Removal of Confounding Factors using GA-SVM Feature Adaptation: Application on Detection of Vocal Fatigue thru sEMG Classification

Yixiang Gao[1] *IEEE Student Member*
G. N. DeSouza[2] *IEEE Senior Member*
Mark Berardi[3]
Maria Dietrich[4]

***Abstract*—As machine learning solutions become increasingly more ubiquitous in medical diagnosis, researchers are becoming equally more aware of the possibility of confounded predictions being produced by these same models. This realization derives, for example, from the observation that sample-wise cross-validation leads to highly underestimated error predictors when compared to subject-wise cross validation. However, without a reliable approach to remove spurious, confounding factors such as age, gender, or even the type/brand of equipment used, these same machine learning solutions will be fated to produce poor results despite the metrics for error estimation employed. In this research, we propose an optimization approach, using genetic algorithms, to adapt the feature vectors in order to maximize the prediction accuracy of a given classifiers, while minimizing the correlation between the features and the potential confounding factors. Our results, when applied to the diagnostic of vocal fatigue, have shown great improvement in terms of the generalization capability of the chosen SVM classifier. The system was evaluated using subject-wise (i.e. leave-one-subject-out) cross-validation, which demonstrated the effectiveness of this new confounding removal approach.**

***Index Terms*—surface electromyography, Genetic Algorithm, Support Vector Machine, Confounding Factor**

I. Introduction

As machine learning techniques have rapidly advanced and higher-quality sensors have become accessible, medical applications have begun to leverage these methods to solve diagnostic problems. For instance, voice signal classification can be employed to identify Parkinson's disease [1], MRIs can be used for HIV diagnosis [2], and sEMG signals can be classified to detect vocal fatigue [3], [4]. Although most of these machine learning diagnostic approaches have been demonstrated to be very effective and accurate using traditional sample-wise cross-validations, the performance of these systems drop significantly when evaluated subject-wise – i.e. when the model is tested on hitherto unseen subjects ([5]). Additionally, as the authors of [6] also mentioned in their study, such diagnostics algorithms, which are trained with repeated measurements, would be likely to cause the issue of "identity confounding", where the models learn to identify subjects instead of the diagnostic or disease pattern itself.

The presence of confounding factors has been a challenging topic for machine learning applications, especially in medical use cases. Previous works have demonstrated that confounding factors such as age and gender can influence the decision made by the classifier, eventually leading to bias in the learning of those diagnostics. However, a few solutions have been proposed to solve this problem of biases from confounding factors. For instance, [7] proposed a causal hidden Markov model that separates the confounding biases through salient (hidden) variables, which were also learned by the model to prevent their influence in the model's diagnosis decision. In [2], the authors applied an adversarial training approach using a deep learning model which extracted features that both maximized the model's classification accuracy and minimized the correlation between the extracted features and confounding variables (in that case, the subject's age or gender). Although both methods have already demonstrated success in removing confounding factors given their particular application, they required the construction of specific model architectures, which are not necessarily easy to adapt to other potential factors.

In this paper, we propose an optimization approach using genetic algorithm (GA) that is capable of removing the confounding factors. In essence, given the confounding variables, the approach seeks to find a linear transformation on the feature vectors thru the use of GA in order to minimize both the classification loss and the correlation between the reweighted feature vectors and the confounding factors. We used two datasets to demonstrate the effectiveness of the proposed approach for confounding-factor removal. The first dataset was the voice sEMG data collected from our previous studies [4]. Our results showed an improvement in terms of the capacity for generalization of a SVM classifier using a subject-wise, or leave-one-subject-out (LOO), validation. The testing accuracy was improved from 65% to 68%. However, in order to show that the results were not dependent on the dataset and/or the specific confounding factor, we tested our confounding removal approach on a second dataset which was proposed to identify depression using vocal acoustic

[1]*yg5d6@umsystems.edu,* [2]*DeSouzaG@missouri.edu,* [3]*mark.berardi@ukbonn.de,* [4]*maria.dietrich@ukbonn.de*
[1,2]*ViGIR Lab, Department of Electrical Engineering and Computer Science University of Missouri, Columbia/MO, USA*
[3,4]*Department of Psychiatry and Psychotherapy, University Hospital Bonn, Bonn, Germany*
[4]*Department of Speech, Language and Hearing Sciences University of Missouri, Columbia/MO, USA*

979-8-3503-1458-8/23/$31.00 ©2023 IEEE

features. After applying the proposed approach, we were able to improve the generalization of the SVM from 64% to 67%.

In the following sections, we provide a brief summary of both datasets and their related studies. We also give some technical background for our proposed GA-SVM pipeline.

II. Background

A. Vocal Fatigue Detection Using sEMG Pattern Recognition

In our previous study [3], the goal was to detect vocal fatigue using sEMG signals collected from subjects performing voice gestures - vowels, syllables, sentences, etc. So, we recruited a total of 92 subjects, which were divided into two groups labeled vocally healthy and vocally fatigued according to self-reported VFI-1 scores — VFI-1 is the first factor of the Vocal Fatigue Index: tiredness of voice and avoidance of voice use [8]. Next, we selected 40 matched subjects (20 vocally healthy and 20 vocally fatigued) and conducted extensive classification experiments using seven different sEMG features as well as the GUSSS ratio which was previously developed in [9]–[11]. The proposed SVM model was cross-validated using a subject-wise, i.e. leave-one-out (LOO), approach. The result revealed that, although the classifier was able to achieve almost perfect performance — 99% accuracy using using sample-wise cross-validation – the testing accuracy based on LOO experiments was much lower. That is because in a sample-wise cross-validation, samples from all subjects are shuffled together before they are divided into training and testing sets, and hence, the training is performed with samples from all subjects. On the other hand, on subject-wise cross validation (LOO), the classifier is not allowed to learn from data from *'future'* subjects. Moreover, our study also indicated that by retaining only 10% of the subjects' testing data in the training (which accounted for less than 1% of our entire training data), the testing accuracy of the proposed SVM classifier could improve significantly. These findings led us to assume that the SVM classifier was potentially learn from individual subject characteristics, rather than the actual vocal fatigue symptoms. This phenomenon was also reported in [6] on using a random forest classifier to diagnosis Parkinson's disease using voice acoustic signals.

B. The Influence of Skinfold Thickness on sEMG Signal

It is assumed that one confounding factor for sEMG signals is the skinfold thickness around the muscle groups where the electrodes were placed (shown in Figure 1). This assumption derives from several studies reporting the influence of skinfold thickness at different muscle groups. For instance, in [12], it was observed that the selectivity of the surface electrodes increased as the skinfold thickness decreased. In [13], the authors have also reported that sEMG cross-talk above the region of inactive muscle increased as the skinfold thickness increased. Moreover, in [14], it was shown that the muscle-electrode distance, measured by ultrasound, explained 33% of the variance of sEMG activity at standardized sub-maximal contraction and 21% for maximal contractions. They also demonstrated that the skinfold measurements from four other sites explained as much as 68% of the inter-subject variance in sEMG amplitude.

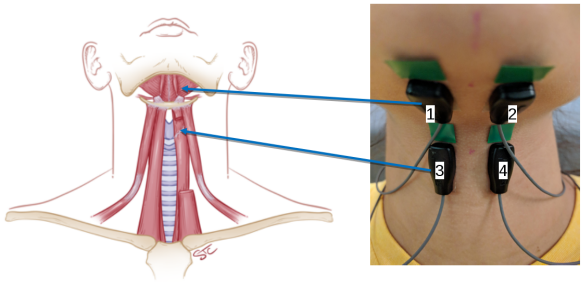


Fig. 1: Electrode placement used in our study for detecting vocal fatigue via sEMG classification. Top pair of electrodes targeted the suprahyoid muscle group, and the bottom pair, the infrahyoid muscle group.

One way to reduce the effects of skinfold thickness is by means of maximal voluntary contraction (MVC) normalization. However, as mentioned in [15], finding appropriate tasks to induce MVC for speech musculature is non-trivial. Similarly, our previous study [4] also showed that applying MVC normalization to the sEMG signals did not improve the accuracy of vocal fatigue detection due to the difficulty of having consistent MVC tasks performed across multiple subjects.

III. Feature Adaptation using GA-SVM

In this section, we introduce our feature adaptation approach based on GA-SVM. The main idea of our method is to remove the influence of a given confounding factors from the extracted feature vectors by reweighting those same features. Traditionally, GA-SVM has been applied to optimize SVM model parameters, such as its penalty factor and kernel function parameters. In [16], the authors included a binary vector inside their chromosome design, and GA was then used to simultaneously optimize hyper-parameters of the SVM as well as the best set of features for its classification task. They have demonstrated that such approach could quickly acquire effective feature sets and SVM model parameters that led to better classification results.

Inspired by their method, we modified the chromosome construction from feature and model selection to feature reweighting, while keeping our SVM model parameters and hyper-parameters fixed. We established the chromosome as a vector, where the elements inside are essentially weights (scalars) associated with each feature dimension for both training and testing samples. Furthermore, we set up our GA algorithm as a multi-objective optimization problem, where the aim was to find a set of chromosomes (weights), which, after being applied to the feature vector, would minimize both the SVM training loss as well as the correlation between the SVM model predictions and the confounding factor. Figure 2 shows the flowchart of our purposed GA-SVM pipeline.

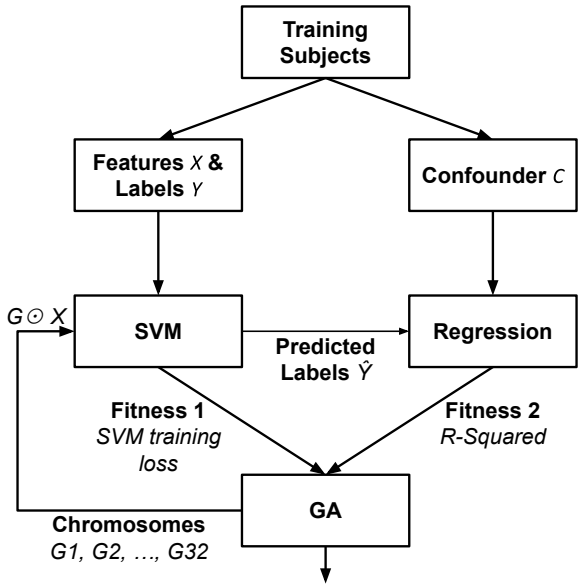


Fig. 2: The proposed pipeline for feature adaptation using GA-SVM.

The following notations are used throughout the rest of this paper: X is sample matrix containing all feature vectors, each with d dimensions; Y is the ground true labels, while $\hat{Y}$ is the predicted labels from SVM; and C represents the confounder.

A. Chromosome Design

Assuming the sample matrix X is an $n \times d$ matrix, where n is the number of samples and d is the dimension of the feature vectors, the chromosome can be then expressed as

$$G = [g^1, g^2, \ldots, g^d] \tag{1}$$

where g is a set of weights applied to each individual feature dimension, and with values between $[-1, 1]$. Once applied, the weighted features in the sample matrix X' becomes

$$X' = G \odot X = \begin{bmatrix} g^1 X^{11} & g^2 X^{12} & \ldots & g^d X^{1d} \\ g^1 X^{21} & g^2 X^{22} & \ldots & g^d X^{2d} \\ \vdots & \vdots & \ddots & \vdots \\ g^1 X^{n1} & g^2 X^{n2} & \ldots & g^d X^{nd} \end{bmatrix} \tag{2}$$

where $\odot$ is used to denote the dimension-wise multiplication between the chromosomes and the feature vectors. The transformed feature vector X' is then fed into the two fitness functions for the GA to optimize.

B. Fitness Function

We first obtain our SVM classifier, denote by $f_{SVM}()$, using the original feature vector X and its ground truth Y. Then we fix the model parameters and hyper-parameters during the GA optimization process to guarantee the consistency of the search over the same space imputed by $f_{SVM}()$. Next, the GA minimizes the loss using the predicted labels from the weighted feature vectors, $\hat{Y} = f_{SVM}(X')$, as well as the correlation between those same predictions and the known confounding factor C. These two proposed fitness functions can be expressed by the following terms:

$$f_1 \sim f_1(X') = loss(\hat{Y}, Y) = loss(f_{SVM}(X'), Y) \tag{3}$$

$$f_2 \sim f_2(X') = corr(\hat{Y}, C) = corr(f_{SVM}(X'), C) \tag{4}$$

where the $loss$ is simply the error rate between predicted labels and the ground truth. However, since $corr$ can be regarded as one of the many correlation measurements, for the scope of this paper we used r-squared obtained from fitting a linear regression model between $\hat{Y}$ and C, which is possible in our case due to the confounding variables being continuous. For discrete confounders, a de-categorization mapping of the same confounders would have to be employed. More details will be explained in Section V.

The rest of GA optimization pipeline follows the standard multi-objective evolutionary optimization framework [17] where the algorithm finds chromosomes on the pareto front that minimizes the two objective functions f_1 and f_2 defined in equation (3) and (4).

Fig. 3: System setup for data collection showing the flow of signals through the equipment.

IV. Experiment Setup

A. Voice sEMG dataset Confounded by Skinfold Thickness

This section briefly describes the subject characteristics, data collection protocol, and system setup used in this study illustrated in Figure 3. The equipment was installed in a soundproof booth (IAC Acoustics, North Aurora, IL) and included: (1) a base station and four wireless Trigno™ mini sEMG sensors with a bandwidth of 20 Hz to 450 Hz (Delsys, Natick, MA); (2) a head-worn microphone (AKG, Model C520, Vienna, Austria); (3) an audio interface (Scarlett 2i2, Focusrite, High Wycombe, UK); and (4) a data acquisition device (PowerLab 16/35, ADInstruments, Dunedin, New Zealand), which performed the synchronized sampling of audio and sEMG signals using LabChart v. 8.1.10 for MS Windows. The sampling rate was set to 4 kHz for the sEMG signals and 20 kHz for audio, both with 16-bit quantization. In addition, audio signals were collected using the software Audacity(R)© v. 2.1.1 for recording and editing at a sampling rate of 44.1 kHz and 16-bit quantization. For the purpose of this study, we focused on classification using only the sEMG data for vowel productions.

1) Subject Characteristics: We selected 40 subjects out of the 92 in the dataset with matching ages and skinfold thicknesses from our original study [4]. The control group had to score ≤ 10 (vocally healthy) on the VFI-1 [8] while early career teachers (within their first 10 years of teaching experience) had to score > 10 (vocally fatigued) on the VFI-1 during the pre-screening. They were all in good general health with no acute or chronic upper respiratory infection or pulmonary disease (including allergies, laryngopharyngeal reflux disease, or asthma that affected their voice at the time of participation). Subjects were asked to perform a series of repetitive voice productions including vowels, syllables, and sentences, etc. For the scope of this study, only the vowel productions (/a/, /u/, /i/) were used in our experiments. The total resulting number of vocally healthy and vocally fatigued samples were 3202 and 3270 respectively.

2) Features and Confounding Variables: Eight features were extracted from all the collected sEMG signal samples. They were: (1) *mean absolute value* (one dimension); (2) *zero crossings* (one dimension); (3) *slope sign changes* (one dimension); (4) *waveform length* (one dimension); (5) *Willison amplitude* (one dimension); (6)*root mean square* (one dimension); (7) the coefficients of a fourth-order *auto regressive* model (four dimensions) and finally (8) *GUSSS ratio* (two dimensions) [9]–[11]. These features were extracted from all four sEMG channels (Figure 1) for a total of 48 dimensions in the feature vector. They were explained in greater detail in our previous work [3], but also in other papers in the literature [18]–[22].

As mentioned previously, in Section II, we clinically identified skinfold thickness as the confounding variable due to its strong influence on inter-individual variance in sEMG amplitude, as demonstrated in [14]. A caliper (Lange Skinfold Caliper, Beta Technology, Cambridge, MD, USA) was used to measure subjects' skin-fold thickness overlying the submental and infrahyoid muscle groups where electrodes were placed: three recordings per site were averaged. Figure 1 shows the electrode placement and the corresponding muscle groups. After averaging measurements from both submental and infrahyoid muscle groups, we obtained an one-dimensional vector which we treated as a continuous confounding variable for computing the correlation term of the fitness function.

B. Vocal Acoustic Features dataset Confounded by Subject IDs

Additionally, we also tested our GA-SVM on a separate dataset in the speech domain released in [23] to diagnose depression using voice features. It included 73 participants (34 healthy individuals and 39 depressed patients) who were between 30 and 60 years old. A totally 714 and 819 samples were collected between healthy and depression patients, respectively. After feature extraction and selection, 36 voice features were kept. Propensity score matching was used to estimate the affect of confounding variables which were the demographic information - age, occupation, and education, all of which attributed to subjects' identity. Therefore, subject IDs were used as the confounder C to compute the $correlation$ in the fitness function.

C. Partial Confounder Test

A statistical hypothesis testing scheme known as Conditional Permutation Test (CPT) was proposed to quantify known confounding bias [24] by measuring the conditional independence $X \perp\!\!\!\perp Y|C$. The test performs random permutation of X based on the probability distribution function $Q(X|C)$, which can be estimated from the given dataset using statistical inference methods. The test consists of obtaining the p-value using all permutated copies $X^{(1)}, \cdots X^{(M)}$ for an arbitrary statistical metric (e.g. correlation, entropy, etc.) $T = T(X, Y, C)$, given by

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(X^{(m)}, Y, Z) \geq T(X, Y, Z)\}}{1 + M} \quad (5)$$

where the higher p-value indicates a *stronger* conditional independence between X and Y. In practice, the threshold for p-value is set at 0.05 for X and Y to be considered conditionally dependent.

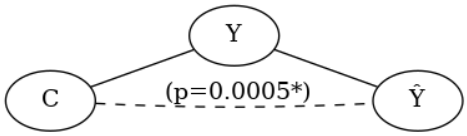
More recently an alternative of CPT was proposed by [25] where the hypothesis $\hat{Y} \perp\!\!\!\perp C|Y$ was used instead for quantifying confounding bias between given model predictions and the known confounder using the same p-value as the metric. We adopted this quantification to demonstrate the effectiveness of our proposed GA-SVM in removing confounding bias along with the classification accuracy using LOO cross-validation.

V. Results

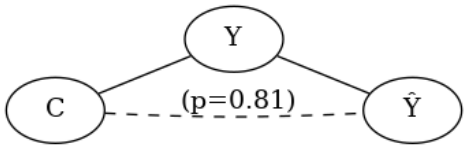
In this section, we present our classification results using GA-SVM for both the voice fatigue, sEMG dataset and the depression, voice acoustic dataset. We compare the performance between the vanilla SVM and our proposed GA-SVM. In order to avoid the underestimation of prediction error raised by [5], [6], we used the subject-wise/LOO approach, where if N is the number of positive subjects and M the number of negative subjects – and under the subject-wise/LOO condition, a total of $N + M - 1$ subjects were selected for training, while the subject *left-out* was used for testing. Finally, we include the p-value metric from CPT to demonstrate that our proposed GA-SVM has indeed removed confounding bias from our model predictions.

A. Voice sEMG dataset

First, we demonstrate our experiment results on confounding bias removal using our voice fatigue, sEMG dataset, where the subjects' skinfold thickness is regarded as a confounding variable. The r-squared value used for the $corr$ term of the fitness function, in Equation 4, was estimated based on fitting a linear regression model of skinfold thickness using binary predictions from SVM and GA-SVM. Figure 4 shows an example using the partial confounder test on our trained classifiers outputs. The goal of using partial confounder test is to quantify the conditional independence between model predictions $\hat{Y}$ and confounder C conditioned on ground truth

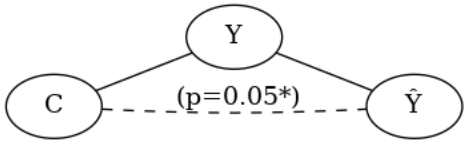


(a) SVM

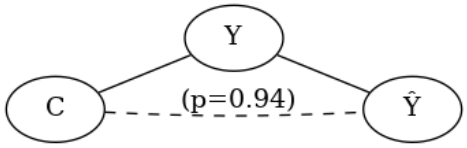


(b) GA-SVM

Fig. 4: Partial confounder test of the hypothesis $\hat{Y} \perp\!\!\!\perp C|Y$ for SVM and GA-SVM trained on voice sEMG data. Y is the ground truth, $\hat{Y}$ is the model prediction, and C is the confounder which in this case is the skinfold thickness of subjects. A p-value (shown in parentheses) less than 0.05 indicates a strong dependence between $\hat{Y}$ and C.



(a) SVM



(b) GA-SVM

Fig. 5: Partial confounder test of the hypothesis $\hat{Y} \perp\!\!\!\perp C|Y$ for SVM and GA-SVM trained on voice acoustic features. Y is the ground truth, $\hat{Y}$ is the model prediction, and C is the confounder which in this case is the subjects IDs. A p-value (shown in parentheses) equals to 0.05 still indicates a strong dependence between $\hat{Y}$ and C.

Y — $\hat{Y} \perp\!\!\!\perp C|Y$. As shown in Figure 4, the p-value obtained from SVM is extremely low (less than .0005) which indicated a strong dependence between its predicted labels and the confounder. On the other hand, the proposed GA-SVM was able to remove that confounding bias, indicated by a much higher p-value (0.81) from the same CPT test, indicating much more independent model predictions.

We then performed partial confounder test on the voice acoustic dataset for depression diagnostic and observed the same trend as shown in Figure 5. There was a strong dependence (p-value=0.05) between the model predictions from SVM and the confounder - subject IDs. After we applied our proposed GA-SVM, the dependence was successfully removed indicated by p-value=0.94.

Next, we evaluated the classification performance of both classifiers (vanilla SVM and GA-SVA) using LOO cross-validation. The reported p-value, training accuracy and testing accuracy in Table I were averaged over 10 experiments with different random initialization. While the vanilla SVM could achieve an almost perfect training accuracy 99%, its generalization capability was extremely poor and its model predictions were completely dependent on the confounder based on the extremely low score of the averaged p-value (.0005). In comparison, the proposed GA-SVM not only was successful at removing the confounding bias indicated by the much higher p-value (.8109), but it also was able to improve the testing accuracy from 65% to 68%. Moreover we observed the reduction in training accuracy as well which confirmed that the SVM classifier was heavily depending on the confounding variable (skinfold thickness) to give its predictions.

TABLE I: Classification performance comparison and corresponding p-values for predicting vocal fatigue using voice sEMG dataset confounded by skinfold thickness.

	p-value	Training	Testing
SVM	$.0005 \pm .0001$	$99.10\% \pm .01\%$	$65.04\% \pm 0.59\%$
GA-SVM	$.8109 \pm .0295$	$86.35\% \pm .83\%$	$68.33\% \pm 3.15\%$

TABLE II: Classification performance comparison and corresponding p-values for predicting depression using voice acoustic features confounded by subject IDs.

	p-value	Training	Testing
SVM	$.0487 \pm .0006$	$75.50\% \pm .00\%$	$63.67\% \pm 0.00\%$
GA-SVM	$.9373 \pm .0252$	$72.33\% \pm .30\%$	$66.94\% \pm 1.54\%$

B. Vocal acoustic features dataset

Next, we evaluated our proposed GA-SVM using the depression, voice acoustic dataset, where the subject IDs were regarded as a confounding variable. The results are summarized in Table II. Once again, we observe a consistent outcome with respect to the previous experiment in terms of producing much more independent model predictions (p-value increased from .0487 to .9373) and better model generalization performance (testing accuracy increased from 63% to 67%). It is important to mention that our experiments using SVM and GA-SVM are consistent, while slightly better than the classification results reported in [23]. Their classification model achieved 63% accuracy when only the voice acoustic features were used, and when both voice acoustic features and demographic data were combined, the testing accuracy of their model also increased to 66%.

VI. Conclusion and Future Works

Our proposed GA-SVM has demonstrated its effectiveness in removing confounding bias via applying re-weighting onto the features that both minimize its correlation to the confounder and training error of SVM. We demonstrated the performance of our proposed approach through two different dataset across different domains: sEMG signals from voice productions w/ and w/o vocal fatigue; and acoustic voice features from subjects w/ and w/o depression. Although the

objectives of our GA-SVM were not conclusively much better with respect to model generalization, it demonstrated its ability to de-correlate confounder and predictions, while still improving LOO testing accuracy in both cases. Furthermore, the proposed GA optimization scheme (here demonstrated only through SVM) can, in theory, be extended to any arbitrary classification or regression models.

Statistically, the p-value needs to be less than 0.05 to reject the conditional independence hypothesis $\hat{Y} \perp\!\!\!\perp C|Y$, implying that the model predictions $\hat{Y}$ are indeed confounded by C as demonstrated in Table I and II for the vanilla SVM. On the other hand, to accept the same hypothesis of conditional independence, the p-value largely exceeded 0.05 in both of our GA-SVM experiments, returning almost *perfect* p-values of $.8109$ and $.9373$. More investigations are needed to better understand the trade-off between p-value and testing accuracy in order to further shorten the gap between training and testing in both cases.

The confounding effects on classifiers, whether from subject identities or from subject attributes (e.g. age, gender, subject identity etc) remain a challenging task for medical diagnostic applications, especially when the confounding variables our unknown. Therefore, our current work focuses on not only mitigating or removing the effects of known confounding sources, but identifying unknown factos as well.

ACKNOWLEDGMENT

Research reported in this publication was supported by the National Institute On Deafness and Other Communication Disorders of the National Institutes of Health under award number R15DC015335 and R01DC018026. Thank you to Ashton Bernskoetter, Taylor Hall, Katherine Johnson, Haley McCabe, Melinda Pfeiffer, and Allison Walker for assistance with data collection. We thank Matthew Page, MD, for reviewing laryngeal videostroboscopies.

REFERENCES

[1] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, Jun 2007.

[2] Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications," *Nature Communications*, vol. 11, no. 1, p. 6010, Nov 2020.

[3] Y. Gao, M. Dietrich, M. Pfeiffer, and G. N. DeSouza, "Classification of sEMG Signals for the Detection of Vocal Fatigue based on VFI Scores," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 5014–5017.

[4] Y. Gao, M. Dietrich, and G. N. DeSouza, "Classification of vocal fatigue using semg: Data imbalance, normalization, and the role of vocal fatigue index scores," *Applied Sciences*, vol. 11, no. 10, 2021.

[5] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *Gigascience*, vol. 6, no. 5, pp. 1–9, May 2017.

[6] E. Chaibub Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, P. Snyder, B. M. Bot, A. D. Trister, S. H. Friend, L. Mangravite, and L. Omberg, "Detecting the impact of subject characteristics on machine learning-based diagnostic applications," *npj Digital Medicine*, vol. 2, no. 1, p. 99, Oct 2019.

[7] J. Li, B. Wu, X. Sun, and Y. Wang, "Causal hidden markov model for time series disease forecasting," *CoRR*, vol. abs/2103.16391, 2021. [Online]. Available: https://arxiv.org/abs/2103.16391

[8] C. Nanjundeswaran, B. H. Jacobson, J. Gartner-Schmidt, and K. Verdolini Abbott, "Vocal Fatigue Index (VFI): Development and Validation," *Journal of Voice*, vol. 29, pp. 433–440, 2015.

[9] L. Rivera and G. N. DeSouza, "Recognizing Hand Movements from a Single sEMG Sensor using Guided Under-determined Source Signal Separation," in *12th IEEE ICORR*, Jun 2011, ETH Zurich, Switzerland.

[10] N. R. Smith, T. Klongtruagrok, G. N. DeSouza, C. R. Shyu, M. Dietrich, and M. P. Page, "Non-invasive ambulatory monitoring of complex sEMG patterns and its potential application in the detection of vocal dysfunctions," in *2014 IEEE 16th Healthcom*, 2014, pp. 447–452.

[11] N. R. Smith, L. A. Rivera, M. Dietrich, C. R. Shyu, M. P. Page, and G. N. DeSouza, "Detection of Simulated Vocal Dysfunctions Using Complex sEMG Patterns," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, pp. 787–801, 2016.

[12] E. J. De la Barrera and T. E. Milner, "The effects of skinfold thickness on the selectivity of surface EMG," *Electroencephalogr Clin Neurophysiol*, vol. 93, no. 2, pp. 91–99, Apr. 1994.

[13] T. A. Kuiken, M. M. Lowery, and N. S. Stoykov, "The effect of subcutaneous fat on myoelectric signal amplitude and cross-talk," *Prosthet Orthot Int*, vol. 27, no. 1, pp. 48–54, Apr. 2003.

[14] C. Nordander, J. Willner, G.-A. Hansson, B. Larsson, J. Unge, L. Granquist, and S. Skerfving, "Influence of the subcutaneous fat layer, as measured by ultrasound, skinfold calipers and BMI, on the EMG amplitude," *Eur J Appl Physiol*, vol. 89, no. 6, pp. 514–519, Apr. 2003.

[15] C. E. Stepp, "Surface electromyography for speech and swallowing systems: measurement, analysis, and interpretation," *J Speech Lang Hear Res*, vol. 55, no. 4, pp. 1232–1246, Jan. 2012.

[16] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "Ga-svm based feature selection and parameter optimization in hospitalization expense modeling," *Applied Soft Computing*, vol. 75, pp. 323–332, 2019.

[17] K. Deb, *Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction*. London: Springer London, 2011, pp. 3–34.

[18] A. J. Young, L. H. Smith, E. J. Rouse, and L. J. Hargrove, "Classification of Simultaneous Movements Using Surface EMG Pattern Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 1250–1258, 2013.

[19] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 7, pp. 848–854, Jul 2003.

[20] M. Zardoshti-Kermani, B. C. Wheeler, K. Badie, and R. M. Hashemi, "EMG feature evaluation for movement control of upper extremity prostheses," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, pp. 324–333, 1995.

[21] T. Fukuda, J. Echeimberg, J. Pompeu, P. Lucareli, S. Garbelotti Junior, R. Gimenes, and A. Apolinário, "Root Mean Square Value of the Electromyographic Signal in the Isometric Torque of the Quadriceps, Hamstrings and Brachial Biceps Muscles in Female," *Journal of Applied Research*, vol. 10, pp. 32–39, 2010.

[22] L. Hargrove, K. Englehart, and B. Hudgins, "A training strategy to reduce classification degradation due to electrode displacements in pattern recognition based myoelectric control," *Biomedical Signal Processing and Control*, vol. 3, pp. 175–180, 2008.

[23] W. Pan, J. Flint, L. Shenhav, T. Liu, M. Liu, B. Hu, and T. Zhu, "Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders," *PLOS ONE*, vol. 14, no. 6, pp. 1–14, 06 2019.

[24] T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth, "The Conditional Permutation Test for Independence While Controlling for Confounders," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 1, pp. 175–197, 10 2019.

[25] T. Spisak, "Statistical quantification of confounding bias in machine learning models," *GigaScience*, vol. 11, 08 2022.