Genetic Algorithms for Feature Selection

Tom Gedeon
Research School of Computer Science
Australian National University
tom@cs.anu.edu.au

based on slides by Nandita Sharma

Overview

- What is feature selection?
- Feature selection for stress recognition
- Evolutionary Algorithms (EAs) for feature selection
- Comparison of feature selection methods

Feature Selection

- A simple data set:

	Class A			
ID	F1	F2	F3	F4
1	0.80	0.50	0.37	0.48
2	0.91	0.54	0.16	0.44
3	0.63	0.88	0.54	0.25
4	0.70	0.52	0.27	0.48
5	0.77	0.03	0.02	0.27
6	0.64	0.36	0.19	0.09

	Class B			
ID	F1	F2	F3	F4
7	0.98	0.12	0.74	0.89
8	0.64	0.38	0.56	0.61
9	0.45	0.20	0.86	0.08
10	0.04	0.26	0.32	0.39
11	0.38	0.07	0.64	0.97
12	0.94	0.81	0.51	0.92

- A classifier could be used to separate the classes using all features - any problems?
  - Would a smaller set of features suffice?

Feature Selection

Class A

ID	F1	F2	F3	F4
1	0.80	0.50	0.37	0.48
2	0.91	0.54	0.16	0.44
3	0.63	0.88	0.54	0.25
4	0.70	0.52	0.27	0.48
5	0.77	0.03	0.02	0.27
6	0.64	0.36	0.19	0.09

Class B

ID	F1	F2	F3	F4
7	0.98	0.12	0.74	0.89
8	0.64	0.38	0.56	0.61
9	0.45	0.20	0.86	0.08
10	0.04	0.26	0.32	0.39
11	0.38	0.07	0.64	0.97
12	0.94	0.81	0.51	0.92

- Which features can determine the different classes above? F3
  - Can these features improve classification performance?

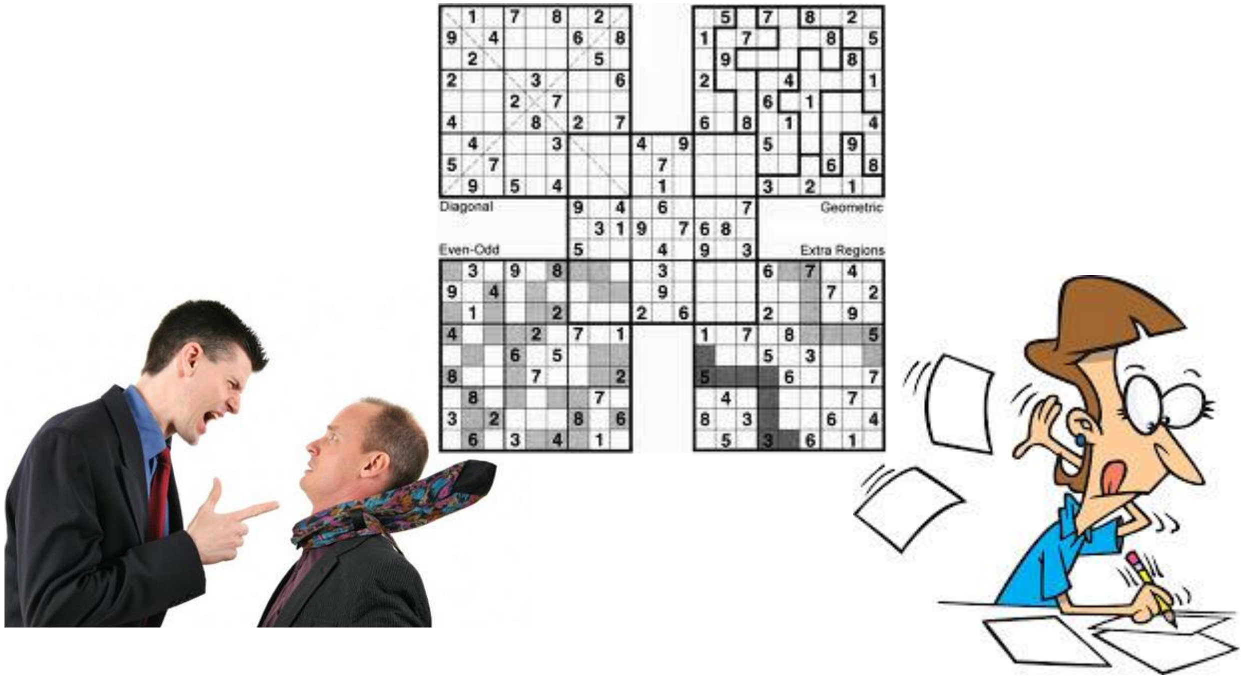
- Now, suppose you have a data set that is 100 times larger.
  - What characteristics do you want in your feature selection method?

Stress Recognition

- Aim: model stress using physiological and physical sensor signals to recognise stress
- Models based on artificial neural networks (ANNs) & support vector machines (SVMs)
- Hundreds of stress features
  - redundant and irrelevant features → motivates feature selection
- Genetic algorithm & correlation methods for feature selection

Stress

- Reaction or response to the imbalance caused between demands & resources available to a person



Stress Measures

Traditional measures

- Interviews, self-assessment reports (subjective)
- Task performance

Physiological measures

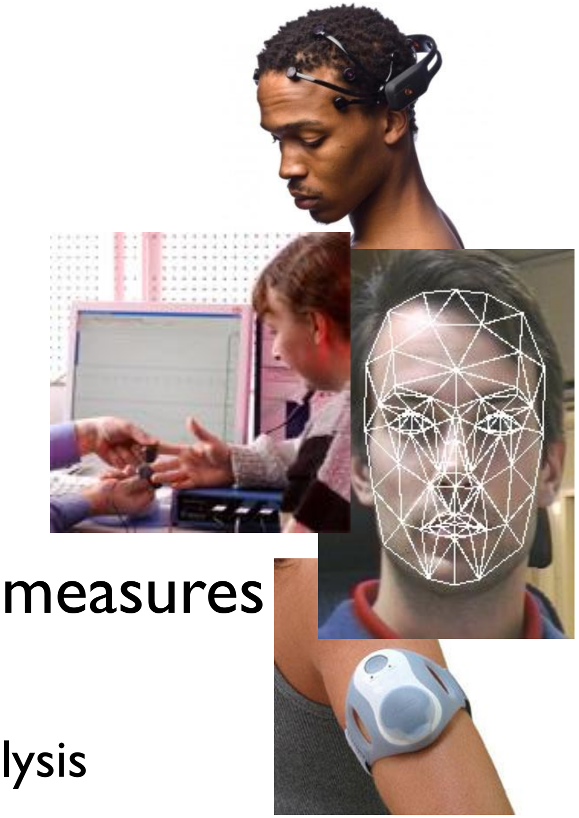
- Heart rate, brain waves, skin conductivity

Physical measures

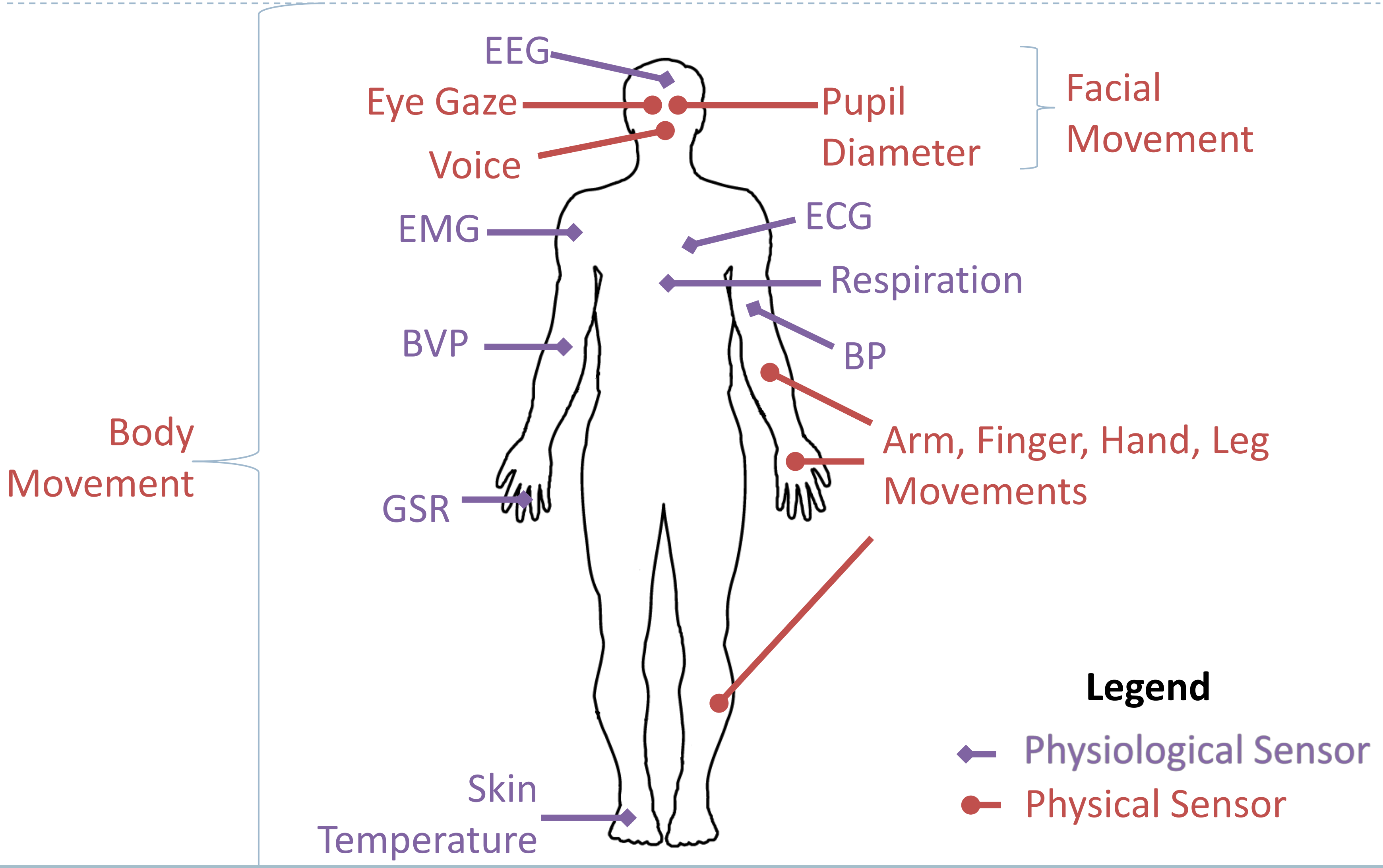
- Body movement, face & eye tracking, voice

Reasons for using physiological & physical measures

- Objective
- Provides data at a higher granularity for detailed analysis



Physiological & Physical Signals



EEG

Eye Gaze

Voice

Pupil Diameter

Facial Movement

EMG

ECG

Respiration

BVP

BP

Body Movement

Arm, Finger, Hand, Leg Movements

GSR

Skin Temperature

Legend

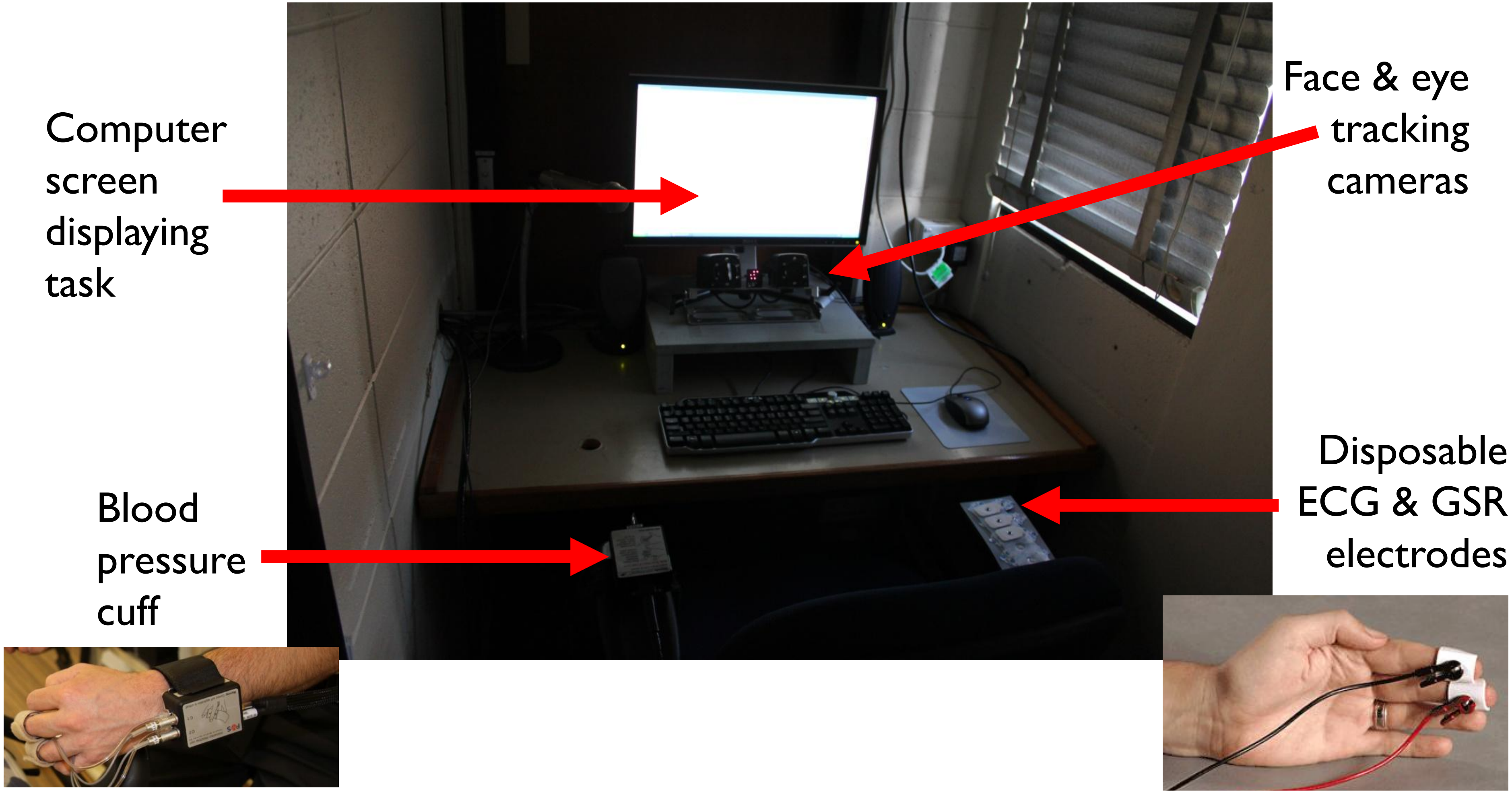
◆— Physiological Sensor

●— Physical Sensor

Stress Data Collection: A HCI Experiment

1. Present experiment requirements to participant
2. Participant provides consent
3. Equipment
   - Physiological signals – ECG, GSR, BP
   - Physical signals – Eye gaze, Pupil diameter
4. Participant does some task
5. Assessment & survey

Participant's Room



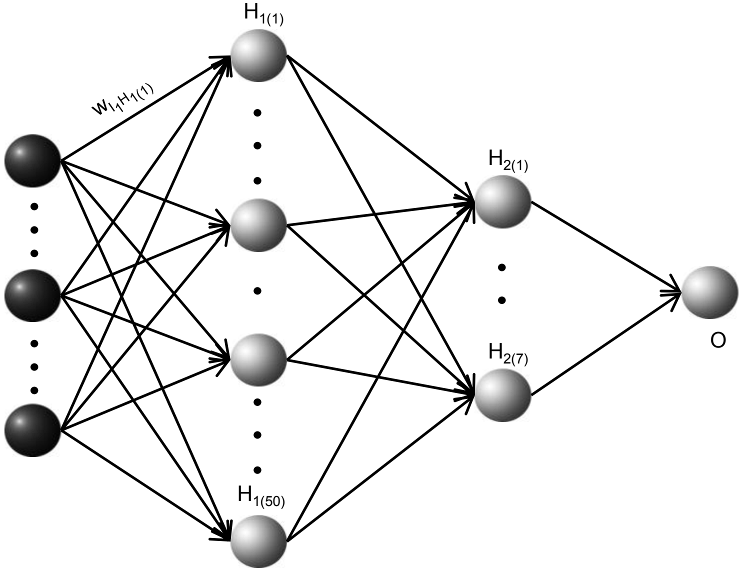
Computer screen displaying task

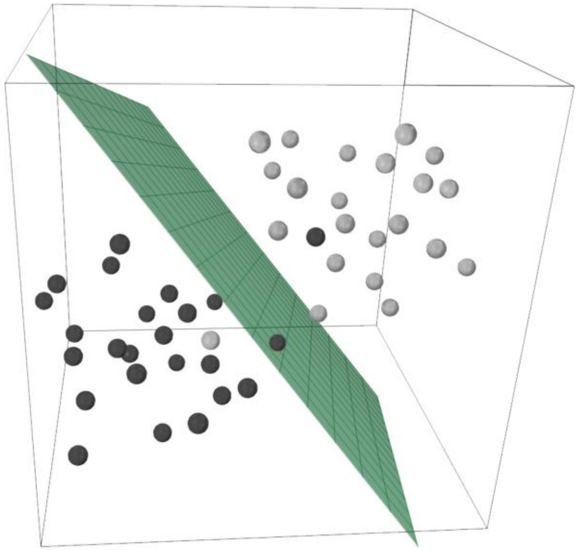
Face & eye tracking cameras

Disposable ECG & GSR electrodes

Blood pressure cuff

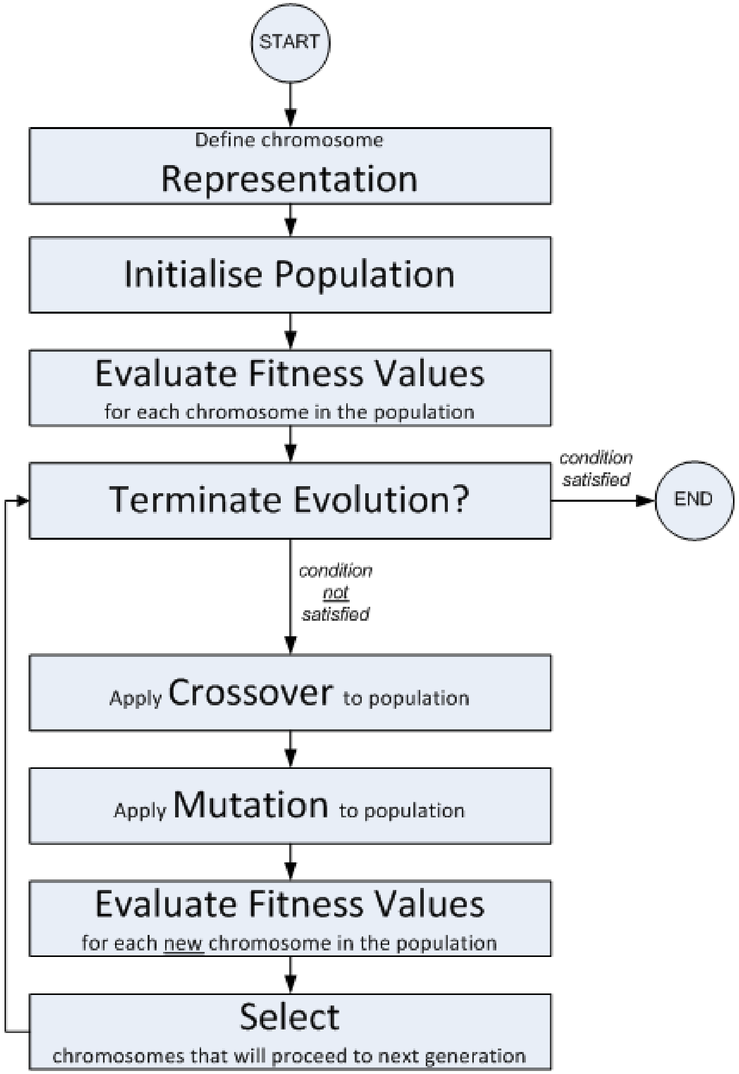
ANN & SVM Stress Models



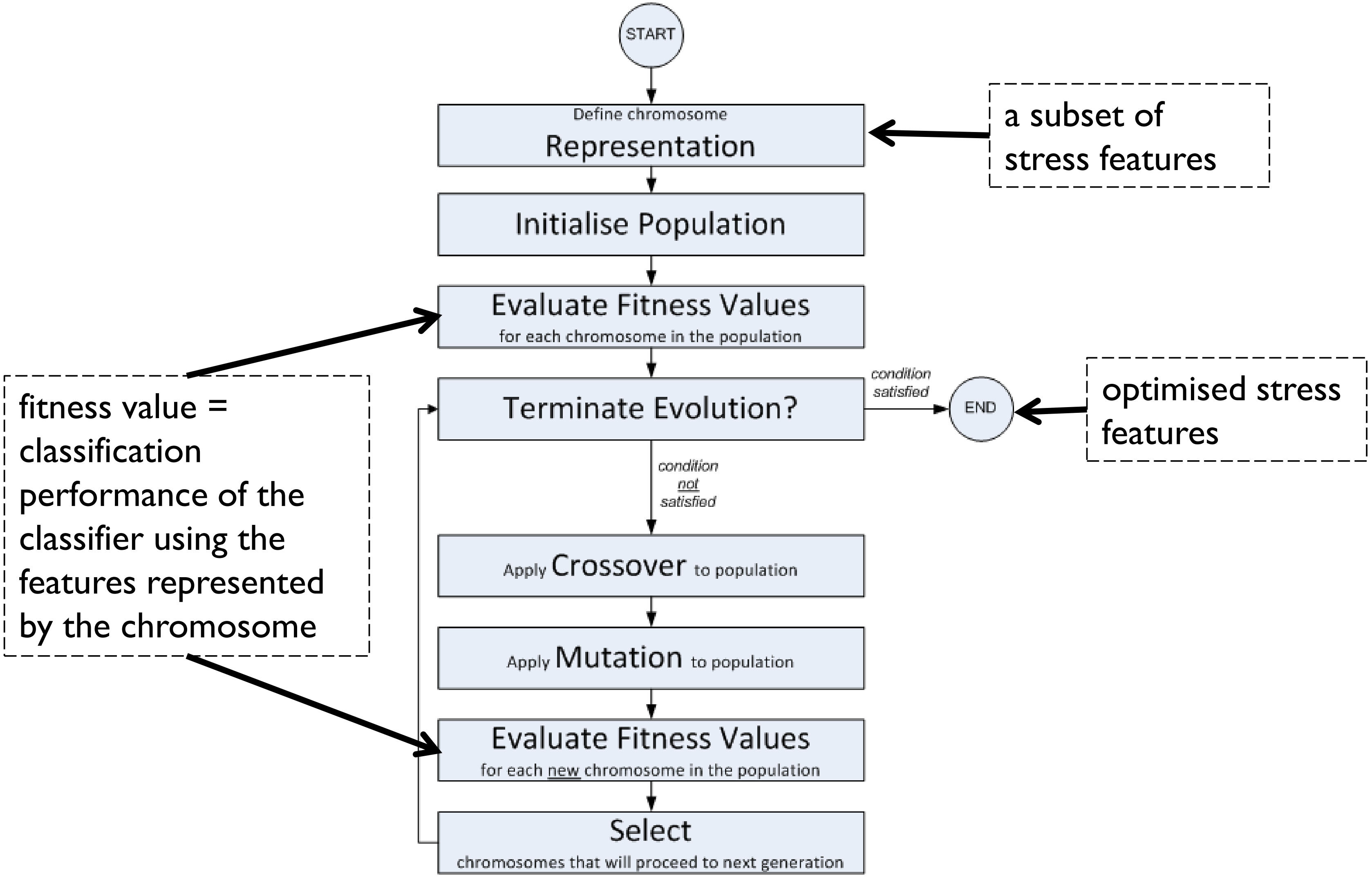


- What are the problems with using all the stress features?
  - Large ANN
  - Could negatively affect model performance
  - Longer computation times
- Solution? Optimise features

Genetic Algorithm


START
Define chromosome
Representation
Initialise Population
Evaluate Fitness Values
for each chromosome in the population
Terminate Evolution?
condition satisfied
END
condition not satisfied
Apply Crossover to population
Apply Mutation to population
Evaluate Fitness Values
for each new chromosome in the population
Select
chromosomes that will proceed to next generation

Genetic Algorithm for Feature Selection


START
Define chromosome
Representation
Initialise Population
Evaluate Fitness Values
for each chromosome in the population
Terminate Evolution?
condition satisfied
END
condition not satisfied
Apply Crossover to population
Apply Mutation to population
Evaluate Fitness Values
for each new chromosome in the population
Select
chromosomes that will proceed to next generation

a subset of stress features

optimised stress features

fitness value = classification performance of the classifier using the features represented by the chromosome

Chromosome

- A chromosome represents a feature subset
- Features in a subset are used as classifier inputs

F1	F2	F3	F4	F5	F6	F7	…	Fn
1	0	0	0	1	1	0	…	1

where Fi = ith feature

indicates F7 feature is not in the subset

indicates Fn feature is in the subset

Correlation Method for Feature Selection

- Pseudo-independent Feature Selection algorithm (PISA)
- Based on *correlation coefficients*
  - measure for strength of linear relationship between features
- Let X & Y be features

  x_t & y_t be feature values at time-step t in X & Y

  σ_X & σ_Y be standard deviations

  r_{XY} = correlation coefficient

$$r_{XY} = \frac{\sum_{t=0}^{T}(x_t - \bar{X})(y_t - \bar{Y})}{(T+1)\sigma_X \sigma_Y} \qquad |r_{XY}| \leq 1$$

Stress Recognition Models

1. **ANN**: all stress features were inputs
2. **PISA+ANN**: ANN with inputs selected by PISA
3. **GA+ANN**: ANN with inputs selected by a GA
4. **SVM**: all the stress features were inputs, like the ANN
5. **PISA+SVM**: SVM with inputs selected by PISA
6. **GA+SVM**: SVM with inputs selected by a GA

Results

- Model performance using 10-fold cross-validation

Classification Performance Measure	ANN	PISA+ANN	GA+ANN	SVM	PISA+SVM	GA+SVM
Accuracy	0.68	0.76	0.82	0.67	0.80	**0.98**
F-score	0.67	0.79	0.82	0.67	0.79	**0.98**

- Classifiers with feature selection methods performed better
- GA hybrid models performed the best

Summary

- Purpose for feature selection
- Feature selection for optimising model performance
- A real-world application of EAs – stress recognition
- EAs are good for selecting the more relevant features & reduce the use of redundant features for modelling