Detection of Universal Cross-Cultural Depression Indicators from the Physiological Signals of Observers

J. F. Plested
T. D. Gedeon
and X. Y. Zhu
Research School of Computer Science
Australian National University
Email: jo.plested@anu.edu.au

A. Dhall
Department of Computer Science and Engineering
Indian Institute of Technology Ropar
R. Geocke
Faculty of Education, Science, Technology and Mathematics
University of Canberra

***Abstract*—We conducted a pilot study experimenting with neural network techniques to use the physiological signals of untrained observers to classify the depression levels of variously depressed people in videos speaking a language the observers did not understand. As the dataset was highly imbalanced, noisy and thus extremely sensitive to relative class sizes, we developed a technique for dynamically oversampling the smaller classes both prior to and during training to approximately align training prediction rates for each class with knowledge of the prevalence of different levels of depression. In predicting the depression levels to a final accuracy of 57.9% over four classes and 78.9% over three classes we demonstrate the likelihood that universal cross-cultural indicators of depression exist. In addition, that some people's automatic physiological responses to these indicators are strong enough that they can be used to predict depression categories of people to a significant degree of accuracy even when the observer does not understand the language the person is speaking. The final accuracy rate is significantly better than the diagnosis rates of doctors speaking to patients in their own language. The results show the potential these techniques have to improve diagnosis of depression, especially in areas with limited access to mental health professionals. This innovative approach demonstrates the importance of further experimentation in this area and research into universal cross-cultural depression indicators.**

1. Introduction

Major depressive disorder, also known as clinical depression, is a mood disorder that is characterized by persistent feelings of sadness, low self-esteem, and loss of interest [1]. According to the World Health Organization (WHO) an estimated 350 million people worldwide are affected by depression, it is the fourth leading cause of disability worldwide and predicted to move to first place by 2030. Although there are known, effective treatments for depression available, fewer than half of those affected in the world and in many countries fewer than 10%, receive such treatments.

Barriers to effective care identified by the WHO include a lack of resources, lack of trained health care providers, and social stigma associated with mental disorders [2]. Aside from the personal toll of depression to the individual, a number of studies have estimated the high financial burdens to economies. The total economic cost of depression in Europe in 2010 was estimated at 92 billion Euros [3]. The cost in lost productivity of workers with depression in the United States in 2002 was estimated at 44 billion dollars [4].

Depression is generally diagnosed via self or clinician assisted questionnaires [5] resulting in diagnoses that are not always consistent and reliable. A meta-analysis of 118 studies [6] showed that general practitioners correctly diagnosed depression in only 47.3% of cases where depression was known to be present and recorded depression in their notes in 33.6%.

Depression has specific effects on certain areas of the brain that can be observed through problems with speech production, how things are said rather than what is said [1] and a range of visual indicators such as differences in facial expression, eye gaze, demeanor and gestures [7], [8]. Given these are the result of problems in the brain, as opposed to learned behaviors, it is likely that many of them could be shown to be universal indicators of depression rather than culturally specific.

It is clear that the identification of universal cross-cultural indicators of depression would assist with both early diagnosis and overcoming the stigma associated with depression, especially in countries with extremely low rates of diagnosis. More accurate and earlier diagnosis together with the use of known effective treatments would provide the opportunity to decrease the burden of disease both individually and economically [9]. The term cross-cultural includes language(s) spoken, and is the aspect of cross-cultural detection of depression we consider here, in line with previous work in this area [10].

Many studies have shown that changes in emotional states are reflected in changes in the physiological signals of the autonomic nervous system (ANS) [11], [12]. Music judged to be calming, neutral or exciting has been shown to have significantly different effects on galvanic skin response (GSR) but not heart rate [13]. Scales of arousal, valence and

978-1-5386-0680-3/17/$31.00

liking in participant's responses to music videos have been predicted at significantly better than chance levels using an electroencephalogram (EEG) and other peripheral physiological signals [14]. A high accuracy has been achieved with a neural network trained to predict emotions using physiological signals recorded while participants watched videos and rated their emotional response into one of four categories [15]. Our previous work has demonstrated that physiological signals can be used to reliably predict observer stress when exposed to videos of known stressfulness as validated by user surveys.

The aim of our current research is to examine physiological signals using neural networks (NNs) to determine whether observers are responding to universal cross-cultural indicators of depression when watching videos of people with varying levels of depression and non-depressed controls speaking in a language the observers do not understand.

NNs are powerful models that have achieved excellent performance on a range of learning tasks, but with limited, skewed and noisy training data they can be difficult to train and suffer from significant overfitting [16], [17]. We developed a technique to approximate optimal oversampling class ratios to help overcome these problems.

2. Experiment Design

2.1. Subjects

Ethics approval to perform the experiment was received from our University's Human Research Ethics Committee.

Ten Masters students were recruited for the experiment, five males and five females. The total being a normal size for publication as a preliminary trial in medicine [18], [19]. The participant cohort was from a range of disciplines across the University, with none from Psychology, Behavioral Science, Medicine or any disciplines likely to be conducive to the recognition of depression status. Five subjects (three male, two female) indicated some level of prior experience coping with depressed people. Their age was 27.5 years on average, with standard deviation 8.3.

2.2. Dataset

We used the AVEC 2014 dataset [20], it consists of videos in Development, Training, and Test partitions. The labels for depression status are from the Beck Depression Inventory-II [21]. We used only videos from the Freeform category, in the Test set.

We chose 29 of the shortest videos, of lengths 0.5 minutes or longer. This selection was done to maximize the number of videos shown while keeping the total experiment length to under one hour, as we have difficulty in recruiting subjects for longer experiments. Ten subjects responding to 29 videos equates to 290 total responses.

There are standard 'broad ranges' in the Beck Depression Index-II [21] which map the depression scores into categories:

- 0-13: indicates no or minimal depression.
- 14-19: indicates mild depression.
- 20-28: indicates moderate depression.
- 29-63: indicates severe depression.

Our videos were selected as described above, from the perspective of human experiment practicality. The 29 videos can be categorized into the four broad ranges with 14, 6, 5, and 4 videos respectively in the ranges from no or minimal to severe.

2.3. Measures and Sensors

2.3.1. Galvanic Skin Response (GSR). Skin conductance, also known as electro-dermal response or psycho-galvanic reflex, measures the electrical conductance of an individual's skin, which varies due to the amount of sweat on the skin. When the individual is under stress, skin conductance will increase; conversely, the skin conductance will reduce when the individual encounters less stress [22]. We used the Neulog GSR Logger Sensor [23].

2.3.2. Heart Rate Variability. Heart rate variability (HRV) is the variation in the interval between heartbeats. Measures used to determine HRV include Electrocardiogram (ECG) and blood pressure, but ECG is considered superior as it excludes unnecessary heartbeats and displays a clear waveform. We used the Neulog ECG Logger Sensor [24]. The sampling rate of the ECG was 10Hz which is the maximum able to be reliably stored by the device.

2.4. Conduct of the Experiment

Subjects filled in a pre-experiment questionnaire to collect their demographic characteristics, ability to speak German (none), and experience with coping with depressed people.

Subjects then watched the 29 short videos and were prompted to select a depression classification at the end of each video, during a five second gap before the next video. The videos were presented in an order balanced way to avoid effects of presentation ordering. Thus, the positioning of each video, near the beginning / middle / end of the experiment, was different for each subject.

3. Neural Network Classification Techniques

The physiological signals recorded while subjects watched the videos were split into segments of length five seconds with an overlap of half a segment. This resulted in 550 segments in total per person across the 29 videos.

The videos were split into test and training sets using 10-fold cross validation. This resulted in nine folds with three videos in the test set and one fold with two longer videos. The test sets were further split into one video for a validation set and two or one (for the final fold) for the test set. The validation and final test sets were rebalanced slightly so they had similar proportions of video segments

from each class. The dataset was split by videos for cross validation purposes rather than by subjects because it is a more useful diagnostic tool to have a system that can use the physiological signals of one or more observers to diagnose the depression level of an individual in a previously unseen video than to predict the depression levels of a set of videos of people with known pre-labelled depression levels using a new observer.

GSR and HRV are well known to be noisy measures and sensitive to influences external to experiments such as the subject's current emotional state and stress levels [11], [15]. Since it was not possible to test the subjects on multiple occasions to account for fluctuations in mood and stress level the resulting data set was small and noisy. As described in section 2.2, the dataset was also highly imbalanced with many more videos in the no or minimal depression category. All these properties of the dataset make it similar to real world situations e.g. a doctor attempting to diagnose a patient in one session, but also presented considerable challenges for using neural networks for classification.

Previous studies have shown that in small unbalanced datasets where the minority class is much smaller than the majority class the naturally occurring class distribution is often not the optimum one for producing the best training results and tends to result in an unacceptably high error rate on the minority class [25], [26]. Accuracy was improved by up to 40% when the optimal class sizes were used over the naturally occurring ones for training with smaller, unbalanced datasets. Having a high error rate on the minority class is a particularly undesirable effect in diagnosis of depression and other illnesses as the goal is for accuracy in diagnosing the illness. The problems created by small unbalanced datasets are significantly compounded in the domain of detection of depression and other illnesses via videos of patients. This is because datasets are generally at least an order of magnitude smaller than the smallest dataset used in [26] due to privacy constraints. The problems are further compounded because of the noise present in the data. It has also been shown that the optimum class sizes to train the best classifier vary significantly with the dataset and are not necessarily related to the true class sizes [26].

Oversampling of minority classes has been shown to be a particularly effective technique for improving accuracy when learning from small complex datasets [27]. However, the ratio of the minority class sizes to the majority class size is generally either set to be 1:1, set at an arbitrary ratio or set heuristically using the results on the validation set [25], [26], [27], [28]. Options one and two do not achieve the precision needed for small, complex datasets and option three would result in overfitting the validation set. Additionally, each fold will have markedly different relative class sizes in a small dataset. For these reasons, it is important that the relative class sizes produced through overfitting be dynamically adjusted in response to training on each individual fold and not adjusted based on the results on the validation set or using an arbitrary policy.

Good estimates of relative class sizes are available for rates of depression and most other diseases [29], [30]. Thus, it is reasonable to use the known depression class sizes, in training, as a heuristic measure of the optimality of each set of oversampling class ratios. In our work we adjusted the class ratios dynamically throughout training in order to train the classifier to predict the classes in approximately the known ratios. Obviously training the classifier to predict the depression classes in approximately the correct ratios results in an increase in the chance value calculated under a policy where depression classes are predicted randomly in these ratios. This is because the most frequently occurring class is predicted more often. For this reason a recalculation of chance for each test is included in the results section.

The noise in the data and bias towards class one was also overcome using neural network ensemble methods and by averaging the final probability output over each segment for each video [31], [32]. We ran the entire ensemble five times and took the average to ensure consistency of results. We averaged the probabilities over segments per subject per video to overcome any classification errors made because of slight fluctuations between segments in the same video for the same person due to external influences. The final reported accuracy was found by taking the average probability of each class taken over all segments for all subjects for each video. The average of the probabilities for all video segments for all subjects for each video was used to overcome any larger external influences that may have had an effect on one person but not the others. This final average was also used because the errors made by classifiers trained on individuals tended to be fairly uncorrelated (classifiers trained on different subjects had different sets of classes that they were able to distinguish between accurately). Thus, a simple majority vote ensemble method was an effective way to make sure that the overall result was more accurate than most of the individual results [32].

3.1. Model Description

The model design consisted of an ensemble method of feed forward neural networks. Three initial networks were trained five times to predict depression classes with each of the ECG, GSR and combined inputs individually. The final network used the 15 sets of class probabilities from the first networks (three different networks trained five times each) as input to learn a final class probability output five times. Figure 1. Both neural networks had one tanh hidden layer of size 100 and a softmax output layer 1. They were trained with gradient descent using backpropagation with the Binary Cross Entropy loss function. The loss function also included large Lasso regularizers to reduce overfitting:

$$Cost = \mathbf{t} log\left(\mathbf{y} + 1e - 10\right) + 1\left(sum\left(abs\left(\mathbf{w1}\right)\right)\right) + 1\left(sum\left(abs\left(\mathbf{w2}\right)\right)\right) \quad (1)$$

To achieve dynamic oversampling that could be adjusted during training a pool, that was twice the size of the majority class, was created using the original and synthetic data

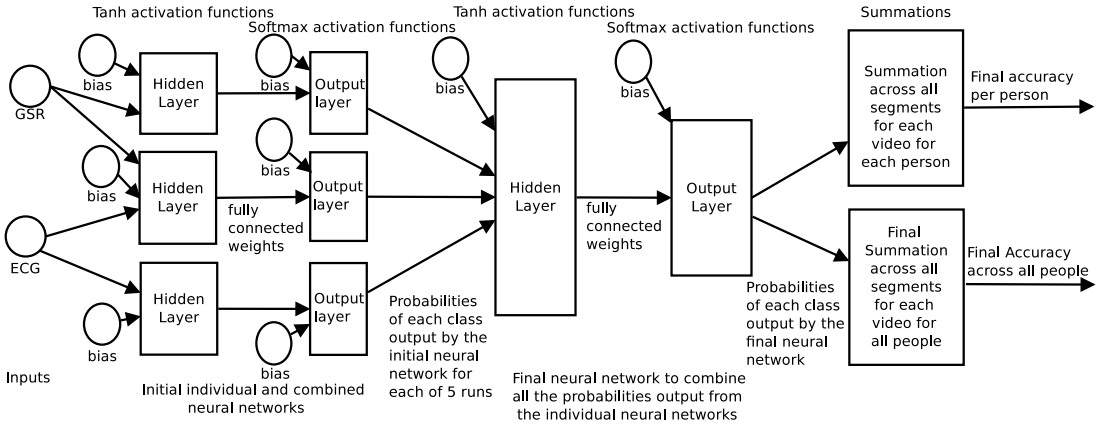

Figure 1. The full neural network ensemble model

for each class. The synthetic data was created by adding Gaussian noise with mean 0 and standard deviation 0.05 to the normalised original data in the minority classes. Adding noise to synthetic data has been shown to make classifiers more robust to noisy test data when trained [33]. The starting dataset size for each class was estimated by training each fold for 500 epochs for various class sizes and adjusting the class sizes, up for under representation or down for the reverse case, until the final prediction rates were within 15% of the known class sizes. The resulting class sizes were then averaged across all folds. The networks were trained on the starting class sizes for 500 epochs, which was observed via experimentation to be the point where the prediction rates had mostly settled. The class sizes were then adjusted every 50 epochs as follows:

$$sizes = sizes + (AS - PR) * SC$$

Sizes is an array of the class sizes as a proportion of class one. AS is the actual class sizes as a proportion of class one, PR is the rate at which each class was predicted in the most recent 50 epochs as a proportion of class one and SC is the rate of change. SC started at 0.1 and reduced by 2% every 50 epochs to prevent sizes from oscillating at the end of training. It was discovered through experimentation that this starting figure was high enough to allow the relative class sizes to quickly adjust to values that were close to optimal as determined by the known class size heuristic, but not so large as to cause large oscillations in prediction rates at the beginning of training.

For each subject the networks were trained for 1,500 epochs in total, with the final 400 epochs of training being fine tuning on only that subject. The relative class sizes were still adjusted every 50 epochs during fine tuning, but based only on that subject's predictions. Validation and testing was done based only on the physiological responses of the subject the classifier was fine tuned on. This can be thought of as pre-training on a related dataset and fine tuning on the actual dataset to be used for classification, as although the data from each subject is in the same format, their individual physiological reactions to the videos are quite different. This is especially pronounced in differences between males and females. When one network was trained for all subjects combined with no fine tuning the final classifier had too many higher depression class predictions for males and the reverse for females. Since all the input data was normalized within subjects this seems to show the males generally had a stronger physiological reaction based on their emotion/stress response to observing more depressed people.

4. Automated Processing of AVEC Videos for Class Prediction

For comparison with fully automated video classification methods we trained a traditional video processing algorithm to classify the videos directly, ignoring the human sensor data. Modern deep learning approaches require very large datasets to learn millions of parameters over the many layers needed to capture features directly from the pixels in videos, so are often not useful in the domain of medical diagnosis via videos of patients when datasets are very small. For this reason a more traditional method with pre-defined features was used following the method in [34].

For this experiment, we trained the model on videos from both the Train and Development sets and tested on the selected 29 video clips from the Test set. This was judged to be a fair way to proceed for a comparison with human performance, as our subjects had all had many years' experience observing fellow humans.

5. Results

5.1. Survey Results

Our observers were not good at classifying the videos by the depression status of the subject in the video. The overall average accuracy was 31.4%, which is just over the prima-facie chance level of 25%. Chance is expected to be 25% in this case as it is unlikely that the subjects had much awareness of the imbalance in the dataset. We note that

TABLE 1. FINAL ACCURACY RESULTS SUMMED PER VIDEO PER PERSON

Person	1	2	3	4	5	6	7	8	9	10
Accuracy	0.421	0.263	0.316	0.316	0.526	0.368	0.421	0.421	0.368	0.316

TABLE 2. FINAL ACCURACY PER CLASS

Class	1	2	3	4
Predictions	10	1	4	4
Accuracy	0.70	0	0.667	1.0

Figure 2. Successful predictions by men (blue, solid curve) vs. women.

the class distribution of 16 in the lowest category and 15 in the higher categories (6, 5, 4 in increasing severity of category), is a distribution which we would have expected to be better for our subjects as it would be more like their normal experience of interacting with fewer people with depression. Our subjects could have achieved a 48% score by simply choosing the lowest category for all cases. Our subjects consistently scored low, ranging from 24 to 38%.

We can see from Figure 2 that the (sorted by score) results for the five men and five women are not very different. Two men scored better than two women and vice versa. The mean correct is 9.2 for men, 9.0 for women. This difference is not statistically significant.

5.2. Neural Network Classification Based on Physiological Signals

The final accuracy for the sum of all subjects and segments for each video combined was 57.9% over the four classes. This is almost double chance, which is calculated at 32.5%. Chance is not 25% because we have deliberately trained the classifier to predict classes at a rate roughly in line with the known class ratios. This makes the calculation of chance under random assignment:

$$(14/29)^2 + (6/29)^2 + (5/29)^2 + (4/29)^2$$

as each class is picked based on its percentage in the dataset, then has that same percentage chance of being correct if it was assigned randomly. The prediction rates and accuracy rates for each class for the final prediction are shown in table 2

The final accuracy result is significantly better than the test subjects' conscious predictions, which was 31.4% over four classes. The accuracy rates on classes three and four are also a great deal higher than doctors' diagnosing patients at an accuracy of 47.3% on the general class of depression as discussed in section 1.

The high accuracy rates on classes three and four shows that the combination of dynamic adjustment of oversampling class ratios with the known class ratios as a heuristic for optimality of a given prediction ratio has had the desired effect of increasing accuracy on the more diagnostically useful minority classes. It has also kept the accuracy rate of the majority class well above chance. The accuracy rate of 0 for class two is concerning however, and is discussed in detail in section 5.3.

There was a large range in the accuracy from using different subject's physiological signals to classify the depression levels of the people in the videos they were watching. The results for each subject are in table 1 and range from roughly chance to significantly higher. As anticipated, the overall result was higher than each of the individual results presumably because it smooths out inconsistencies in the individual data. It is not surprising that individual results vary so much, especially considering the subjects were tested in only one session. The individual results would be affected by any external stress or emotion the subjects were feeling during testing. It is likely that some individuals have a stronger physiological response to seeing signs of depression, but more data incorporating testing over various sessions at various times would be needed to confirm that.

5.3. Analysis of Classification Errors

The types of mistakes the system made are of particular interest. Table 3 shows all mistakes the final classification algorithm made with the true class, predicted class and underlying raw depression scores. The final classification results misclassified all class two videos. Class two is the narrowest class in terms of raw depression score values, having only six possible values, compared to the next narrowest class which has half again at nine values and the widest class which has 35 raw values. In the final results, the videos with labels of class two and raw scores less than the middle value of the class were all classified as class one, whereas those with raw scores above the middle value were classified as class three or four. The probability of this happening by chance under random assignment is only 1.11%. This accounted for four out of the eight misclassifications in the final test sets. Only one video was incorrectly classified as class two, but because of the dynamic class size adjustments making sure all the neural network classifiers had approximately the correct ratio of class two predictions in the raw video segment classifications this added to the

TABLE 3. CLASSIFICATION ERRORS

True Class	3	1	1	3	2	1	1	1	2	3	1	1	1	1	2	1	4	4	2
Predicted Class	3	3	2	3	3	1	1	1	1	1	1	1	1	1	1	4	4	4	4
Depression score	21	0	2	25	19	12	0	3	15	25	7	5	10	3	15	0	33	30	17

TABLE 4. COMPARISON OF CLASS PROBABILITIES OUTPUT BY THE FIRST AND FINAL NEURAL NETWORKS. THE PROBABILITIES OUTPUT BY THE FIRST NETWORK AND THE FINAL NETWORK FOR THE FIRST 4 SEGMENTS OF A VIDEO WITH TRUE CLASS 4 AND ANOTHER WITH TRUE CLASS 1 ARE SHOWN.

Class	1	2	3	4
True class 4	0.296	0.300	0.130	0.274
first NN	0.294	0.242	0.0940	0.370
	0.310	0.269	0.115	0.306
	0.311	0.284	0.115	0.290
True class 4	0.223	0.330	0.029	0.419
final NN	0.206	0.344	0.023	0.426
	0.210	0.343	0.026	0.421
	0.237	0.348	0.033	0.381
True class 1	0.361	0.355	0.126	0.158
first NN	0.386	0.333	0.137	0.144
	0.391	0.320	0.142	0.147
	0.352	0.300	0.140	0.208
True class 1	0.362	0.347	0.111	0.181
final NN	0.379	0.338	0.112	0.171
	0.377	0.339	0.112	0.172
	0.357	0.357	0.110	0.176

noise in the final predictions and could have resulted in other misclassifications as shown in table 4.

It should also be noted that out of the four classification errors that were not from class two, three of them were in the eight shortest videos out of the 19 in the final test set. This may mean that the observers need a longer period of time to fully react to the videos.

We also measured the out-by-one accuracy rates, where the system predicted the wrong class, but the true class was only one higher or lower than the prediction. The final out by one accuracy rate for all subjects combined was 84.2% with chance being 64.3% The final number of errors where the prediction was more than one class different to the actual class was only 3 out of 19 videos.

It is also interesting to compare the outputs from the initial NNs with the final combined NN. Despite the predictions being reasonably close to the desired class ratios the raw probability figures for each class output by the initial NNs still showed a strong skew towards class one, with the probability figures generally being very similar for class one and the actual class when the actual class was class three or four. This was overcome with the final NN, with the network strongly preferring the minority classes in cases where their probabilities were similar to those of class one after the initial NNs. An example of this is shown in table 4. This table also shows that the networks are not able to predict class two at all, the predictions for this class are just adding noise to the final classifications and that this is not helped by the second NN. The final probabilities for class two are almost identical when the true class is one or four.

5.4. Comparison with Video Processing Algorithm

TABLE 5. VIDEO PROCESSING ALGORITHM ACCURACY PER CLASS

Class	1	2	3	4
Accuracy	0.714	0.167	0.60	0.50

The final accuracy result obtained with the video processing algorithm is 55.2%. Again this is significantly better than the human prediction by our subjects, and the chance value which is 33.3%.

The per class breakdown of the results obtained using the video processing algorithm are similar to our results based on the physiological signals. The algorithm is most correct on class one, with 71.4% correct, and also classifies classes three and four well, but is particularly bad at class two where it is only 16.7% correct. The results for each class are in table 5.

5.5. Experiments with Class Two Removed

Because of the above results showing the difficulty in classifying class two using both neural networks trained on physiological signals and automated video analysis, and the narrowness of class two compared to the other classes, we reran all the neural network algorithms using only three classes. Class two was absorbed by classes one and three depending which class their raw figures were closest to. The final accuracy with class two redistributed was 78.9% over the three classes. Chance for the three class classification problem was 43.7%.

The accuracy results for each subject over the three classes are shown in table 6. Again there is a large range in accuracies, but this time all but two subjects are above chance. In this case the two subjects whose results were just at chance level also showed no significant differences between the results for four classes and those for three classes. It can be concluded from this that they are either not able to pick up on universal cross-cultural depression indicators at all, or that they had other external influences that were affecting their stress or emotion levels during testing and there was too much noise in their readings.

The out-by-one accuracy rate for the three class scenario was 100%. No videos were misclassified by more than one class. Obviously this is less meaningful than in the four class scenario, as there are only two possibilities out of the nine combinations of true class and predicted class where the prediction could be out by more than one. However, the out by one accuracy of 100% is still significantly higher than our calculation of chance in this case which is 83.8%.

TABLE 6. FINAL ACCURACY RESULTS WITH CLASS TWO REMOVED SUMMED PER VIDEO PER PERSON

Person	1	2	3	4	5	6	7	8	9	10
Accuracy	0.474	0.526	0.526	0.526	0.842	0.579	0.421	0.579	0.421	0.632

5.6. Comparison of Results

TABLE 7. COMPARISON OF RESULTS.

	4 Class	Video Processing	3 Class
Accuracy	0.579	0.552	0.789
Chance	0.325	0.333	0.437
Reduction in Error Rate	0.376	0.328	0.625

Table 7 shows the results from the four class and three class classification based on physiological signals and the classification performed by the video processing algorithm. We have calculated the reduction in error rate, being the difference between the error rate produced by chance and the error rate produced by the algorithm as a proportion of the error rate produced by chance for each result. For example for the four class results based on the physiological signals the calculation would be:

$$\frac{0.579 - 0.325}{1 - 0.325} = 0.376$$

This allows us to compare the three different results with three different initial chance values. This shows that the percentage reduction in error rate over chance is much more significant for the three class scenario than the other two results. This was in line with our expectations given the analysis of the errors the algorithm was making in the four class results.

6. Conclusion

In the domain of detection of depression and other illnesses via videos of patients, datasets are generally extremely small due to privacy constraints. This makes the use of many traditional machine learning techniques problematic. The method presented in this paper combines human beings with the benefit of their many years of observation of human behaviors with machine learning methods to interpret each subject's basic physiological signals as reactions to their observations with significantly more accuracy than the subjects themselves could.

We have overcome the difficulties of training a neural network on a small noisy dataset with a large class imbalance in a novel way. We adjusted the oversampling class ratios both prior to and dynamically throughout training using known depression class sizes as a heuristic for optimal performance. This has been shown to be an effective way to train a neural network to fit a small, noisy and imbalanced dataset that is highly sensitive to relative class sizes. The techniques used added to the problem of overfitting which resulted in a large difference between the training accuracy and the final testing accuracy. However, due to the use of 10-fold cross validation allowing 19 videos to be used in the final test set without being used in the training or validation sets combined with the fact that the final test accuracies for some subjects were close to chance while others were significantly above chance, with none being significantly below chance, it is unlikely that the final test results are due to overfitting.

We have demonstrated the potential for depression severity to be predicted to significantly greater than chance levels by using neural networks to analyze the physiological responses of certain untrained observers watching videos of people with various levels of depression speaking in a language they do not understand. The results show the importance of more research in two areas 1) the existence of cross-cultural universal indicators of depression that are unaffected by language and cultural barriers, and 2) the possibility that there are some people with innate depression recognition capability whose automatic physiological reactions to observing universal depression indicators are pronounced enough that we can use them to classify depression levels to a significant level of accuracy even when the observers have no ability to consciously classify depression levels. Relating this work back to previous studies showing that physiological signals can be used to predict a subject's stress levels or emotional response to stimuli it seems clear that some of the observers are feeling a level of either stress or empathy in response to the body language and how the person being observed is speaking, with these levels directly related to the severity of the depression of the person being observed. A limitation of our work is that we cannot separate potential comorbidity with anxiety, so we may have found cross-cultural depression and anxiety indicators. This is a task for our future work. We note that all video processing algorithms would also suffer from the same issue for the entire AVEC competition.

The final accuracy figure result using the physiological signals of observers to predict depression levels over four classes was similar to but slightly higher than the result from the video processing algorithm trained on the entire AVEC dataset. The breakdown of accuracy levels for each class was also surprisingly similar. This seems to show that the subjects and the video processing algorithm are responding to similar information in the videos, which again makes it extremely unlikely the results are due to chance or noise in the dataset despite the small number of subjects. It appears that in this case the subject's innate response to universal cross-cultural indicators of depression was more effective at diagnosing depression than the specific training in classifying videos in the same language and style as

the test set which was utilised in the video processing algorithm. This is because the video processing algorithm is analyzing the primary data source, being the original videos, and the algorithm trained on the physiological signals is analyzing noisy secondary data. Thus the fact that the two very different methods produced such similar results is a strong indication that the use of cross-cultural depression indicators has the potential to be a useful tool in diagnosing depression, both for human in the loop algorithms and video processing algorithms.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[2] W. H. Organization, "Who depression factsheet," http://www.who.int/mediacentre/factsheets/fs369/en/, 2017, accessed: 2017-1-16.

[3] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, and B. Jönsson, "The economic cost of brain disorders in europe," *European Journal of Neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[4] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, "Cost of lost productive work time among us workers with depression," *Jama*, vol. 289, no. 23, pp. 3135–3144, 2003.

[5] U. of Pittsburgh, "Inventory of depressive symptomatology ids and quick inventory of depressive symptomatology qids," http://www.ids-qids.org/index2.html, n.d, accessed: 2017-1-26.

[6] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *The Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[7] P. Waxer *et al.*, "Nonverbal cues for depression," *Journal of Abnormal Psychology*, vol. 83, no. 3, pp. 319–322, 1974.

[8] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[9] D. Semple and R. Smyth, *Oxford handbook of psychiatry*. Oxford university press, 2013.

[10] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[11] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[12] W. Creemers, "On the recognition of emotion from physiological data," Ph.D. dissertation, Edith Cowan University, 2013.

[13] G. H. Zimny and E. W. Weidenfeller, "Effects of music upon gsr and heart-rate," *The American journal of psychology*, vol. 76, no. 2, pp. 311–314, 1963.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[15] C. Lee, S. Yoo, Y. Park, N. Kim, K. Jeong, and B. Lee, "Using neural network to recognize human emotions from heart rate variability and skin resistance," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*. IEEE, 2006, pp. 5523–5525.

[16] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 647–660, 2013.

[17] S. Wang, L. L. Minku, and X. Yao, "Online class imbalance learning and its applications in fault detection," *International Journal of Computational Intelligence and Applications*, vol. 12, no. 04, p. 1340001, 2013.

[18] E. A. Gehan, "The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent," *Journal of chronic diseases*, vol. 13, no. 4, pp. 346–353, 1961.

[19] R. Simon, "Optimal two-stage designs for phase ii clinical trials," *Controlled clinical trials*, vol. 10, no. 1, pp. 1–10, 1989.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[21] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *Journal of personality assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[22] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 70–70.

[23] Neulog, "Neulog, gsr logger sensor nul-217," https://neulog.com/gsr/, 2011, accessed: 2017-2-17.

[24] ——, "Neulog, electrocardiogram logger sensor nul-218," https://neulog.com/electrocardiogram/, 2011, accssed: 2017-2-17.

[25] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural networks*, vol. 21, no. 2, pp. 427–436, 2008.

[26] G. M. Weiss and F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[27] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Intl Conf. on Artificial Intelligence*, 2000.

[28] S. Barua, M. M. Islam, X. Yao, and K. Murase, "Mwmote–majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014.

[29] R. J. Anderson, K. E. Freedland, R. E. Clouse, and P. J. Lustman, "The prevalence of comorbid depression in adults with diabetes," *Diabetes care*, vol. 24, no. 6, pp. 1069–1078, 2001.

[30] D. C. Steffens, I. Skoog, M. C. Norton, A. D. Hart, J. T. Tschanz, B. L. Plassman, B. W. Wyse, K. A. Welsh-Bohmer, and J. C. Breitner, "Prevalence of depression and its treatment in an elderly population: the cache county study," *Archives of General Psychiatry*, vol. 57, no. 6, pp. 601–607, 2000.

[31] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[32] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[33] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[34] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 255–259.