I used genetic programming (GP) to transform labels from a multilabel dataset into new labels. This is needed because multilabel classification often works with datasets with a large number of labels.

The primitive functions I used are the arithmetic addition, subtraction, multiplication, and protected division, where x / y is 1 when y = 0. The terminal nodes are the class labels.

In the evaluation function, for each label i I iterated through all other labels j, and in a list I stored the Hamming distance between i and j as well as the new label j, as well as the absolute value between the new label i and label j. From this list, I could find the 5 nearest original labels by sorting by the Hamming distance, as well as the 5 nearest new labels by sorting by the absolute values of the differences between the new labels. Then, for each label in the 5 nearest original labels not in the 5 nearest new labels, I added 1 to the fitness.

The goal of the algorithm was to minimize the fitness function. The parameters I used were tournsize = 7, maximum height of tree = 7, population = 500, number of iterations = 50. I ran the eaSimple algorithm on the ‘emotions’ and ‘medical’ train datasets, and the results are shown below.

Dataset	Min. fitness at generation 0	Min. fitness at generation 50
‘emotions’	28	23
‘medical’	118	49