image objects was first fine-tuned to predict TP (first transfer learning) and then further fine-tuned to predict TW (second transfer learning). Similarly, The ResNet-18 model pre-trained to recognize image objects was first fine-tuned to predict TW and then fine-tuned to predict TP. This approach is hereafter referred to as double transfer learning.

The outcomes of this experiment are given in Table 4. An improvement could be observed compared to the baseline results (Table 3). Again, the accuracy varied from team to team, but the standard deviation remained small (3.7% for TP and 4.28% for TW). The TP prediction accuracy was 58.79% - 69.44% showing 1.28% - 1.32% improvement with respect to the baseline. At the same time, the TW prediction accuracy range was 55.96% - 67.79%, with a 1.32% – 1.42% increase compared to the baseline.

These results provided our first empirical support for the research hypothesis by confirming that pre-requisite TW information embedded into the network facilitated better TP prediction outcomes. Similarly, pre-requisite TP information embedded into the network facilitated better TW prediction outcomes.

C. TP AND TW PREDICTION USING MULTI-CHANNEL DECISION-MAKING SYSTEM

In the third experiment, the research hypothesis from Section III-A was tested once again, but this time the TP and TW information was integrated into the prediction model using the two-channel decision-making system illustrated in Fig. 2. The input speech spectrogram was first passed to two separate ResNet-18 models, one trained to predict TP and the other to predict TW (using only a single transfer learning). The prediction outcomes (probability vectors) from both classifiers were concatenated and conveyed to the second stage of classification consisting of two shallow NNs, one trained to predict TP and one to predict TW. Although each NN was making a different type of prediction (TP or TW), both networks were provided with combined TP and TW information. The outcomes of this experiment are presented in Table 5.

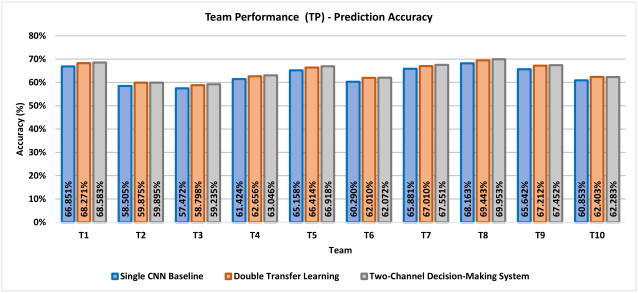

FIGURE 5. TP prediction accuracy (%) for each team using three different TP prediction methods.

A further improvement over the baseline results (in Table 3) and the double transfer learning results (in Table 4) could be observed. Model accuracy still varied across teams only by a small amount, with the standard deviation of 3.82% for TP and 4.29% for TW. The TP prediction accuracy was 67.45% - 62.28% showing 1.76% - 1.79% improvement with respect to the baseline. The TW prediction accuracy, on the other hand, was 56.46% - 68.29% denoting a 1.82% - 1.92% increase compared to the baseline.

The outcomes of this experiment provided our second empirical support for the research hypothesis by confirming that simultaneous TP and TW information can be used to make more accurate TP and TW predictions.

D. COMPARISON BETWEEN PREDICTION METHODS

The experimental work described in previous sections created a classification baseline for the prediction of TP and TW. Two different methods of improving the baseline outcomes were tested. The TP and TW information was combined within a single CNN model via double transfer learning in the first approach. The second approach achieved integration within a multichannel classification system. Fig. 5 and Fig. 6 compare the TP and TW prediction, respectively, between all tested scenarios across all teams.

Consistent differences as a function of method-dependent performance can be observed across all teams. The baseline method was always the worst performer. At the same time, the best results were given by the two-channel decision-making system, which outperformed the baseline and the

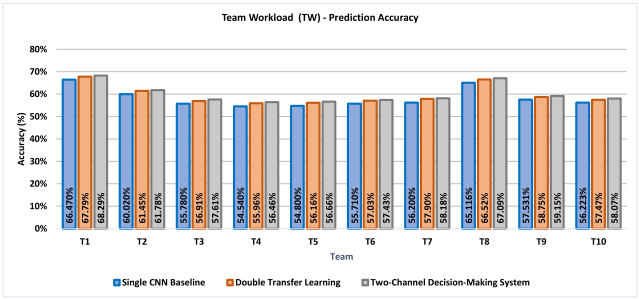

FIGURE 6. TW prediction accuracy (%) for each team using three different TW prediction methods.

double transfer learning approaches. In comparison with the baseline, the improvement in performance for the two-channel decision-making system was about 1.3% for TP and 1.4% for TW; in the case of the double transfer learning, the improvement was only 0.5%.

E. COMPARISON WITH RELATED STUDIES

As shown in Table 6, the results reported in this study are consistent with previously reported speech classification based on objective measures [22]. However, most previous studies applied subjective labels to categorize either cognitive load [21]–[23] or interpersonal trust [7]. This could potentially be due to the fact that the injection of objective workload ratings into the task increases workload due to having to complete an extra sub-task during the experiments. In [22], for example, objective labels of cognitive load were based on the number of aircraft being controlled (8, 12, or 16 aircraft). The results were less accurate than those obtained when cognitive load classifications were based on subjective labels. Therefore, the follow-up study described in [23] focused primarily on subjective labeling. Apart from undertaking the more challenging objective labeling option, the current study used about five times fewer spectrogram images compared to the numbers used to predict interpersonal trust in [7] or cognitive load in [22]. This was due to the very sparse speech communications between team members. In addition, the classification presented in the current study was not person-dependent but rather team-dependent. As shown in [22]–[23], the best classification outcomes were achieved when the classifiers were trained using person-dependent methods, and a significant reduction was observed when the training was done in a person-independent fashion. Further consistency with previous studies was shown for double transfer learning. Similar to [7], where the pre-requisite team familiarity information supported better outcomes for trust recognition and vice versa, the current study observed a dependency between team workload and team performance.

VI. REAL-TIME TW PREDICTION VIDEO

The current study is accompanied by a video demonstrating an example of real-time TW prediction. Speech recordings of a single team were processed block-by-block and

TABLE 6. Classification accuracy – A comparison with related studies.

Authors	Prediction Task	Person-dependent.	Labels		Acc (%)
			Objec.	Subjec.	
Vukovic et al. [23]	Cognitive load (speech, 2 categ.)	Yes	No	Yes	79.7
Vukovic et al. [22]	Cognitive Load (speech, 3 categ.)	Yes	No	Yes	83.7
Vukovic et al. [22]	Cognitive Load (speech, 3 categ.)	Yes	Yes	No	65.7
Sandoval et al. [7]	Interpersonal Trust (speech, 2 categ.)	No	No	Yes	89.0
Appriou et al. [20]	Mental Workload (EEG, 2 categ.)	Yes	No	No	72.7
Benerradi et al. [21]	Mental Workload (fNIRS, 3 categ.)	Yes	No	Yes	48.5
Benerradi et al. [21]	Mental Workload (fNIRS, 3 categ.)	No	No	Yes	50.9
Current Study	Team Performance (speech, 2 categ.)	No	Yes	No	**69.95**
Current Study	Team Workload (speech, 3 categ.)	No	Yes	No	**68.29**

passed through the two-stage classification system trained as described in Section V-C. A classification label indicating one of the three levels of workload (low, medium, and high) was generated block by block as the speech was streaming through the classifier. The block duration was 1 second; the average computational times needed to process each block and generate the TW label are presented in Table 7. The computations were performed using MATLAB 2020b on an HP Z8 G4 Workstation with the Intel Xeon Silver CPU and 128 GB RAM. The computational time was determined using MATLAB function *timeit* [27]. For a given 1-second block, the TW label was generated within 47 to 50 milliseconds. The time needed for the inference process was about 11 milliseconds, and it was longer than the total time required to generate the features (about 3.2 milliseconds).

The longest processing time (approximately 34 milliseconds) during the feature generation stage was needed to calculate the magnitude spectrogram arrays, whereas the time required to convert these arrays to RGB images was only 3.2 milliseconds. Fig. 7 shows a screenshot of the graphical user interface (GUI) used in our demonstration video. It allows choosing the streaming speech source and the team model. The top two display windows show the streaming speech spectrogram (left) and the time waveform (right). The bottom window displays the instantaneous prediction results in the form of probability bars for low, medium, and high

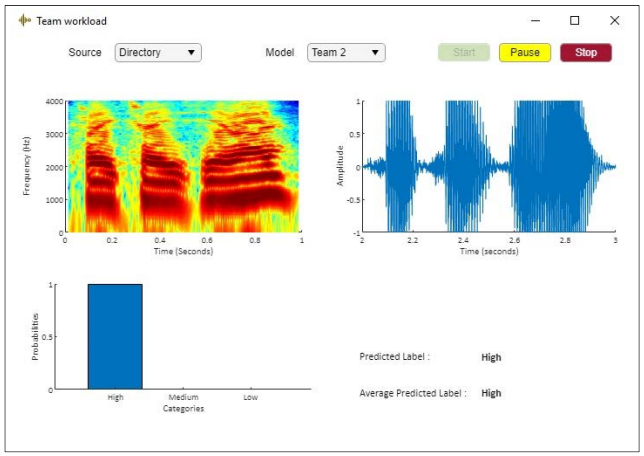

FIGURE 7. A GUI screenshot for the real-time TW prediction demonstration.

TABLE 7. Real-time TW prediction - Computational Time needed to generate a TW label for a single 1-second block of speech.

Spectral Magnitude [ms]	Conversion into RGB Images [ms]	Inference [ms]	Total [ms]
11	3.2	34	48.2

workload, as well as the prediction label for the current block and the average prediction outcome.

VII. CONCLUSION

This study addressed the tasks of TP and TW prediction based on objective labels of speech. It was hypothesized that the addition of related pre-requisite knowledge to the prediction model could improve classification outcomes for the task at hand without the need to increase the size of the training data or the complexity of the classification model. Experiments were conducted to test the research hypothesis by integrating the TP and TW information either within a single CNN model via double transfer learning or within a multichannel decision-making system of parallel TW and TP classifiers. In both cases, an improvement in prediction accuracy for the TP and TW states was observed, confirming the hypothesis. The study introduced an efficient novel multichannel TP and TW prediction method, outperforming the single CNN baseline classification by 1.76% for TP and 1.79% for TW. The research outcomes were consistent with previous related studies in two aspects. Firstly, the classification accuracy was comparable with previous outcomes of speech classification based on objective labels outperforming [16], [3] by 2.5%. Secondly, the classification improvements achieved by double transfer learning of TP and TW were consistent with previously described improvements achieved when predicting interpersonal trust and familiarity [4]. Future research will include both objective and subjective labels. Other information source modalities such as ECG and EEG will be added to speech to support the team performance and workload monitoring process.

REFERENCES

[1] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learn. Instruct.*, vol. 4, no. 4, pp. 295–312, Jan. 1994, doi: 10.1016/0959-4752(94)90003-5.

[2] D. A. Waldman and B. J. Avolio, "A meta-analysis of age differences in job performance," *J. Appl. Psychol.*, vol. 71, no. 1, p. 33, Feb. 1986, doi: 10.1037/0021-9010.71.1.33.

[3] D. V. Day and S. B. Silverman, "Personality and job performance: Evidence of incremental validity," *Personnel Psychol.*, vol. 42, no. 1, pp. 25–36, Mar. 1989, doi: 10.1111/j.1744-6570.1989.tb01549.x.

[4] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2018, doi: 10.1109/ACCESS.2018.2883213.

[5] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Appl. Sci.*, vol. 10, no. 11, p. 3843, May 2020, doi: 10.3390/app10113843.

[6] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 1, pp. 60–70, Jan. 2022, doi: 10.1109/MAES.2021.3115198.

[7] C. S. Rodriguez, A. R. Panganiban, M. N. Stolar, R. S. Bolia, and M. Lech, "Prediction of inter-personal trust and team familiarity from speech: A double transfer learning approach," *IEEE Access*, vol. 8, pp. 225437–225447, 2020, doi: 10.1109/ACCESS.2020.3044285.

[8] A. Bruggen, "An empirical investigation of the relationship between workload and performance," *Manage. Decis.*, vol. 53, no. 10, pp. 2377–2389, Nov. 2015, doi: 10.1108/MD-02-2015-0063.

[9] S. M. Jex, *Stress and job performance: Theory, research, and implications for managerial practice*. Thousand Oaks, CA, USA: Sage, 1998, p. 143.

[10] J. A. Lepine, N. P. Podsakoff, and M. A. Lepine, "A meta-analytic test of the challenge stressor–hindrance stressor framework: An explanation for inconsistent relationships among stressors and performance," *Acad. Manage. J.*, vol. 48, no. 5, pp. 764–775, Oct. 2005, doi: 10.5465/amj.2005.18803921.

[11] E. Kahya, "The effects of job characteristics and working conditions on job performance," *Int. J. Ind. Ergonom.*, vol. 37, no. 6, pp. 515–523, Jun. 2007, doi: 10.1016/j.ergon.2007.02.006.

[12] E. Shepard and T. Clifton, "Are longer hours reducing productivity in manufacturing?" *Int. J. Manpower*, vol. 21, no. 7, pp. 540–553, Nov. 2000, doi: 10.1108/01437720010378999.

[13] C.-I. Chu and Y.-F. Hsu, "Hospital nurse job attitudes and performance: The impact of employment status," *J. Nursing Res.*, vol. 19, no. 1, pp. 53–60, Mar. 2011, doi: 10.1097/JNR.0b013e31820beba9.

[14] B. Xie and G. Salvendy, "Prediction of mental workload in single and multiple tasks environments," *Int. J. Cognit. Ergonom.*, vol. 4, no. 3, pp. 213–242, Sep. 2000, doi: 10.1207/S15327566IJCE0403_3.

[15] S. Loft, "Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications," *Hum. Factors*, vol. 49, no. 3, pp. 376–399, 2007, Jun. 2007, doi: 10.1518/001872007X197017.

[16] X. Li, W. Fang, and Y. Zhou, "Mental workload prediction model based on information entropy," *Comput. Assist. Surg.*, vol. 21, no. 1, pp. 116–123, Oct. 2016, doi: 10.1080/24699322.2016.1240298.

[17] K. Leiden, K. R. Laughery, J. Keller, J. French, W. Warwick, and S. D. Wood, *A Review of Human Performance Models for the Prediction of Human Error*. Moffett Field, CA, USA: Ames Research Center NASA, 2001, p. 125.

[18] L. J. Trejo and M. J. Shensa, "Feature extraction of event-related potentials using wavelets: An application to human performance monitoring," *Brain Lang.*, vol. 66, no. 1, pp. 89–107, Jan. 1999, doi: 10.1006/brln.1998.2026.

[19] A. Yuan, K. Pfeuffer, and Y. Li, "Human performance modeling with deep learning," in *Human-Computer Interaction Series, Artificial Intelligence for Human Computer Interaction: A Modern Approach*. Cham, Switzerland: Springer, 2021, pp. 3–31, doi: 10.1007/978-3-030-82681-9_1.

[20] A. Appriou, A. Cichocki, and F. Lotte, "Towards robust neuroadaptive HCI: Exploring modern machine learning methods to estimate mental workload from EEG signals," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–6, doi: 10.1145/3170427.3188617.

[21] J. Benerradi, H. A. Maior, A. Marinescu, J. Clos, and M. L. Wilson, "Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks," in *Proc. Halfway Future Symp.*, Nov. 2019, pp. 1–11, doi: 10.1145/3363384.3363392.
[22] M. Vukovic, M. Stolar, and M. Lech, "Cognitive load estimation from speech commands to simulated aircraft," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1011–1022, 2021, doi: 10.1109/TASLP.2021.3057492.
[23] M. Vukovic, V. Sethu, J. Parker, L. Cavedon, M. Lech, and J. Thangarajah, "Estimating cognitive load from speech gathered in a complex real-life training exercise," *Int. J. Hum.-Comput. Stud.*, vol. 124, pp. 116–133, Apr. 2019, doi: 10.1016/j.ijhcs.2018.12.003.
[24] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017, doi: 10.1016/j.neunet.2017.02.013.
[25] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers Comput. Sci.*, vol. 2, p. 14, May 2020, doi: 10.3389/fcomp.2020.00014.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90. [Online]. Available: https://ieeexplore.ieee.org/document/7780459/
[27] B. McKeeman. *MATLAB Performance Measurement*. Accessed: Oct. 25, 2021. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/18510-matlab-performance-measurement

CATHERINE SANDOVAL (Member, IEEE) received the B.E. degree in electronic engineering from Javeriana University, Bogota, Colombia, in 2001, and the M.S. degree in electronic engineering and the Ph.D. degree in engineering from the Royal Melbourne Institute of Technology University (RMIT), Melbourne, Australia, in 2015 and 2021, respectively. Currently, she is working as a Research Assistant with the School of Engineering, RMIT University. Her research interests include machine learning, deep learning, digital humanities, and image processing.

MELISSA N. STOLAR received the B.E. degree in science and electronic engineering and the Ph.D. degree in electrical engineering from the Royal Melbourne Institute of Technology University (RMIT), Melbourne, Australia, in 2013 and 2017, respectively. From 2016 to 2019, she was a Research Fellow at the School of Engineering, RMIT University. Currently, she is a Data Scientist with the Defence Science and Technology Group, Melbourne. Her research interests include audio-visual signal processing, neural networks, and machine learning.

SIMON G. HOSKING received the B.App.Sci. (Hons.) and Ph.D. degrees in psychology from Deakin University, in 1999 and 2006, respectively. He is currently the Discipline Leader of the Training Effectiveness Research in Human Factors, Aerospace Division, Defence Science and Technology Group. His research interests include physiological and behavioral indices of team performance, team-based spatio-temporal metrics, and human-autonomy teaming.

DAWEI JIA received the B.S. degree (Hons.) in computing, in 2006, and the Ph.D. degree in human-computer interaction from Deakin University, Melbourne, Australia, in 2011. She is currently a Human Factors Research Scientist at the Defence Science and Technology Group. Her research interests include human cognition, communication, eye movements, human performance measurement, human-machine interface design, and evaluation.

MARGARET LECH (Member, IEEE) received the M.S. degree in physics from Maria Curie-Skłodowska University, Poland, the M.S. degree in biomedical engineering from the Warsaw University of Technology, Poland, and the Ph.D. degree in electrical engineering from the University of Melbourne, Australia. She is currently a Professor in signal processing and artificial intelligence with the School of Engineering, RMIT University, Australia. Her research interests include machine learning applications in speech and image processing, system modeling, and optimization.