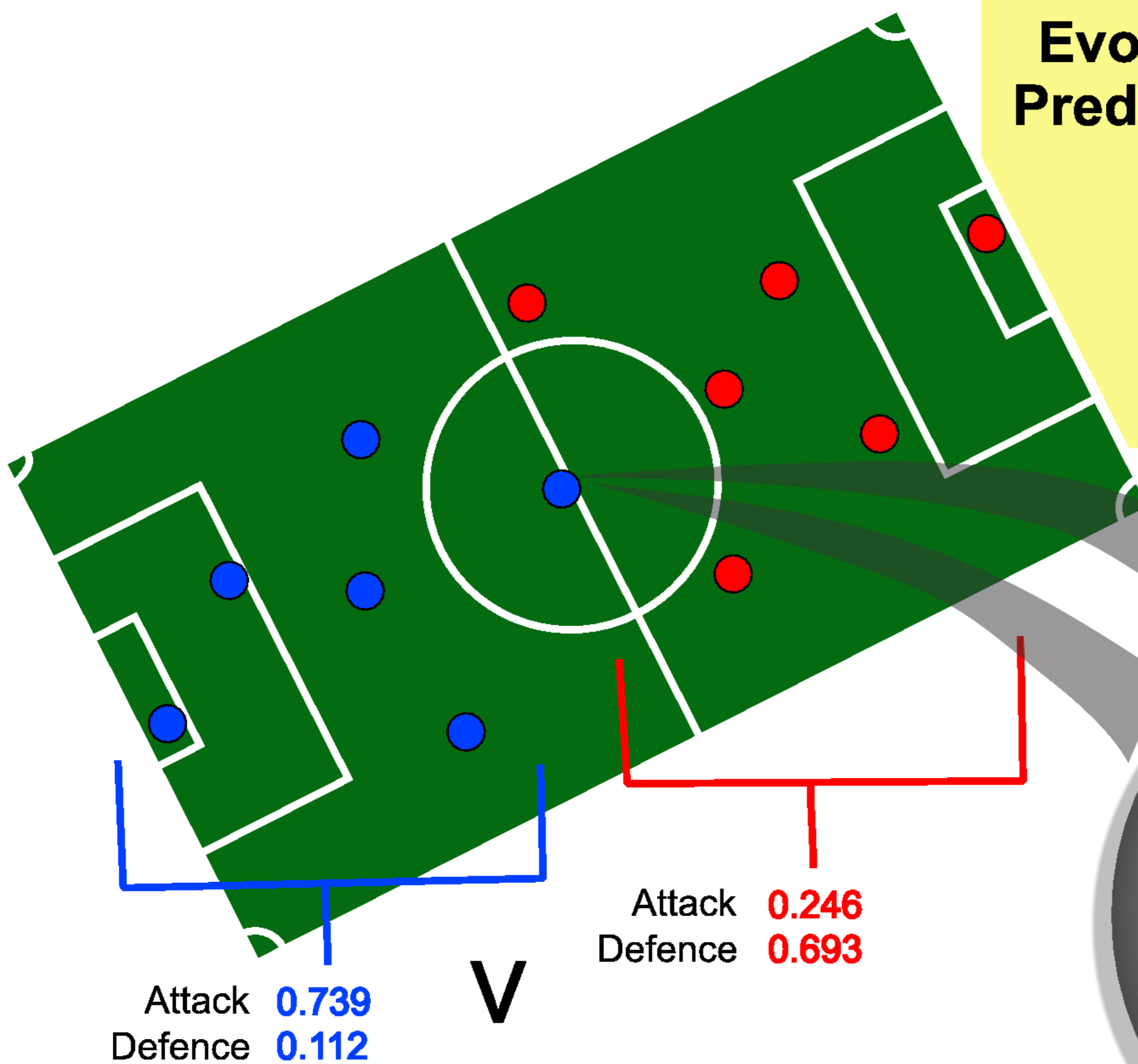


Evolved Representations For Prediction Of Football Matches

Mark Rowan

School of Computer Science,
University of Birmingham

msc99msr@cs.bham.ac.uk,
mark@tamias.co.uk



Showing how it is possible to utilise previous performances of football teams to predict future outcomes.

Introduction

For many centuries, mankind has sought to capitalise on the unpredictability of certain stochastic events; whether this be casting lots on dice, playing roulette, forecasting the stock markets, or betting on the outcome of sporting matches such as football.

Some of these challenges rely more on "luck" than skill, although there is almost always an element of both required to be successful in predicting outcomes of future events. If the element of skill can be reproduced computationally and improved beyond human skill (as has been well documented in many fields, such as the game of chess), then it may prove possible to succeed to some extent in predicting the outcome of football matches given past performance.

Defining a representation

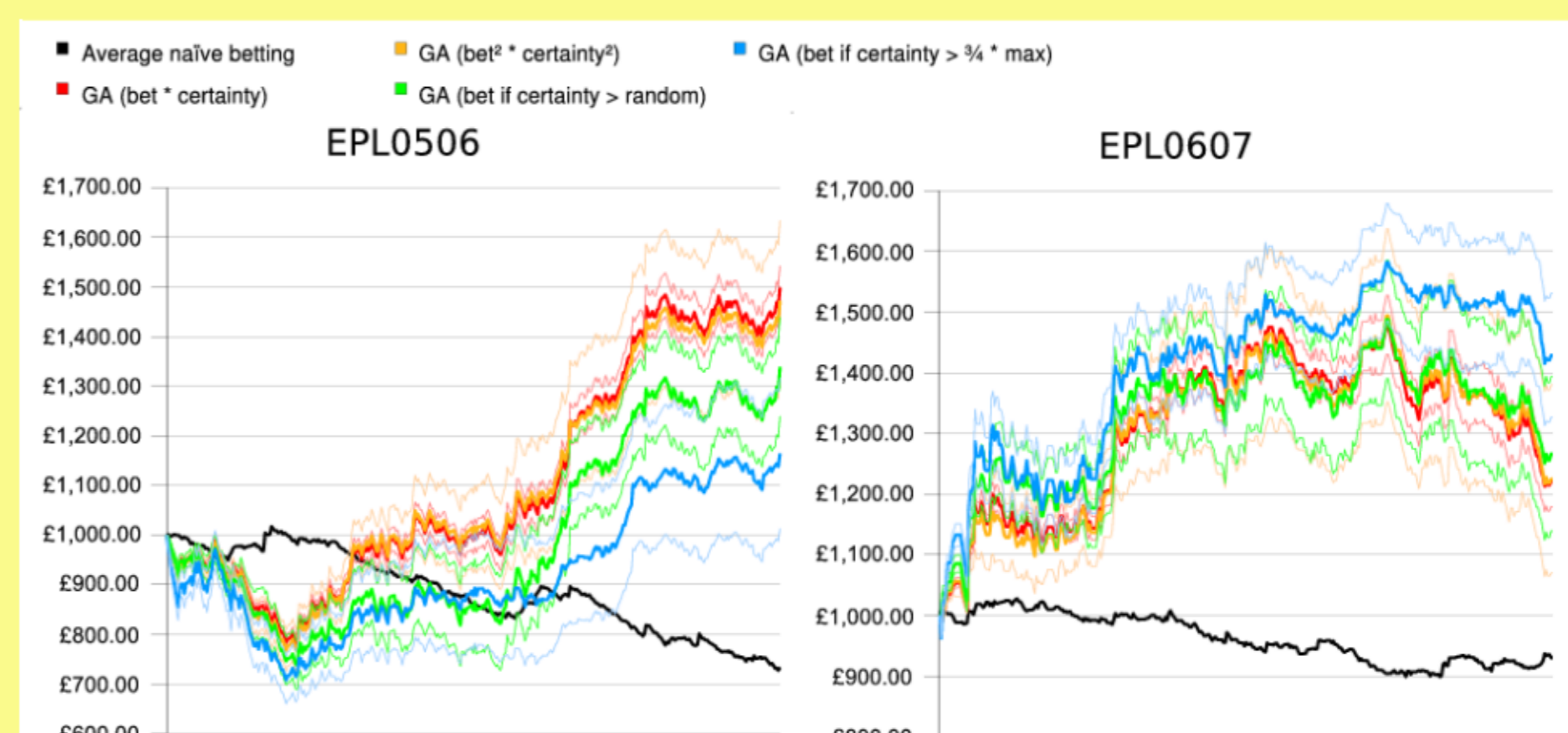
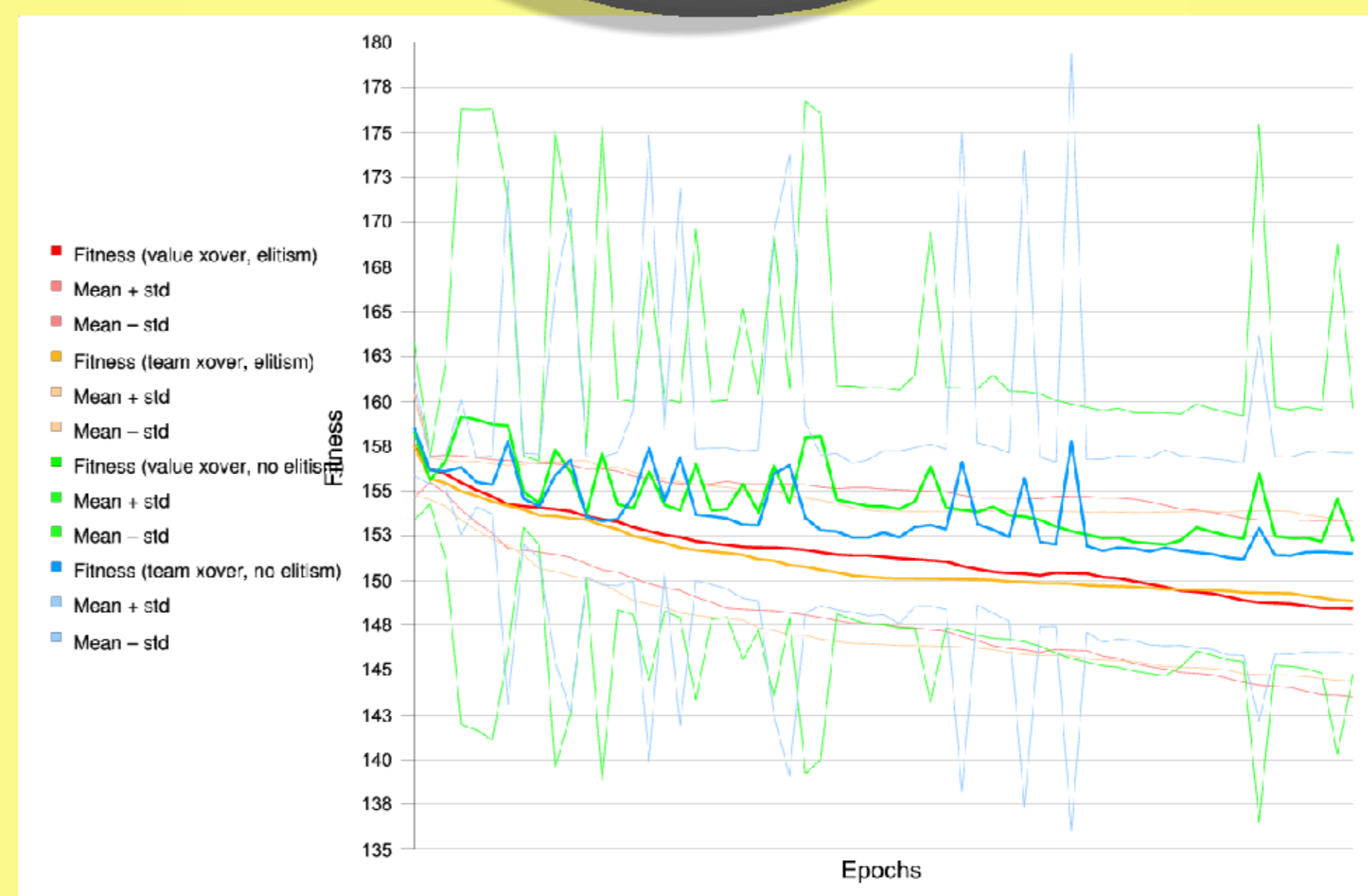
Before an algorithm is able to process historical data and use it to generalise to future performance, a method of representing the performance of two competing football teams must be devised. Traditionally, representations focus on fuzzy logic inference, leading to logical statements such as "if team *a* beat team *b*, and team *b* beat team *c*, then team *a* will beat team *c*" (Tsakonas, *et al.*, 2002). Proposed in this project is a more precise and flexible mathematical representation, based around numerical scores given to each team denoting their relative attack and defence strengths when playing at home and away from home (see main figure above).

Optimising the representation

The chosen method of optimising the scores is by using a genetic algorithm. The scores are stored as binary values using reflective Gray coding, allowing thorough manipulation without unintentionally informing the GA of its upper and lower bounds for the values. The GA also encodes its own mutation rate in the individuals, allowing for self-adaptation, and a ratio of home wins : away wins : draws is self-evolved simultaneously to enable the differences between two teams' scores to be quantified into one of these three outcomes.

Variations on the GA and its parameters were then implemented and tested, including population size, number of epochs of training, use of elitism, precision of the binary representation, different crossover schemas, and initial mutation rate values.

Results



Means and standard deviations for the fitness of the best individual in the population on each epoch of training over twenty runs were recorded and plotted (see top graph above). A clear fitness improvement trend is evident, showing that the GA is able to optimise the values in the representation and improve its performance on the training set. Means and standard deviations for the actual performance of the trained model on a number of testing sets using different training techniques were then also plotted against the results of average naive betting (see bottom graph above), showing the dramatic improvement in betting returns of the trained model over naive betting, which invariably loses money to bookmakers' profits.