Feature Selection on Thermal-stress Dataset

Xuyang Shen[1]

[1]Research School of Computer Science, Australian National University
[1]Xuyang.Shen@anu.edu.au

Abstract. Physical symptoms caused by high stress commonly happens on our daily life, which leads to the importance of stress recognition system. The goal of this study was to improve stress classification by selecting appropriate features from Thermal-stress data, ANUstressDB. We explored three different feature selection techniques including correlation analysis, magnitude measure, and genetic algorithm. In addition, SVM (Support vector machine) and ANN (artificial neural network) models are involved to measure three algorithms. Our result indicates genetic algorithm by applying ANN model can improve the predication accuracy around 20% compared with the baseline which without any feature selection. Moreover, magnitude measure is the best option regarding to the balance of computation time and performance.

Keywords: feature selection, genetic algorithm, neural network, magnitude measure, thermal-stress dataset

1 Introduction

Stress as an emotional language of humans represents the body's reaction to the environment. Light stress is actually beneficial to our body [1], while high stress may let us feel low energetic, even headaches. According to the latest report from the American Institute of Stress [2], job stress takes 46 percentages of stress in American adults and that it has escalated gradually over the past decades. Correctly judging employees' stress not only helps the leaders to adjust workload but also is favorable for the mental health of employees.

Traditional stress recognition [3] and detection system require physiological signals collected from special devices like blood pressure cuff, which is not suitable to utilize in the above circumstance. Therefore, Irani et al. purposed a real-time stress recognition system based on the physical appearance of objects, which only requires one RGB camera and one thermal camera [4]. The experiment result indicates the system with SVMs as main classifiers can achieve 89% accuracy, refreshing the state-of-the-art stress recognition system. In this paper, we focus on the same dataset as [4], but mainly focus on which feature selection algorithms can sufficiently improve the stress classification for SVMs and artificial neural networks (ANNs) respectively.

Reducing data resources from several physiological devices into two cameras improves the feasibility of the recognition system in real application but requires stricter data pre-processing. Feature selection as one pre-processing sub-task helps to filter out meaningless image information and hence improves the model performance [5]. We firstly explored correlation analysis from information theory which aims to measure the strength of the linear relationship between features [6]. Apart from that, we also employed the magnitude measure that is originally applied to prune neural networks [7]. It helps us to consider the feature relationship from learned neurons. Considered that the accuracy of magnitude measure is affected by the local-minima issue of ANNs, genetic algorithms are introduced as the third method. Compared to the ANNs, genetic algorithms use stochastic search to approach the global optimization, which theoretically can present the most accurate feature selection result.

2 Method

2.1 Thermal-stress Dataset

There are two accessible versions of the thermal-stress dataset from ANUStressDB [4], the raw data, and PCA processed result. The dataset was generated from an HCI experiment about stress stimulator, which involved 31 subjects and their facial information was recorded by a Microsoft webcam and FLIR camera at 30 frames per second at 640x480 pixels [4]. As the raw data is video-based and lacks classification labels, we have to give up this version in the experiment.

The second version of the thermal-stress dataset contains 620 records with 5 features extracted from RGB facial image and 5 features extracted from the thermal camera as described above. Specifically, these 10 features vectors derive from the result of principal component analysis (PCA) to the original raw data. Additionally, it also gives us a classification label of each row that is either stressful or calm. The labels were manually constructed in the experiment environment setting and validated by the questionnaire survey.

Although the dataset is well-balanced between two classes, it is still hard for the learned model to make a correct prediction. From two visualization (Fig. 1), it indicates a high-class similarity between stressful and calm data. Most of the data points from two different classes are overlapped. It might be caused by applying the PCA to the original raw data. Most spatial information is lost after the main feature extraction, as well as troubling the network prediction.

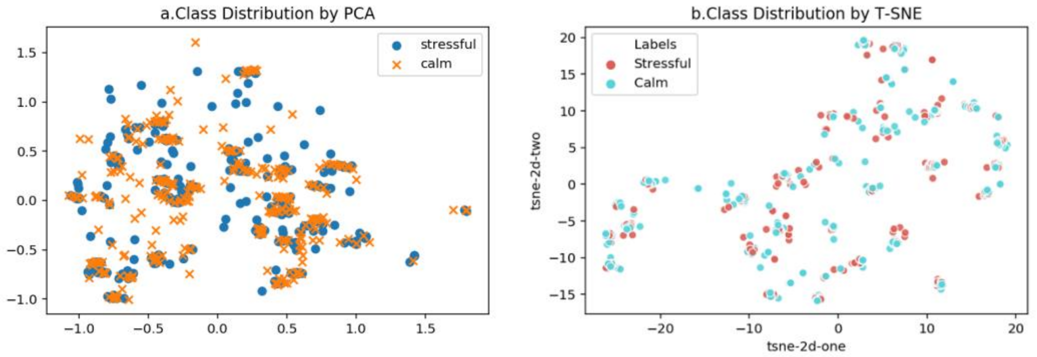

Fig. 1. PCA and t-SNE dimensional reduction techniques are applied to visualize the original ten-dimension data

The results of experiments also prove that the trained model fails in prediction (Table 1). In 10-18-16-8-2 model, high training accuracy indicates the networks learn several features from training data, but still failed in the prediction. If we decrease neurons and reduce the hidden layer into 2 layers, the 10-10-6-2 model cannot generalize well on the given train set, and hence test accuracy is worse than the previous one.

Table 1. Training and Testing Accuracy of two mdoels and validation methods (report the average result among 20 times running). 10-18-16-8-2: 10-input neurons; 18-hidden1 nurons; 16-hidden2 nurons; 8 hidden3 nurons, 2 output neurons. 10-10-6-2: 10-input neurons; 10-hidden1 nurons; 6-hidden2 nurons; 2 output neurons.

Model	Validation Measure	Training Accuracy	Test Accuracy
10-18-16-8-2	Training: 0.7 Test: 0.3	92.84%	53.50%
10-18-16-8-2	5-Fold Cross-validation	91.71%	52.10%
10-10-6-2	Training: 0.7 Test: 0.3	67.3%	51.2%
10-10-6-2	5-Fold Cross-validation	63.2%	49.7%

2.2 Network Structure and Hyper-parameters

2.2.1 Support Vector Machine, SVM

SVM is a supervised machine learning model that performs both linear classification and non-linear classification through kernel trick. Considered that SVM can be successfully applied in various applications with different vision algorithms [8], we decide to explore it as a baseline with ANNs.

The SVM utilized in this paper is from the Sklearn package with leaving all hyper-parameters as default except the random state (Table 2). Since to maintain the fairness of each generation in genetic algorithms, we manually fix the random state.

Table 2. Hyper-Parameter of SVM

Kernel	Degree	Gamma	Degree	Random State
rbf	3	auto	3	22

2.2.2 Artificial Neural Network, ANN

Based on the property of the dataset, we attempted several configurations of fully connected neural networks to recognize the stress (same as traditional multi-layer perceptron, MLP). The MLP described here is to distinguish from the convolutional neural network whose weight is shared in local connected layers.

The first typical network topology is 10-10-6-2, being 10 input neurons, 2 hidden layers, and 2 output neurons, but it fails in generalizing to data (Table 1.). Considered that the raw data is image-based, which contains a complex relationship between local pixels. Therefore, a shallow neural network cannot achieve an ideal performance. The second one is 10-18-16-8-2 (Fig. 2), which is added a few neurons in each hidden layer and one extra hidden layer compared to the previous shallow MLP. Partial results of experiments indicate it performs better in prediction but appear to overfit to the training data (Table 1.).

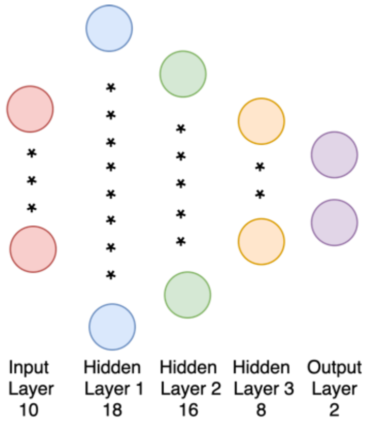

Fig. 2. Four-Layer neural network (ANN).

Table 3. Hyper-Parameter of ANNs

Optimization	Learning Rate	Weight Decay	Epochs
Adam	0.0008	0.0003	10000
Validation	**Activation Function of Hidden Layer**	**Activation Function of Output Layer**	**Random Seed**
Training: 0.7 Test: 0.3	ReLU	SoftMax	7

One major issue of neural network design is always to optimize the hyper-parameter to achieve the balance between overfitting and generalization. Considered that the shallow network cannot well generalize to thermal-stress data, we finally decide to use second network topology in the later experiment (Fig. 2). The configuration of the hyper-parameter of ANNs is also adjusted to improve the efficiency of training in genetic algorithms (Table 3). For instance, (Table 1) indicates the train-test separation is able to reflect similar test accuracy as the cross-validation, but the former one highly reduces the computation time. Besides, compared to the SGD (Static Gradient Descent), Adam can gain better performance in the non-convex problem in a shorter time [9].

2.3 Feature Selection Techniques

2.3.1 Correlation Analysis

Correlation indicates the dependence or relationship between two data variables from a statistical perspective. The result of the correlation formula is located between -1 and 1, which represents the dependence between two variables from negative correlation into positive correlation.

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_x\ \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y\ ]}{\sigma_x\ \sigma_Y} \tag{1}$$

$$where\ E\ means\ the\ excepted\ values, and\ cov\ meas\ convariance$$

We applied the correlation analysis to perform feature selection by filtering high correlated variables from the statistical property of the dataset, and the result is displayed in the third section.

2.3.2 Magnitude Measure

Gedeon (1997) [7] purposed the magnitude measure of input neuron to output neuron based on the measure of input neuron to hidden layers [10].

$$Q_{ik} = \sum_{r=1}^{nh}(P_{ir} \times P_{rk})\ ,where\ P_{ij} = \frac{|W_{ij}|}{\sum_{p=1}^{ni}|W_{pk}|}\quad P_{jk} = \frac{|W_{jk}|}{\sum_{r=1}^{nh}|W_{rk}|} \tag{2}$$

P_{ij} refers the influence of input neuron i to the hidden neuron j, and P_{jk}measure the effect of hidden neuron j to the output neuron k.

It is a static analysis applied for the input neuron on the output neuron. Compared with the correlation analysis, magnitude measure can analyze deeply into the weight between layers which directly indicates the impact of input neurons to others. As a result, we can perform feature selection based on the ranking of input variables. It also concludes a highquality result in the experiments, which is beyond our expectations. On the other hand, the calculation of the hidden layer to the hidden layer is complicated for the neural networks which have more than one hidden layer.

2.3.3 Genetic Algorithm

From the experiments of magnitude measure, we found it has large limitations including it is not applicable to SVMs and highly non-determined in ANNs, since it only performs the local search. Therefore, we employed the genetic algorithm with SVMs and ANNs respectively to further analyze the feature selection.

Genetic algorithms as one large class of evolutionary algorithms perform stochastic search for an optimal solution based on 5 components (described in Fig. 3). Fitness function is the main module among these 5 components, which responsible to evaluate each chromosome (feature selection). Typically, hybrid values from train accuracy and test accuracy of models are selected as fitness scores, whereas it is not suitable in our experiment since the model training on thermal-stress dataset easily causing overfitting issues. As a result, we only pick test accuracy, in the meanwhile, we also control all the random values in model initialization and train-test dataset separation.

Not only we explore different fitness function as commented above, but we also experimented three different selection and crossover techniques:

1. Proportional selection one parent and randomly select the other parent: the principle of this selection algorithm is chromosomes with large fitness value having a higher probability to pass their genes to offspring.

$$\varphi_s(x_i(t)) = \frac{f_r(x_i(t))}{\sum_{l=1}^{n_s} f_r(x_l(t))} \tag{3}$$

$$where\ n_s is\ the\ total\ number\ of\ indiviuals\ in\ the\ population;\ \varphi_s(x_i) is\ the\ probability$$
$$that\ x_i\ will\ be\ selected;\ f_r(x_i) is\ the\ scaled\ fitness\ value\ of\ x_i$$

2. Tournament selection for both parent: the best individual in the group of n_ts chromosomes will become the parents ($set\ n_{ts} = 0.6 * popultion$). The principle of tournament selection is to limit the chance of the best individual to dominate.
3. Hall of Fame selection one parent and the proportional selection for the other parent: only best individual of each generation is selected to be inserted into the hall of fame which becomes a parent pool for crossover operator.

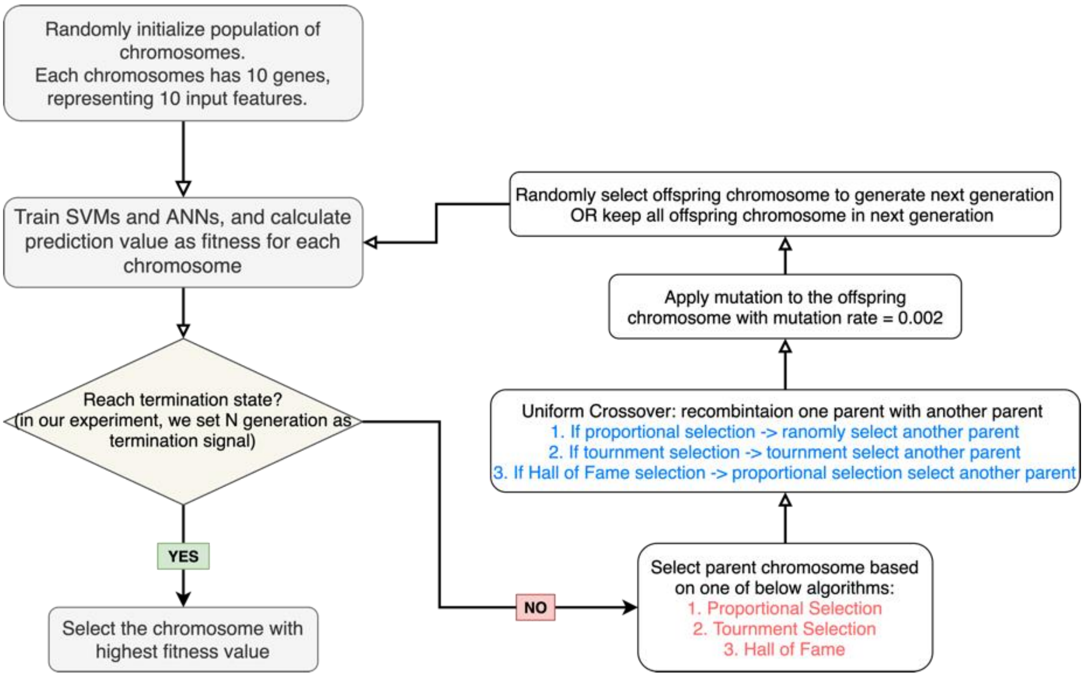
Randomly initialize population of chromosomes.
Each chromosomes has 10 genes, representing 10 input features.
Train SVMs and ANNs, and calculate prediction value as fitness for each chromosome
Reach termination state?
(in our experiment, we set N generation as termination signal)
YES
Select the chromosome with highest fitness value
NO
Select parent chromosome based on one of below algorithms:
1. Proportional Selection
2. Tournment Selection
3. Hall of Fame
Uniform Crossover: recombintaion one parent with another parent
1. If proportional selection -> ranomly select another parent
2. If tournment selection -> tournment select another parent
3. If Hall of Fame selection -> proportional selection select another parent
Apply mutation to the offspring chromosome with mutation rate = 0.002
Randomly select offspring chromosome to generate next generation
OR keep all offspring chromosome in next generation

Fig. 3. Combine SVMs and ANNs with Genetic Algorithm to explore feature selection

3 Results and Discussion

3.1 Baseline of Feature Selection Experiment

In order to reflect the influence of different feature selection algorithm, we compute the baseline of SVM and ANN model, which takes all of ten features as inputs (Table 4). The data in the table is the result of one run since all random states were set in a fixed value.

Table 4. Baseline of SVM and ANN.

Model	Final Train Accuracy	Test Accuracy	Platform
SVM	100%	49%	MacOS
ANN	95.78%	54.12%	Linux
ANN	95.33%	46.47%	MacOS

The result indicates that the SVM model takes ten input features with default hyperparameters that can reach 100% final training accuracy and 49% test accuracy. Additionally, the test accuracy of the ANN model is 54% and 47% on two different platforms respectively. Considered the genetic algorithms with the ANN model need large computation time, we prepared two computer resources to train them. Despite fixed random states being configured, the actual model training still depends on the platform and the version of the main packages. Nevertheless, it does not violate the fairness of this experiment since the difference between them is slight (Fig .4).

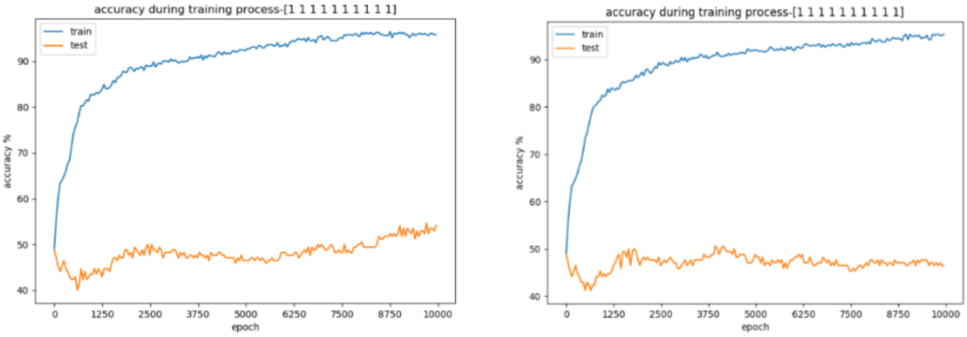

Fig. 4. Training monitor of ANN model in Linux platform (left) and MacOS platform (right)

3.2 Correlation Analysis on Feature Selection

In this section, features are selected based on their correlation to labels. The overall correlation among these ten features are all lower than 0.1, and there are 4 attributes even lower than 0.005. Theoretically, the absolute value of correlation less than 0.19 is considered as low correlation, hence, all input features can be abandoned based on it.

In order to quantify the performance of correlation analysis, we further explore the influence of removing low correlated features to prediction accuracy (Fig. 5). The result indicates correlation analysis cannot help to gain better performance of SVM, and it contributes few to the ANN as well. Therefore, we conclude that the correlation analysis cannot guide the feature selection in positive ways.

Table 5. Correlation between different feature and labels, ranking from low correlation to high correlation.

Thermal_4	Thermal_5	RGB_4	Thermal_1	RGB_1	RGB_2	RGB_3	Thermal_3	Thermal_2	RGB_5
-0.0015	0.0025	0.0039	-0.0040	0.0054	0.0055	-0.0093	0.010	0.015	0.023

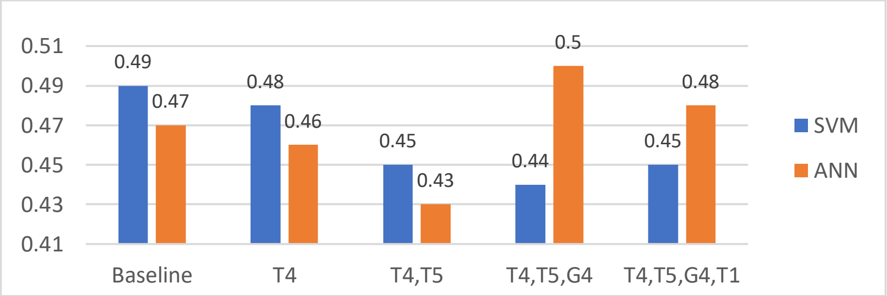

Fig. 5. Test accuracy of removing the lowest correlated feature to top 4 lowest correlated features T4: remove Thermal_4 feature; T4,T5: remove Thermal_4 and Thermal_5 features, and etc. Besides, the training platform of ANN is Linux.

3.3 Magnitude Measure on Feature Selection

Compared to the correlation analysis, the magnitude measure focuses on the impact of dataset property on the behaviour of neural networks. Based on the formula in section 2.3, we calculated the average magnitude value of input neurons behaviour to the hidden neurons in 20 runs (Table 6) and applied the results to feature selection experiments (Fig. 5).

The result indicates that “RGB_3”, “Thermal_1”, “RGB_1” negatively contribute to the ANN training and prediction, which are shown as less important attributes in the magnitude measure. On the contrary, RGB_4 is incorrectly marked as less important data to hidden neurons by the magnitude measure. The reason causing this issue might be the low stability of magnitude measure. This method targets the weight’s calculation of trained networks, which can be easily affected by the initial weights assigned to each hidden neurons and final termination judgment. Overall, magnitude measure is able to provide valuable guidance to feature selection for ANN models.

Table 6. Magnitude of each attribute, ranking from low to high. To obtain a reasonable result, we set all random states in random values.

RGB_3	Thermal_3	RGB_4	Thermal_1	RGB_1	Thermal_5	RGB_2	Thermal_4	Thermal_2	RGB_5
1.693	1.745	1.761	1.773	1.793	1.811	1.838	1.843	1.850	1.887

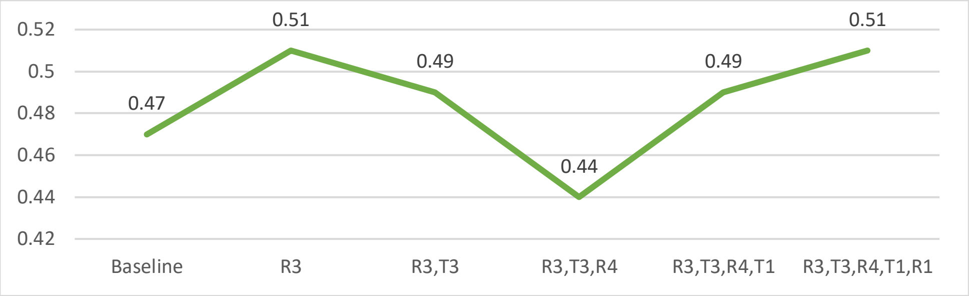

Fig. 6. Test accuracy of removing least top1 – top 5 important variables. The training platform of ANN is Linux.

3.4 Genetic Algorithm on Feature Selection

Through the previous experiments, we found the magnitude measure is sensitive to the neuron weights which are updated through gradient descent in ANN training. As it is known, backpropagation with gradient descent is a local search and easily trouble with local-minima issue. Besides, the magnitude measure also is limited to the learning model, which cannot be applied to SVMs in this experiment. Therefore, this feature selection algorithm seems not to be a sensible choice for the thermal-stress dataset. In the next stage, we explored a global search to feature selection which is genetic algorithms from evolutionary algorithms. Follow the procedures introduced in section 2.3, we apply genetic algorithms into SVMs and ANNs respectively to further analyse the feature selection. Evaluating each chromosome with the test accuracy of ANN model requires 10 seconds, which means 60 population in one generation needs 10 minutes. Hence, we employed two computer resources, which causes different baseline in (Table 7).

From the result of the experiment (Table 7), the GA algorithm applied with the ANN model has a large improvement to the test accuracy, which achieves 20% at maximum. One reason behind this could be the stochastic global search with a high initial diversity has a higher probability to find the best optimization than the gradient descent. Additionally, the thermal-stress dataset is low correlated between attributes and labels (Table 5), which exaggerates the difference between local search and global search. Apart from that, we also found that the initial diversity of the population has different influences regarding each selection strategy of genetic algorithms. Specifically, the proportional selection is more sensitive to the initial diversity than then tournament selection, hence, properly dominating superior genes is beneficial to the stability of genetic algorithms. Furthermore, it is not difficult to find out that there is more than one combination of feature selection resulting in higher prediction accuracy than baseline, they all commonly select “RGB_2”, “RGB-4”, “Thermal_2”, and “Thermal_3”.

Table 7. GA algorithm applied with ANN. DNA order: [RGB_1, RGB_2, RGB_3, RGB_4, RGB_5, Thermal_1, Thermal_2, Thermal3, Thermal_4, Thermal5] 1 means this attribute is selected and vice versa

GA	**Proportional selection**		**Hall of Fame selection**	**Tournament selection**	
Population | Generation	80 | 78	60 | 100	60 | 100	60 | 100	80 | 78
DNA	01010 01101	01101 11011	00011 11110	11010 11110	11111 10011
Test Accuracy	0.56	0.51	0.55	0.54	0.53
Baseline	0.47	0.54	0.54	0.47	0.47
Improvement	+0.09	- 0.03	+0.01	+ 0.07	+ 0.06

Apart from applying genetic algorithms with the ANN model, we also explored the SVM model (Table 8). Performing feature selection with genetic algorithms can improve 8% test accuracy of SVM at maximum. In conclusion, the genetic algorithm is the best option for either ANN models or SVM models. Not only it has wide applicability, but it also improves the model performance.

Table 8. GA algorithm applied with SVM. DNA order: [RGB_1, RGB_2, RGB_3, RGB_4, RGB_5, Thermal_1, Thermal_2, Thermal3, Thermal_4, Thermal5] 1 means this attribute is selected and vice versa

GA	Proportional selection	Hall of Fame selection	Tournament selection
Population | Generation	60 | 100	60 | 100	60 | 100
DNA	01110 10010	01111 11011	01011 00011
Test Accuracy	0.52	0.53	0.50
Baseline	0.49	0.49	0.49
Improvement	+0.03	+0.04	+ 0.01

4 Conclusion and Future work

Feature selection as one of the essential parts of data pre-processing can improve the prediction accuracy for both SVN and ANN models. Different feature selection algorithms may contribute differently to the improvement of model performance. As we explored in the experiments, correlation analysis focuses on the statistical property of the dataset to provide selection guidance, which takes the least computation power but is sensitive to the dataset. Since the correlation between attributes and labels is low in the thermal-stress data, this algorithm cannot guide sufficient information for us to choose features. The second algorithm we explored is magnitude measure which is commonly applied to prune hidden neurons. This algorithm is moderate among the three algorithms we discussed in section 3 with a balanced ratio of performance and efficiency. However, the same as correlation analysis, magnitude measure only provides feature selection guidelines to us, and we have to manually decide which feature to keep or remove. Eventually, genetic algorithms are able to select the most appropriate features for both SVM and ANN models, but require huge computation power. For instance, 1000 generations with 100 initial population of genetic algorithm with ANN might take 10 days to simulate the evolution. However, we also discovered that the genetic algorithm with tournament selection is less sensitive to population and generation, which helps us to gain an ideal performance in a limited time.

To further explore the balance between time and performance in feature selection by the genetic algorithm can be considered as future work. Apart from different selection or crossover configuration of genetic algorithm, we can also consider this from the perspective of fitness functions. In addition, feature selection is also sensitive to the dataset. More experiments on different stress dataset are also required to validate our results.

5 References

[1] "Stress Symptoms: Effects of Stress on the Body," 2020. [Online]. Available: https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body#1.

[2] "Stress Effects - workplace stress," 22 2 2020. [Online]. Available: https://www.stress.org/workplace-stress.

[3] T. Chen, P. Yuen, M. Richardson, G. Liu and Z. She, "Detection of Psychological Stress Using a Hyperspectral Imaging Technique," in *IEEE Transactions on Affective Computing*, 2014.

[4] R. Irani, K. Nasrollashi, A. Dhall, T. B. Moeslund and T. Gedeon, "Thermal super-pixels for bimodal stress recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Oulu, Finland, 2016.

[5] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[6] J. Gareth, W. Daniela, H. Trevor and T. Robert, An Introduction to Statistical Learning, Springer, 2013.

[7] T. D. Gendeon, "DATA MINING OF INPUTS: ANALYSING MAGNITUDE AND FUNCTIONAL MEASURES," *International Journal of Neural Systems,* vol. 08, no. 02, pp. 209-218, 1997.

[8] M. Tanini, G. Zhao, S. Z. Li and . M. Pietikainen, "Facial expression recognition from near-infrared video sequences," in *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA, 2008.

[9] D. P. Kingma and J. L. Ba, "AAdam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980.*, 2014.

[10] P. M. Wong, T. D. Gedeon and I. J. Taggart, "An improved technique in porosity prediction: a neural network approach," 1995.