



Health prediction for king salmon via evolutionary machine learning with genetic programming

Fangfang Zhang, Yuye Zhang, Paula Casanovas, Jessica Schattschneider, Seumas P. Walker, Bing Xue, Mengjie Zhang & Jane E. Symonds

To cite this article: Fangfang Zhang, Yuye Zhang, Paula Casanovas, Jessica Schattschneider, Seumas P. Walker, Bing Xue, Mengjie Zhang & Jane E. Symonds (14 Mar 2024): Health prediction for king salmon via evolutionary machine learning with genetic programming, Journal of the Royal Society of New Zealand, DOI: [10.1080/03036758.2024.2329228](https://doi.org/10.1080/03036758.2024.2329228)

To link to this article: <https://doi.org/10.1080/03036758.2024.2329228>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 14 Mar 2024.



Submit your article to this journal [↗](#)



View related articles [↗](#)




View Crossmark data [↗](#)

RESEARCH ARTICLE



Health prediction for king salmon via evolutionary machine learning with genetic programming

Fangfang Zhang ^a, Yuye Zhang^a, Paula Casanovas^b, Jessica Schattschneider^b,
Seumas P. Walker^b, Bing Xue^a, Mengjie Zhang^a and Jane E. Symonds^b

^aCentre for Data Science and Artificial Intelligence & School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand; ^bCawthron Institute, Nelson, New Zealand

ABSTRACT

King (Chinook) salmon is the only salmon species farmed in Aotearoa New Zealand and accounts for over half of the world's production of king salmon. Determining the health status of king salmon effectively is important for farming. However, it is a challenging task due to the complex biotic and abiotic factors that influence health. Evolutionary machine learning algorithms have shown their superiority in learning models for challenging tasks. However, they have not been investigated for health prediction in king salmon farming. This paper focuses on data processing and machine learning algorithm design to develop king salmon health prediction models in Aotearoa New Zealand. Particularly, this paper proposes a king salmon health prediction method based on genetic programming which is an evolutionary machine learning algorithm. The results show that genetic programming achieves the best overall performance among all examined typical machine learning algorithms for most trials. Further analyses show that genetic programming can automatically detect important features for learning classifiers for king salmon health classification tasks effectively, and can also learn potentially interpretable models. Our results are an important step forward in developing health prediction tools to automatically assess health status of farmed king salmon in Aotearoa New Zealand.

ARTICLE HISTORY

Received 4 December 2023
Accepted 7 March 2024

KEYWORDS

Evolutionary machine learning; genetic programming; king salmon; health prediction; classification

Introduction

King (Chinook) salmon (*Oncorhynchus tshawytscha*) are often heralded as the best salmon species in terms of taste, texture and nutritional quality, and Aotearoa New Zealand is the largest producer of farmed king salmon in the world (NZKS 2020). Health is an important factor for king salmon farming, and the health prediction of king salmon during production is a priority for improving farming sustainability (Stead and Laird 2002; Buschmann and Muñoz 2019). If farmers can predict the health status of king salmon reliably, they can monitor health status during production

CONTACT Fangfang Zhang  fangfang.zhang@ecs.vuw.ac.nz

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

and implement proactive health management strategies more easily. This can improve farming sustainability and resilience, which can bring great benefits for king salmon farming, especially under climate change challenges (Feddern et al. 2023). Various features such as temperature, feeding frequency, husbandry practices and the presence of pathogens can affect the health of king salmon. It is non-trivial for farmers to know which measurable set of variables (features) are important for the prediction of king salmon health.

The Cawthron Institute in Aotearoa New Zealand works closely with the salmon industry and have collected data on king salmon health and growth-related variables, as part of different projects, in its Finfish Research Centre (FRC). Figure 1 shows an example king salmon from the FRC. This paper uses king salmon data available from three trials conducted in the FRC between 2018 and 2020. Each fish sample contains different features and a health label, i.e. healthy or unhealthy, which is naturally a *binary classification task*.

Machine learning (Theobald 2017; Zhou 2021), as a subfield of artificial intelligence (Winston 1984), focuses on model learning to discover patterns and relationships in data, and make predictions based on the learned models. In machine learning, models are learnt from training data, and the learned models are then applied to unseen data to measure their effectiveness and generalisation ability. Genetic programming (GP) as an evolutionary machine learning method (Poli et al. 2008), has been widely used to learn classifiers for classification tasks in health (Espejo et al. 2009) such as breast cancer diagnosis classification (Dhahri et al. 2019; Devarriya et al. 2020), heart disease diagnosis (Reddy et al. 2020) and skin cancer diagnosis (Ain et al. 2022). There are three main advantages of using GP for classification. First, GP can implicitly detect important features during the classifier learning process. Second, the learned classifiers, which are normally with tree-like structures, are easier to be interpreted and understood by human. Last, it is efficient to use learned classifiers obtained by GP to predict the class label, which is very important for real-world applications. However, GP has not been explored for king salmon health predictions.



Figure 1. An example king salmon reared in the Cawthron Institute's Finfish Research Centre.

The goal of this paper is to propose a new GP algorithm to predict king salmon health. The three main contributions of this paper are:

- (1) The king salmon data for all three trials has been effectively preprocessed, and seven unbalanced classification tasks have been generated based on different datasets of information on king salmon health prediction for each trial. Furthermore, imputation has been performed to fill the missing values on the datasets. This lays the foundation for future research on king salmon health with machine learning techniques.
- (2) GP has been investigated for king salmon health classification. For comparison, this paper has implemented a number of commonly used machine learning algorithms such as K-Nearest Neighbour (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF), on the extracted king salmon health classification tasks. The results show that our proposed GP algorithm achieves the best overall performance on the unbalanced classification tasks for most trials.
- (3) The complexities of the classification tasks and the accuracy difference of the different trials are also illustrated with visualisations of the results. In addition, important features have been identified for different king salmon health classification tasks in different trials, which can significantly help the understanding of king salmon health in farming.

The rest of this paper is organised as follows. Section ‘Background and related work’ presents the background and related work for this paper. Section ‘Data processing’ provides the details of data processing and the extracted datasets for three trials. Detailed descriptions of the proposed algorithm are given in Section ‘The proposed GP algorithm’. The experiment designs are shown in Section ‘Experiment designs’. Results and discussions are presented in Section ‘Results and discussions’ followed by further analyses including feature importance for classification tasks in Section ‘Further analyses’. Section ‘Conclusions’ concludes this paper.

Background and related work

Background

Unbalanced classification

Unbalanced classification refers to a classification problem in which one class greatly outweighs the number of instances belonging to the other class or classes in terms of the number of instances (Krawczyk et al. 2016; Kim et al. 2020). Unbalanced classification is found in many real-world applications such as disease diagnosis where there are often fewer diseased samples than normal samples (Ahsan and Siddique 2022; Liu et al. 2023). In unbalanced classification, the majority class known as the negative class, and dominates the dataset, while the minority class typically known as the positive class is underrepresented. The unbalanced classification tasks are more challenging than balanced classification tasks.

The king salmon health classification data in this paper is a typical case of unbalanced classification, where the numbers of healthy fish samples and unhealthy fish samples are

different in all the three trials. *Identifying the unhealthy king salmon accurately* is more important for farmers so that farmers can manage the health issues proactively. Thus, this paper sets *unhealthy king salmon as the positive class* for classification algorithms. Since the fish health status labels are available (i.e. two labels, healthy and unhealthy) for training classifiers with machine learning algorithms, the investigated problem in this paper is a *supervised unbalanced binary classification task*.

K-Nearest neighbours

K-Nearest Neighbours (KNN) is a commonly used classification algorithm, a simple non-parametric algorithm that works based on the principle of similarity (Zhang and Zhou 2007). KNN predicts the label of an unseen instance by finding similar k instances based on the calculated Euclidean distances from training instances, where k is the number of nearest neighbours that are considered when making a prediction. For classification tasks, KNN normally assigns the class label that is most frequent among the k nearest neighbours to an unseen instance.

Naive Bayes

Naive Bayes (NB) is based on Bayes' theorem, which states that the probability of a hypothesis (class) given the observed evidence (features) is proportional to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis (Salmi and Rustam 2019). Bayes' theorem states the relationship which can be expressed as: $P(y|x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n|y) * P(y)}{P(x_1, \dots, x_n)}$, where y is the class variable and x_i indicates a feature value. NB assumes variables are conditionally independent to each other.

Support vector machine

Support Vector Machines (SVMs) aim to find a line, a plane or a hyperplane that linearly separates the data into two different classes by increasing the dimensionality of the data (Cervantes et al. 2020). Particularly, SVM targets on maximising the distance between the decision boundary and the support vectors, i.e. the data points which are the closest to the decision boundary.

Decision tree

Decision Tree (DT) is a tree-like model for decision marking, where each node represents a decision point, and each branch represents a possible decision/output (Banzhaf et al. 1998; Charbuty and Abdulazeez 2021). For constructing a decision tree, it normally starts with the root node, and then add nodes and branches as decisions. DT typically uses statistical measures including entropy and information gain to determine the best splitting criteria for each node in the tree.

Random forest

Random Forest (RF) is an ensemble learning method that uses multiple decision trees to make predictions (Speiser et al. 2019). The key idea of RF is to learn an ensemble of decision trees, where each tree is trained on a different subset of the training data with a random subset of features. RF makes predictions by aggregating the decisions of all trees, e.g. voting for classification.

Genetic programming

Genetic Programming (GP) is one of the most popularly used evolutionary algorithms (Pei et al. 2021; Santoso et al. 2021). GP starts with a randomly initialised population with a number of individuals, and the solutions/individuals are improved generation by generation. At the beginning, all individuals will be evaluated with a fitness function, e.g. classification accuracy. During the evolutionary process, tournament selection is used to select parents and genetic operators such as crossover, mutation and reproduction, are used to generate offspring for forming a new population. When the stopping criterion is met, GP will output the individual with the best fitness as the final solution for the given problem. GP has been successfully used to learn classifiers for unbalanced classification in different studies (Bhowan et al. 2012; Kumar et al. 2020; Pei et al. 2022).

An example of a binary classifier learned by GP for classification task is shown in Figure 2. The learned classifier of GP is a tree-like model, which can be regarded as a function. The expression of the classifier in Figure 2 is $0.1 * F3 + (F1 - F3) / F8$, where $F1$, $F3$ and $F8$ are three features. When predicting the label of an instance, the output of the GP classifier will be calculated. If the output is less than a threshold, e.g. 0 is a commonly used threshold, the instance will be predicted as positive class; Otherwise, it belongs to negative class. This is how this paper uses GP to handle classification tasks.

Related work

In a previous study, statistical analyses were conducted to investigate the relationships between blood biochemistry and haematological indicators with king salmon sampled from freshwater and seawater farms (Casanovas et al. 2021). The results show that there are significant differences between the two environments for some parameters, including haematocrit and haemoglobin. In addition, the results show that some blood parameters are significantly correlated with fish size. The effects of temperature and fasting on king salmon at different life stages were studied in Araújo et al. (2022). The results show that body weight is not significantly impacted by fasting at 13°C. However, fasting at 17°C at all three stages has a negative impact on fillet weight and total fatty acid daily loss. In Lulijwa et al. (2021), the studies show that temperature stress-activated leukocyte apoptosis induces a minor immune response, influencing blood ion profiles indicative of osmoregulatory perturbation, regardless of how well a

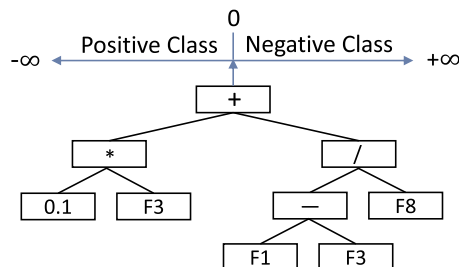


Figure 2. An example of a binary classifier learned by Genetic Programming. The expression of the classifier here is $0.1 * F3 + (F1 - F3) / F8$, where F is the abbreviation of 'Feature', and $F1$, $F3$ and $F8$ are three features.

fish grows. Conversely, fish displaying poor growth performance irrespective of temperature exhibited numerous biomarker shifts including haematology indices and cellular-based enzyme activities. A study (Zhao et al. 2021) shows that water temperature and feed ration play a minor role in affecting the salmon faecal microbial community, but the increased temperature could have affected the fish faecal appearance score. In addition, the faecal microbiomes changes are not associated with corresponding changes in the microbiota of the water and feed. The investigations by Esmaeili et al. (2022) show that growth rate and feed conversion ratio are not significantly different between higher food intake and lower food intake fish.

These studies focussed their investigation on a specific aspect of king salmon health, such as feeding efficiency, rather than a comprehensive study of multiple aspects of king salmon health. They use statistical analyses to compare the experimental groups with the purpose of determining the effect of variables on the investigated targets or specific phenotypes. The relationship between multiple variables, e.g. fish features, and targets, e.g. health, have not been studied.

Data processing

This paper uses king salmon data available from three trials conducted in the FRC between 2018 and 2020. The trials were conducted in freshwater (FW, Trial 1 and Trial 3) or seawater (SW, Trial 2), and at different temperatures. Rations fed to the fish also varied. The experimental trials and the sampling events within each trial are summarised in Table 1. The salmon were raised in trials that tested different control variables such as salinity, feed ration and temperature. The fish were sampled at designated

Table 1. FRC trial information and details for each sampling event indicated by WT*.

Trial	Sampling Event	Salinity	Satiation Ration(s)	Temperature(°C)	Start Date	End Date
1	Arrival	N/A	N/A	N/A	21-Aug-18	21-Aug-18
	WT2	FW	100	15	11-Sep-18	14-Sep-18
	WT4	FW	60, 80, 100	13, 17	15-Oct-18	23-Oct-18
	WT7	FW	60, 80, 100	13, 17	26-Nov-18	06-Dec-18
	WT10	FW	60, 80, 100	17	21-Jan-19	23-Jan-19
	WT14	FW	60, 80, 100	17	12-Mar-19	28-Mar-19
2	Arrival	N/A	N/A	N/A	17-Dec-18	18-Dec-18
	WT2	SW	100	17	31-Jan-19	01-Feb-19
	WT3	SW	100	17	12-Feb-19	13-Feb-19
	WT4	SW	100	17	15-Apr-19	18-Apr-19
	WT5	SW	100	17	10-Jun-19	27-Jun-19
	WT6	SW	100	17	29-Jul-19	12-Aug-19
	WT7	SW	100	17	30-Sep-19	22-Oct-19
	WT9	SW	100	17	18-Nov-19	03-Dec-19
	WT11	SW	100	17, 19	17-Feb-20	27-Feb-20
	Arrival	N/A	N/A	N/A	06-May-20	25-May-20
3	WT2	FW	100	14	08-Jun-20	10-Jun-20
	WT3	FW	100	14	15-Jun-20	17-Jun-20
	WT4	FW	100	8, 12, 16, 20	06-Jul-20	16-Jul-20
	WT5	FW	100	8, 12, 16, 20	05-Aug-20	18-Aug-20
	WT6	FW	25	8, 12, 16, 20	26-Aug-20	08-Sep-20
	WT7	FW	25	8, 12, 16, 20	16-Sep-20	29-Sep-20
	WT8	FW	0	8, 12, 16, 20	14-Oct-20	28-Oct-20

Notes: For Satiation Ratio, 100 = fish fed to satiation, 80, 60, 25 = fish fed to 80%, 60% or 25% of the satiation ration respectively. 0 satiation ration = fish were not fed.

time points throughout the trials to collect health information in relation to the different experimental conditions and to monitor health over time.

Figure 3 shows the overall data collections and the number of features assessed for each dataset. There are nine data collections that indicate different aspects of king salmon information, i.e. blood biochemistry and haematology, body chemistry composition (e.g. fatty acids), feeding (e.g. feed intake), biometrics (e.g. body measurements), growth (weight, fork length, girth), sample assessments (e.g. external and internal appearance, kidney score, organ indices, spinal anomalies, swim bladder and stomach abnormalities), histology (multiple tissues), trial information (e.g. temperature and ratio) and health classification. Different collections contain various numbers of observations for king salmons, i.e. some have more observations than others. In addition, each collection is conducted during three different trials, and the number of observations of each trial under the same collection is slightly different. The health classification datasets highlighted in blue consist of information on whether the king salmon individuals examined are presumed healthy or unhealthy, which takes into account different aspects of observed fish health variables at the time of sampling (see Table 2). This paper considers *seven datasets, blood biochemistry and haematology, body chemistry composition, feeding, growth, sample assessment, histology and biometrics*, with the collection *health classification*, to form seven classification tasks. Collection trial information highlighted in orange contains the environment information such as temperature ($^{\circ}\text{C}$) and the ration the fish were fed (e.g. satiation ration or reduced ration) at the time of sampling. This paper adds water *temperature_celsius* and *satiation_ration* as two features into the investigated datasets by using the fish ID and event as a key. In general, the dataset extractions have four main steps:

- (1) *Expand each dataset by using fish ID and event as the key, and the unique set of observation variables as features.* Note that a fish may have health labels at different events, to get accurate data for health classification tasks, we only consider the last event to construct the classification tasks. In addition, not all the fish are examined by all criteria listed in Table 2. In addition, not all the fish are examined by all criteria listed in Table 2 due to different experiment design considerations for various observation purposes. To have accurate health labels, this paper only uses king salmon samples so that their health is identified according to all criteria listed in Table 2.

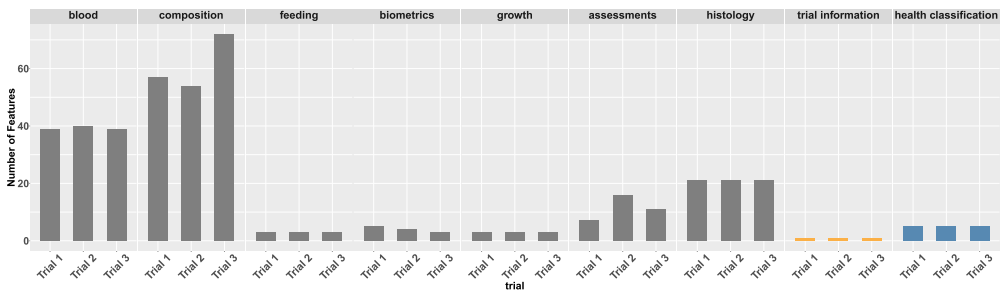


Figure 3. Overall data sets and the number of features assessed for each dataset of king salmon from Cawthron Institute.

Table 2. Health classification criteria.

Dataset	Parameter	Unhealthy classification criteria
Growth	Weight loss Condition factor	Lost weight between weight assessments Condition factor less than 1.1
Haematology	Blood cell appearance: Leucocytes, erythrocytes, thrombocytes Percentage of white blood cells	Presence of abnormalities Percentages: lymphocytes<87%, neutrophils>10%, monocytes>2%
Health assessment	Swim bladder Stomach Stomach width Kidney Faecal appearance Liver index	Presence of fluid in the swim bladder, abnormal if volume: >1 ml (fish < 500 g) or >2 ml (fish>500 g) Abnormal based on visual assessment Abnormal if width: >20 mm (fish<500g) or >35 mm (fish>500 g) Visual nephrocalcinosis score>3 Faecal appearance score>= 3 Liver index<0.75
Histology	Total histology score, sum of scores for all individual tissues Gastrointestinal tract inflammation score Inflammation score	Total score>12 Score>5 Score>10
Comments	Observations during external and internal visual assessments	Health related abnormalities recorded

In addition, not all the fish are examined by all criteria listed in Table 2. To have accurate health labels, this paper only uses data from the salmon that have measured with all criteria listed in Table 2. An observation variable can be examined multiple times on different body parts such as general health in the kidney and stomach. In this case, this paper creates a new feature formatted as observation_bodypart such as generalhealth_kidney and generalhealth_stomach to make every instance unique.

- (2) *Add two new features temperature_celsius and satiation_ration from collection trial information for each dataset.* Specifically, this paper utilises fish ID and event as the key to find the values of two new features, i.e.temperature_celsius and satiation_ration from the collection trial_dates, and combine them with the datasets extracted according to step (1).
- (3) *Add class label for instances.* This paper uses the fish ID as the key to finding the class label of each sample from collection health classification.
- (4) *Replace string feature values with numeric numbers.* Machine learning algorithms typically handle numeric feature values, and this paper needs to convert non-numeric values into numeric values. The main replacements are for event variables, where in trial 1, this paper sets 'tag' to 0, 'WT2' to 1, 'WT4' to 2, 'WT7' to 3, 'WT10' to 4, 'WT14' to 5; in trial 2, there are 'tag' to 6, 'WT2' to 7, 'WT3' to 8, 'WT4' to 9, 'WT5' to 10, 'WT6' to 11, 'WT7' to 12, 'WT9' to 13, 'WT11' to 14; in trial 3, there are 'tag' to 15, 'WT2' to 16, 'WT3' to 17, 'WT4' to 18, 'WT5' to 19, 'WT6' to 20, 'WT7' to 21, 'WT8' to 22, where WT indicates the sampling period. In addition, this paper uses 0 to indicate unhealthy king salmon and 1 to represent healthy king salmon.

Table 3 shows the sizes represented by the number of instances, the number of features, and class unbalanced ratio calculated as the division of the number of unhealthy fish and healthy fish of the extracted datasets for each trial. For example, the size (103,37) of the blood dataset in trial 1 indicates there are 103 fish samples and 37 features on this

Table 3. Sizes of datasets represented by the number of samples and features, and the imbalance ratio of unhealthy class and healthy class for different trials.

No.	Dataset	Trial1		Trial2		Trial3		#Class
		Size	Imbalance Ratio	Size	Imbalance Ratio	Size	Imbalance Ratio	
1	blood	(103,37)	1.51	(445,38)	2.53	(376,36)	0.29	2
2	composition	(103,116)	1.51	(275,115)	3.44	(376,127)	0.29	2
3	feeding	(99,7)	1.48	(126,7)	1.14	(113,7)	0.19	2
4	growth	(103,5)	1.51	(445,5)	2.53	(376,5)	0.29	2
5	assessment	(103,5)	1.51	(445,14)	2.53	(376,18)	0.29	2
6	histology	(103,36)	1.51	(445,36)	2.53	(376,36)	0.29	2
7	biometrics	(103,14)	1.51	(444,15)	2.55	(376,9)	0.29	2

dataset. The number of unhealthy king salmon is larger than the number of healthy fish in trial 1 and trial 2. On the contrary, trial 3 has less unhealthy fish than healthy fish. The dataset presents missing values, and we use KNN (Alianso et al. 2022) with $k=5$ to impute them, where the parameter k is chosen according to preliminary investigation.

Note that the costs or complexities of obtaining the different features across the datasets varied. The feeding, biometrics, growth and assessment data were collected by the Cawthron team during the assessment events. The blood analyses, composition and histology were the most expensive methods and required samples to be submitted to specialised analytical laboratories. The histology scoring was carried out by a trained histopathologist (Casanovas et al. 2021). For blood information, blood samples were collected from the caudal vein immediately following euthanasia, and tested by International Accreditation New Zealand (IANZ), an accredited commercial laboratory (Gribbles Veterinary, Christchurch, New Zealand) for a targeted and quantitative analysis of all biochemistry and haematology analyses. King salmon food intake was measured by X-radiography using feed containing X-ray opaque beads (Esmaeili et al. 2021; Elvy et al. 2022).

The proposed GP algorithm

The flowchart of GP on king salmon health classification

Figure 4 shows the flowchart of GP on king salmon health classification with an example of using a GP individual for king salmon healthy classification. GP starts with population initialisation, and all GP individuals are represented as trees in the evolutionary process. All the individuals are then evaluated to get their fitness during the individual evaluation stage. For individual evaluation, an important step is to predict the classification labels of king salmon instances, and then classification accuracy can be calculated based on the comparison between predicted class labels and the real class labels. For predicting the class label of a king salmon instance with a GP individual $(F1-F4)/F5$, one will calculate the output of a GP individual with the feature values of a king salmon instance. In this example, GP automatically selects important features $F1$, $F4$ and $F5$ to build the classifier. If the output of a GP individual is smaller than or equal to 0, this paper will set the predicted class label as 0 which indicates an unhealthy fish. For example, for the first instance in Figure 4, the output of the classifier $(F1-F4)/F5$ is $(10-20)/2$. i.e. -5 . Since -5 is smaller than 0, the corresponding king salmon of this instance is predicted as an unhealthy king salmon. Otherwise, this algorithm will give a label 1 which represents

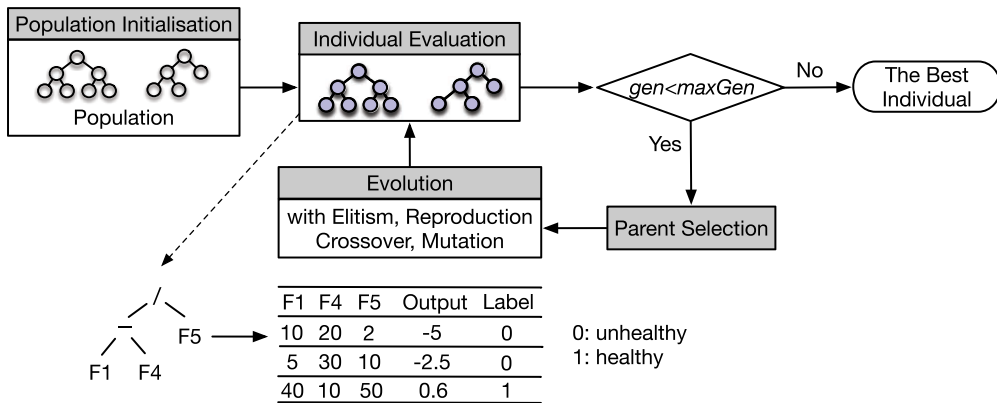


Figure 4. The flowchart of GP on king salmon health classification with an example of using a GP individual for king salmon health classification.

healthy fish. If the stopping criterion is not met, e.g. the maximal number of generations, the individuals with high classification accuracy will be selected to generate a new population during the evolutionary process via genetic operators, i.e. elitism, reproduction, crossover and mutation. If the stopping criterion is met, the best individual learned will be the output of the GP algorithm.

Fitness function

The fitness function is to evaluate the performance of each tree (i.e. a candidate solution) in order to guide the search of GP. Table 4 shows all possible situations of a classification task by the predicted class and actual class, i.e. the confusion matrix. According to whether the predicted label is the same as the actual one, there are four cases which are true positive (TP), false positive (FP), false negative (FN) and true negative (TN). This paper uses a commonly used fitness function for unbalanced classification, i.e. F1 score which is a combination of precision and recall as shown in Equation (1). The calculations of precision and recall are shown in Equations (2) and (3), respectively. The precision indicates the ratio of truly predicted positive instances among all instances that are predicted as the positive class. The recall represents the ratio of truly predicted positive instances among all instances in positive class. A larger F1 score value indicates a better performance.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

Table 4. Possible prediction and actual class for a classification task.

		Actual Class	
		1	0
Predicted Class	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Evolution

GP is an evolutionary machine learning algorithm that mimics the evolution in nature to improve the quality/adaption of its individuals/solutions generation by generation. The output of the GP algorithm is the best classifier/tree at the last generation. Crossover and mutation are two main genetic operators for generating new solutions or offspring during the evolutionary process for GP. For crossover, two parents are selected, and subtrees from them are randomly swapped to generate two offspring. An example can be found in Figure 5. The generated offspring have genetic materials from both parents. For mutation, one parent is selected, and one subtree is randomly chosen and replaced with a newly generated subtree. An example can be found in Figure 6. Except for crossover

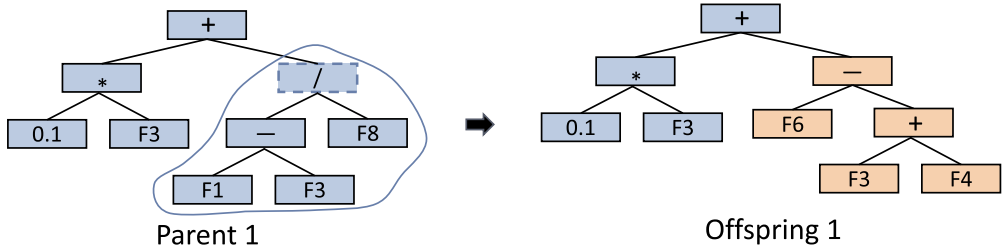


Figure 6. An example of mutation to generate offspring, where F_i indicates feature/variable i .

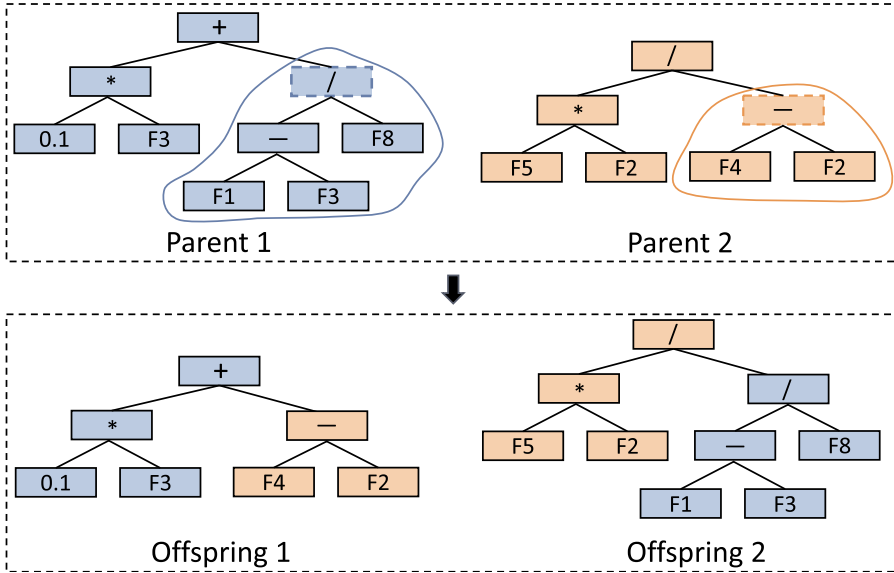


Figure 5. An example of crossover to generate offspring, where F_i indicates feature/variable i .

and mutation, GP also uses reproduction and elitism operators to keep good individuals into the next generation. Reproduction operator keeps the selected parents to the next generation. Elitism keeps the best individual(s) which is(are) chosen according to their fitnesses obtained with the fitness function, from the current generation to the next generation to avoid losing good found individuals.

Experiment designs

Training and test

Each dataset is divided into a training set and a test set. The training data are used to learn a model which is a classifier for a classification task, and the test data are used to measure the performance of the learned classifiers. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample (Fushiki 2011). The procedure has a single parameter called k that refers to the number of groups that a given data set is to be split into. Specifically, this paper uses 5-fold cross-validation to split the data into 5 folds, and use 4 folds as training data to calculate the training accuracy (80% of data for training) and the classification accuracy of the rest fold as the test accuracy (20% of remaining data for test). The test fold is unseen during the training process. This process will repeat five times so that each fold will be used once only, the average of the five test accuracies will be used as the overall test performance. This process will repeat five times so that each fold will be used once only, the average of the five test accuracies will be used as the overall test performance. Figure 7 shows how to use 5-fold cross-validation to obtain the final test classification accuracy as the performance. This paper uses the average test accuracy in each split as the final test accuracy of the learned classifiers.

All the data are standardised before applying machine learning algorithms, and the standard score of a sample x is calculated as:

$$z = \frac{x - u}{s} \quad (4)$$

where u is the mean of the training samples, and s is the standard deviation of the training samples. The same mean and standard deviation obtained from training are then used to normalise test data.

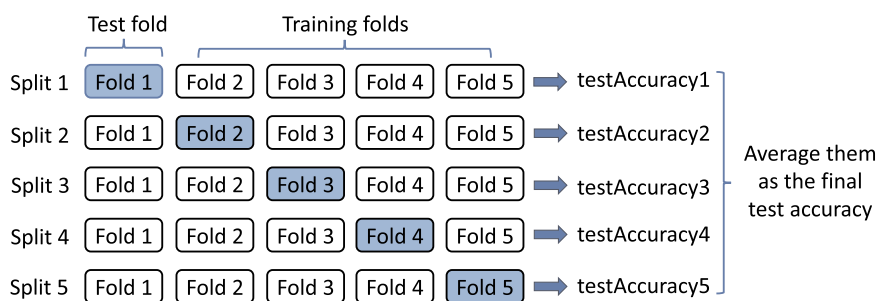


Figure 7. Fivefold cross-validation for getting the final test classification accuracy.

Parameter settings

GP individuals consist of terminals and functions. The terminal set contains all features for each dataset, and this paper uses arithmetic functions, i.e. +, −, *, protected /, as the function set. For the protected /, when the denominator is 0, it returns 1. Other main parameter settings of GP including population size and number of generations, are shown in Table 5, and are typically used for GP (Poli et al. 2008). Crossover, mutation and reproduction rates are the probabilities of applying different genetic operators, which are 0.6, 0.3 and 0.1, respectively. The best individual at each generation, i.e. the number of elite equals 1, will be kept to the next generation as elite. Ramped half and half method is used to initialise the population with individuals that have a depth between 2 and 4. During the evolution process, the depths of individuals should not exceed 6. This paper uses tournament selection with a size of 7 to select parents for generating new offspring.

For KNN algorithm, the *k* is set as 5, which means KNN depends on the closest 5 neighbours to predict the label of an instance. SVM uses radial basis function kernel. DT and RF set the maximum depth of tree to 3, the minimum number of samples required to split an internal node to 10, and the minimum number of samples required to be at a leaf node to 3. In addition, RF sets the number of trees in the forest to 80, and criterion to ‘entropy’. These settings have shown good performance based on our preliminary investigations.

Performance metrics

Considering that detecting unhealthy king salmon as unhealthy fish (rather than identifying healthy king salmon as healthy fish) is more valuable for farming, this paper sets the unhealthy class as the positive class. F1 score as shown in Equation (1) in Section ‘Fitness function’ is provided as an overall performance information. In addition, with a focus on detecting the true unhealthy king salmon samples accurately, this paper uses recall shown in Equation (3) in Section ‘Fitness function’, as metrics for measuring the performance of the involved machine learning algorithms.

Results and discussions

The average accuracy of 5-fold crossover-validation is reported as the results for each run, and this paper runs each algorithm for 30 independent runs. The performance of the algorithms based on the 30 independent runs is ranked using Friedman’s test with a significance level of 0.05. The ‘Rank’ represents the algorithm’s average ranking on all examined datasets. A smaller rank value indicates better performance. The best

Table 5. Parameter settings of GP.

Parameter	Value	Parameter	Value
Population size	1000	Number of generations	50
Crossover rate	0.6	Mutation rate	0.3
Reproduction rate	0.1	Number of elites	1
Initial minimal/maximal tree depth	2/4	Maximal tree depth	6
Initialisation method	Ramped half and half	Tournament size	7

classification accuracy for detecting the unhealthy fish (i.e. recall) achieved on each dataset is highlighted in bold.

Trial 1

Test performance

Table 6 shows the mean and standard deviation of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 1 according to the 30 independent runs on the seven datasets. For trial 1, GP achieves the best performance with a rank of 2.2 among all compared algorithms followed by SVM with a rank of 3.15. Further looking at the test recall in trial 1 of all involved algorithms as shown in Table 7, the results show that GP has the best overall accuracy with a rank of 1.65 among all involved algorithms, and achieves the best accuracy on five out of the seven datasets.

Figure 8 shows the violin plots of test recall of KNN, NB, DT, RF, SVM and GP according to the 30 independent runs in trial 1. In general, GP is superior to the other algorithms in determining the king salmon health classification in trial 1 for 5 out of the 7 datasets, i.e. <blood>, <feeding>, <growth>, <assessment> and <histology> with higher test recall value distributions. Although GP is not the best one on datasets <composition> and <biometrics>, GP performs the second best.

Training vs test

This section investigates the generalisation ability of involved algorithms by looking at their *training recall* and *test recall*. Figure 9 shows the scatter plots of the recall on training and test in trial 1. Most algorithms have a good generalisation ability that their

Table 6. The mean (standard deviation) of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 1 over the 30 independent runs on the seven datasets.

Dataset	KNN	NB	DT	RF	SVM	GP
blood	64.20(3.10)	60.97(2.85)	52.78(5.32)	58.64(3.15)	59.92(2.81)	66.03(3.69)
composition	61.00(2.86)	72.33(3.24)	57.19(4.02)	62.95(1.98)	63.47(2.95)	67.37(5.15)
feeding	64.67(3.43)	64.58(3.38)	58.81(5.37)	67.45(1.41)	70.57(1.97)	69.22(2.99)
growth	60.97(2.83)	69.51(1.68)	65.63(4.65)	56.58(2.94)	55.41(3.08)	67.47(3.81)
assessment	62.87(3.02)	32.84(3.08)	49.61(7.79)	60.93(4.72)	64.31(3.14)	69.11(3.59)
histology	64.40(2.94)	33.52(4.18)	59.37(5.07)	66.57(2.49)	63.86(3.40)	64.48(3.36)
biometrics	68.62(3.39)	61.99(2.99)	58.30(5.73)	66.11(1.64)	67.29(2.75)	66.65(3.44)
Rank	3.22	3.9	4.92	3.6	3.15	2.2

Table 7. The mean (standard deviation) of recall of KNN, NB, DT, RF, SVM and GP in trial 1 over the 30 independent runs on the seven datasets.

Dataset	KNN	NB	DT	RF	SVM	GP
blood	66.07(4.51)	55.21(3.58)	48.32(6.55)	55.42(3.75)	56.03(3.45)	74.68(6.92)
composition	65.02(3.58)	84.24(5.43)	53.92(5.35)	59.49(2.46)	60.12(3.25)	74.06(8.11)
feeding	65.10(4.55)	62.27(4.22)	58.28(6.97)	65.07(1.98)	71.76(2.52)	75.93(4.98)
growth	64.80(3.74)	77.26(2.45)	68.15(8.01)	55.25(4.22)	50.67(4.68)	78.71(6.23)
assessment	64.93(4.02)	27.74(2.91)	45.19(9.73)	59.44(5.87)	65.82(4.72)	77.96(5.27)
histology	69.88(4.13)	24.58(5.91)	56.45(7.18)	68.25(3.63)	64.24(4.61)	71.59(5.71)
biometrics	73.41(4.43)	58.79(3.33)	53.12(7.72)	62.59(2.30)	66.37(3.29)	70.92(5.94)
Rank	2.69	4.08	4.95	4.25	4.06	1.65

Bold numbers indicate the highest recall for a given dataset.

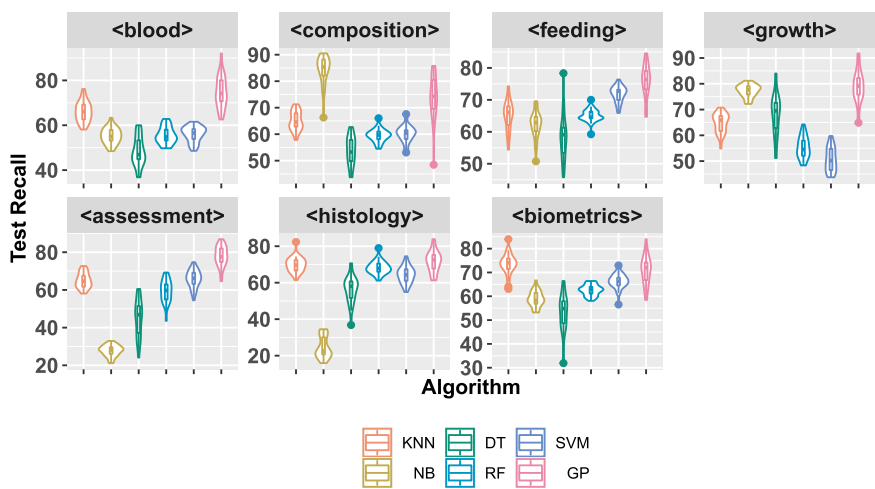


Figure 8. Violin plots of test recall of KNN, NB, DT, RF, SVM and GP in trial 1.

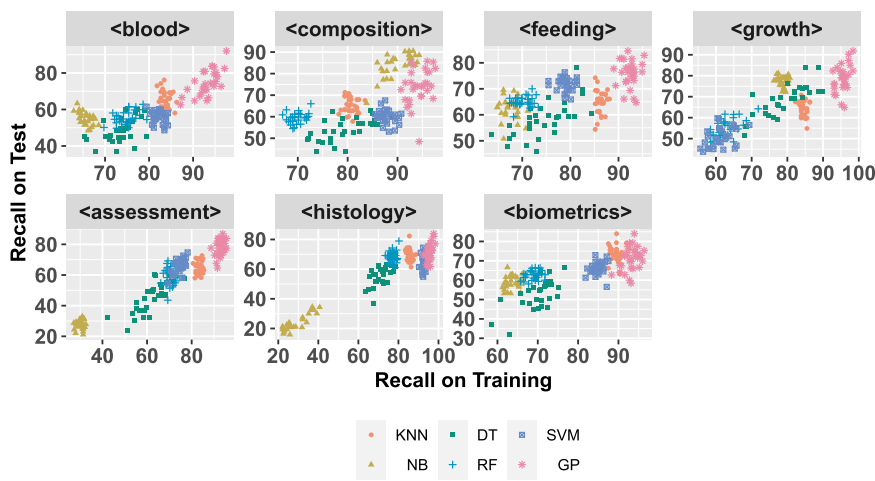


Figure 9. Scatter plots of the recall on training and test in trial 1.

training and test recall values have high correlations. However, some algorithms have poor generalisation ability, i.e. a high training accuracy but a low test accuracy, such as DT on datasets <blood>, <composition>, <feeding> and <biometrics>. This might also be the reason why DT performs the worst in trial 1. Overall, GP has a good generalisation ability, and performs the best on training and test on most of the datasets.

Trial 2

Test performance

Table 8 shows the mean and standard deviation of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 2 according to the 30 independent runs on the seven datasets.

Table 8. The mean (standard deviation) of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 2 over the 30 independent runs on the seven datasets.

Dataset	KNN	NB	DT	RF	SVM	GP
blood	78.77(0.99)	56.18(0.96)	73.95(2.33)	69.75(0.75)	77.18(1.08)	83.48(0.15)
composition	82.35(1.12)	64.14(3.84)	63.72(3.60)	69.13(2.11)	74.49(1.26)	86.68(0.63)
feeding	67.50(3.32)	58.38(2.21)	54.48(4.85)	56.07(2.81)	54.01(2.86)	66.05(3.18)
growth	78.29(0.92)	65.95(1.25)	61.49(3.23)	62.91(2.44)	55.87(0.87)	82.91(0.57)
assessment	76.46(1.07)	76.14(5.02)	55.65(2.56)	61.42(1.46)	59.30(1.02)	83.09(0.53)
histology	74.63(1.37)	39.24(1.59)	56.44(3.63)	62.31(0.69)	68.70(0.98)	82.71(0.49)
biometrics	75.40(0.95)	66.18(1.22)	63.27(4.68)	66.48(1.47)	68.09(1.11)	83.51(0.25)
Rank	2.01	4.45	5.03	4.36	4.06	1.09

Table 9. The mean (standard deviation) of recall of KNN, NB, DT, RF, SVM and GP in trial 2 over the 30 independent runs on the seven datasets.

Dataset	KNN*	NB	DT	RF	SVM	GP
blood	80.80(1.61)	39.98(0.84)	64.70(3.32)	56.98(1.05)	70.89(1.41)	99.84(0.53)
composition	85.46(1.93)	55.98(4.70)	52.67(4.71)	59.82(2.83)	67.93(1.72)	98.44(1.35)
feeding	68.74(3.83)	58.03(2.49)	54.84(6.32)	48.65(2.77)	47.21(3.13)	77.71(5.65)
growth	81.17(1.28)	56.66(1.48)	50.61(4.62)	51.08(2.90)	40.54(1.06)	98.11(1.40)
assessment	80.25(1.80)	84.85(7.05)	42.55(3.23)	49.05(1.74)	46.73(1.15)	97.58(1.05))
histology	73.23(1.90)	25.16(1.19)	44.11(4.80)	48.88(0.89)	59.16(1.08)	97.96(1.03)
biometrics	77.10(1.49)	56.08(1.40)	53.34(6.42)	55.59(1.91)	57.39(1.27)	99.70(0.62)
Rank	2.1	4.3	4.82	4.59	4.19	1

Bold numbers indicate the highest recall for a given dataset.

The results show that GP achieves the best classification accuracy with a rank of 1.09 in trial 2 followed by KNN with a rank of 2.01. Table 9 shows the mean and standard deviation of recall of KNN, NB, DT, RF, SVM and GP in trial 2 according to the 30 independent runs on the seven datasets. The results show that GP performs the best among all algorithms with a rank of 1, followed by KNN with a rank of 2.1. In addition, GP performs the best on all datasets.

Figure 10 shows the violin plots of test recall of KNN, NB, DT, RF, SVM and GP according to the 30 independent runs in trial 2. Although other algorithms may have different performance on different datasets, GP is consistently significantly better than all other compared algorithms on all datasets.

Training vs test

Figure 11 shows the scatter plots of the recall on training and test in trial 2. Overall, all algorithms perform consistently between training and test on all datasets, and clearly GP performs the best (in the very right top position) on all datasets than other algorithms.

Trial 3

Test performance

Table 10 shows the mean and standard deviation of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 3 according to the 30 independent runs. Overall, the results show that SVM is the best among all involved algorithms with a rank of 1.59 followed by GP with a rank of 3.07, which is different from our observations in trial 1 and trial 2. The reason might be that the current GP algorithm cannot cope with highly unbalanced

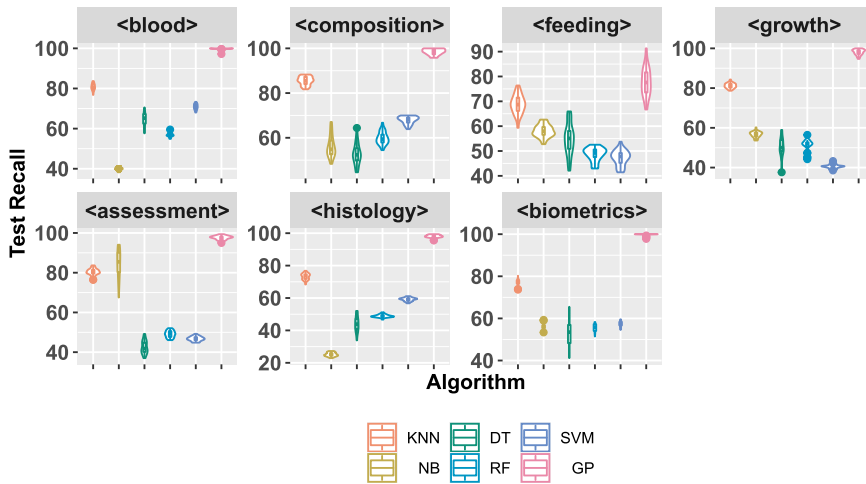


Figure 10. Violin plots of test recall of KNN, NB, DT, RF, SVM, and GP in trial 2.

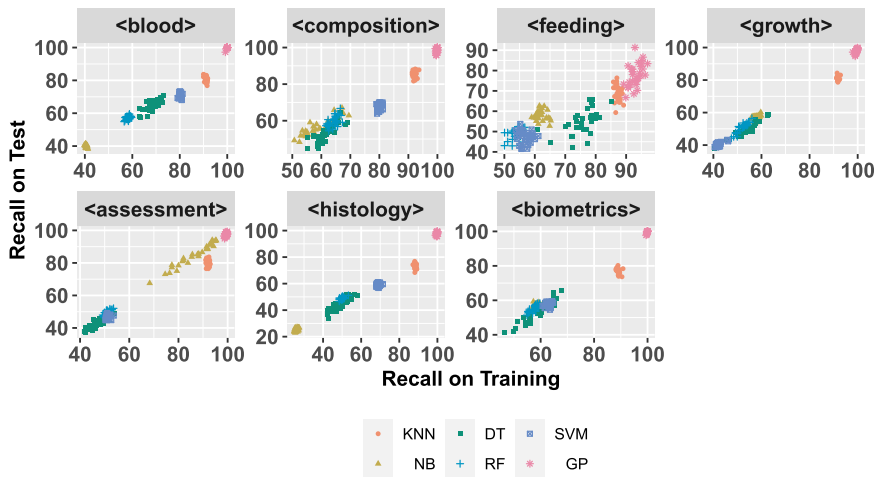


Figure 11. Scatter plots of the recall on training and test in trial 2.

classification tasks well, i.e. trial 3 has a high imbalance ratio. Table 11 shows the mean and standard deviation of recall of KNN, NB, DT, RF, SVM and GP in trial 3 over the 30 independent runs on the seven datasets. It shows that SVM is the best algorithm among all algorithms with the best rank of 1.68. GP with a rank of 2.76 has similar performance with DT which is ranked as 2.72. GP performs the best on the dataset <blood>, and DT performs the best on the dataset <composition>.

Figure 12 shows the violin plots of test recall of KNN, NB, DT, RF, SVM and GP on the seven datasets. It shows that the obtained 30 test objective values of SVM indicate a better performance than all other algorithms on 4 out of 7 datasets, i.e. <growth>, <assessment>, <histology> and <biometrics>.

Table 10. The mean (standard deviation) of test F1 score of KNN, NB, DT, RF, SVM and GP in trial 3 over the 30 independent runs on the seven datasets.

Dataset	KNN	NB	DT	RF	SVM	GP
blood	51.51(2.62)	44.06(2.08)	40.83(3.56)	45.20(1.90)	49.51(1.80)	46.10(2.87)
composition	39.74(3.65)	40.32(3.36)	43.19(3.46)	40.44(3.99)	46.44(2.40)	41.25(4.14)
feeding	4.58(3.80)	14.97(6.99)	26.15(3.90)	34.31(3.69)	28.02(4.32)	23.83(6.16)
growth	35.20(2.11)	23.98(2.64)	42.44(2.78)	37.88(3.57)	47.39(1.93)	44.81(2.91)
assessment	58.36(3.34)	25.82(2.21)	54.70(3.39)	56.00(2.46)	60.32(2.03)	58.61(2.58)
histology	36.09(4.31)	43.35(3.89)	50.19(2.92)	55.67(2.49)	62.22(1.60)	56.41(2.90)
biometrics	43.49(2.36)	37.69(3.52)	48.33(2.36)	40.87(3.96)	52.28(2.13)	45.63(3.58)
Rank	4.12	5.16	3.61	3.45	1.59	3.07

Table 11. The mean (standard deviation) of recall of KNN, NB, DT, RF, SVM and GP in trial 3 over the 30 independent runs on the seven datasets.

Dataset	KNN	NB	DT	RF	SVM	GP
blood	46.00(2.36)	39.76(2.13)	51.02(6.11)	48.00(2.30)	51.92(2.33)	56.04(4.55)
composition	34.59(3.50)	48.63(3.26)	56.51(7.74)	34.47(3.06)	53.65(3.29)	49.29(6.20)
feeding	4.28(3.30)	13.72(6.41)	48.67(7.45)	58.61(4.97)	46.67(6.89)	34.56(9.11)
growth	29.49(1.80)	17.14(2.18)	62.16(8.16)	43.96(5.75)	67.73(3.63)	56.16(5.50)
assessment	48.82(3.07)	15.80(1.44)	45.69(2.49)	51.65(3.16)	68.82(2.69)	52.20(2.34)
histology	25.73(3.63)	41.76(4.15)	59.61(5.33)	61.77(2.47)	66.67(1.73)	60.27(4.67)
biometrics	35.22(2.43)	32.08(3.31)	61.49(6.36)	38.78(4.07)	66.82(2.72)	54.71(5.63)
Rank	5.06	5.29	2.72	3.47	1.68	2.76

Bold numbers indicate the highest recall for a given dataset.

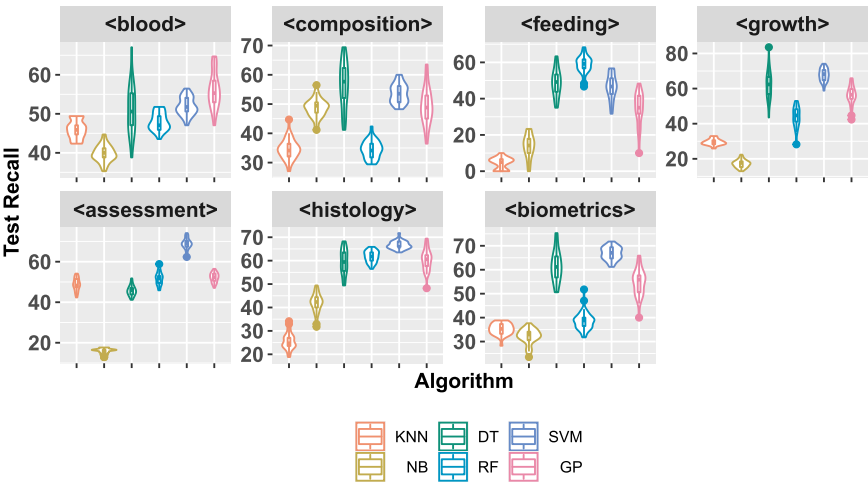


Figure 12. Violin plots of test recall of KNN, NB, DT, RF, SVM and GP in trial 3.

Training vs test

Figure 13 shows the scatter plots of the recall on training and test in trial 3. Although SVM has the best overall performance, SVM does not generalise well on dataset <blood> and <composition>. In addition, KNN does not have a good generalisation ability between training and test, where KNN performs slightly better on training than on test on datasets <blood>, <composition>, <feeding>, <growth>, <histology> and

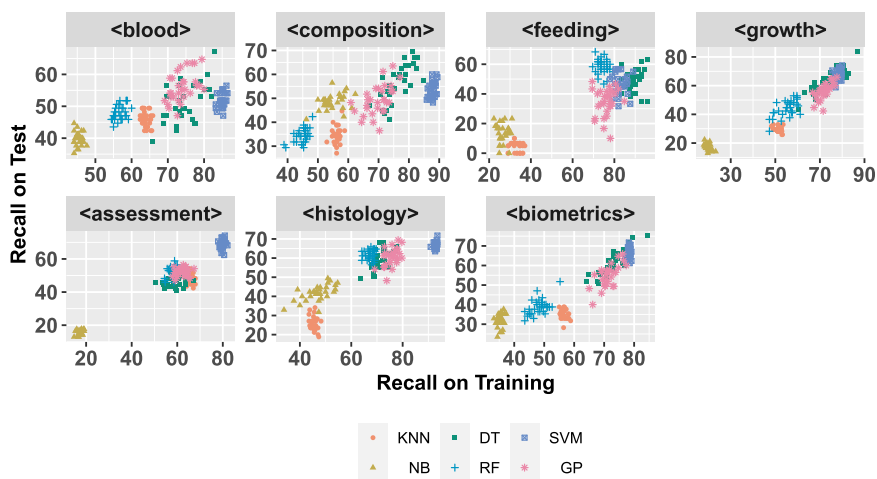


Figure 13. Scatter plots of the recall on training and test in trial 3.

(biometrics). Although GP does not perform the best, i.e. the second best, GP has a good generalisation ability on all datasets.

Discussions on performance across different trials

In general, GP performs the best in most trials. This shows the effectiveness of using GP to learn classifiers for king salmon health classification tasks, and our proposed GP algorithm has a promising ability to detect unhealthy king salmon. Different algorithms might have different performances in different trials, and the obtained classification accuracy in trial 2 is the best followed by trial 1. The obtained classification accuracy in trial 3 is the lowest among all trials. This section takes the dataset <blood> across different trials to investigate the task difficulties in different trials.

The t-distributed stochastic neighbour embedding (t-SNE) is a statistical method for visualising high-dimensional data by giving each data point a location in a two or three-dimensional map. This section uses t-SNE to visualise the dataset <blood> in the three different trials (Figure 14). The number of unhealthy instances is larger than healthy ones in trial 1 and 2, while the number of unhealthy instances is smaller than healthy

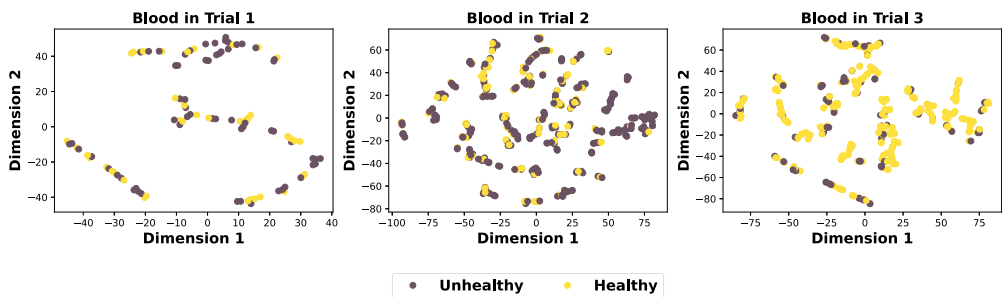


Figure 14. t-SNE visualisation of blood dataset for the three trials.

ones in trial 3. Compared with trial 1 and trial 2, trial 3 has a more complex mixture of healthy and unhealthy instances. This is a possible reason that the classifier taking the unhealthy class as the positive class, predicting unhealthy fish as unhealthy fish, in trial 3 is smaller than in trial 1 and trial 2. Compared with trial 1 and trial 3, trial 2 has a larger number of unhealthy instances, which makes it easier than trial 1 and trial 3 to achieve higher classification accuracy that takes unhealthy class as the positive class. This is consistent with other findings of the experiment comparisons in this Section.

Further analyses

Performance of GP on different folds

This paper uses 5-fold cross-validation to learn classifiers for king salmon health classification tasks, and this section investigates the variances of fitness obtained by GP in different folds. Figure 15 shows the curves of training fitness, i.e. F1 score, along with generations of GP for different folds in trial 1. The fitnesses of GP along with generations on all datasets are similar. This indicates that there are no large variations among the GP performance within different folds. This shows the stability of the performance of GP. The same pattern is observed on all datasets in all three trials. Due to the page limit, this paper does not show the curves of the training fitness in trial 2 and trial 3.

Feature importance

This section investigates which features are important for particular datasets by using the feature importance score. GP is chosen for this investigation since it has overall good performance for king salmon health classification. For GP, the feature importance score is defined as the frequency of features appeared in the best learned classifiers (Qi et al. 2019). This section chooses the dataset <blood> in trial 1 with 37 features to show the feature importance. Table 12 shows the details of features on the dataset <blood> in trial 1.

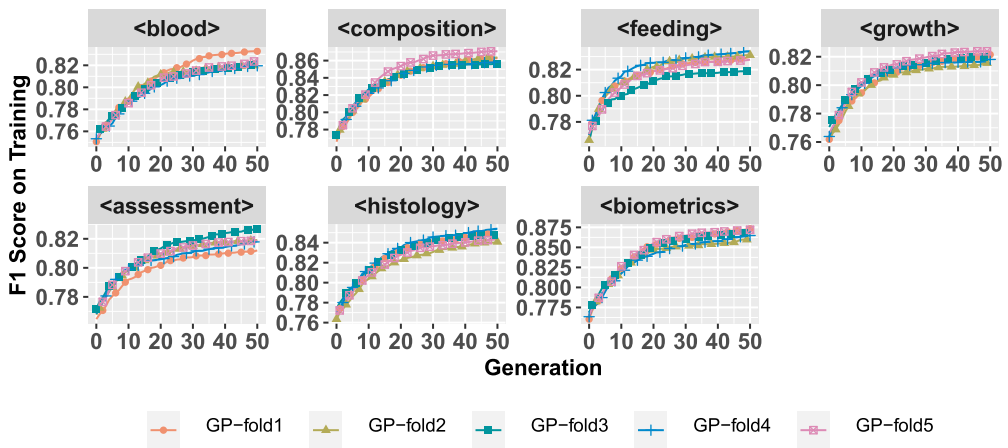


Figure 15. Curves of fitness, i.e. F1 score, along with generations of GP for different folds in trial 1.

Table 12. Detailed features on the blood dataset in trial 1.

ID	Name	ID	Name	ID	Name	ID	Name
1	alanine_aminotransferase	2	albumin	3	alkaline_phosphatase	4	aspartate_aminotransferase
5	bilirubin	6	buffy_coat_thickness	7	c-reactive_protein	8	calcium
9	chloride	10	cholesterol	11	cortisol	12	creatinine_phosphokinase
13	creatinine	14	globulin	15	glucose	16	glutamate_dehydrogenase
17	haematocrit	18	haemoglobin	19	haptoglobin	20	lymphocytes_abs
21	magnesium	22	monocytes_abs	23	neutrophils_abs	24	phosphate
25	potassium	26	prostaglandin_e2	27	sodium	28	thick_buffy_coat
29	total_protein	30	triglycerides	31	urea	32	white_blood_cell_count
33	temperature_celsius	34	satiation_ratio	35	ratioAlbuminGlobulin	36	ratioNeutrophilsLymphocytes
37	meanCorHaeCon						

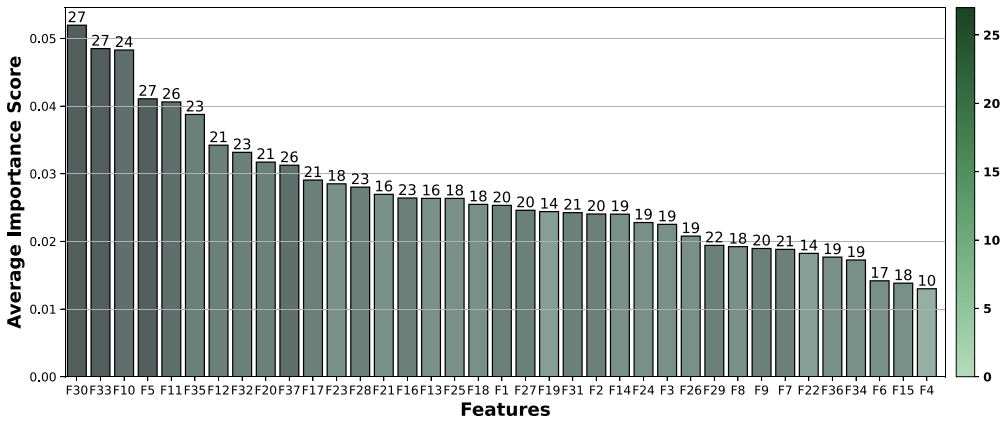


Figure 16. Feature importance on the blood dataset of GP in trial 1 where F_i indicates the i th feature.

Figure 16 shows the average feature importance score on the dataset (blood) according to the 30 independent runs obtained by GP. GP considers F30 (triglycerides), F33 (temperature_celsius), F10 (cholesterol), F5 (bilirubin) and F11 (cortisol) as the top five important features in trial 1. In addition, we can see that important features do not necessary to appear for all best learned classifier, which indicates that the effectiveness of classifiers learned by GP is not only dependent on the important features, but also related to feature construction, i.e. how features are used with functions. This shows that GP can detect important features automatically to generate classifiers for king salmon healthy classification tasks.

Figure 17 shows the boxplots of feature values of the three top important features in Figure 16, i.e. triglycerides (F30), cholesterol (F10) and cortisol (F11) grouped by healthy and unhealthy fish of blood dataset in trial 1. It is clear that the detected important feature values have different distributions in the unhealthy and healthy fish groups according to triglycerides, cholesterol and cortisol. Specifically, unhealthy king salmon have lower triglycerides, cholesterol and cortisol than healthy fish. This reflects the effectiveness of GP to detect important features to learn classifiers. Note that Figure 17 does not include temperature_celsius (F33) and bilirubin (F5), since these two features have the same values between unhealthy and healthy king salmon on blood dataset in trial 1. We can see that GP may choose features with constant values as important features to build classifiers, however, features with constant values is not helpful to distinguish classes. We will further investigate this to improve the effectiveness of GP in our future work.

Learned classifier by GP

To investigate the learned classifier, Figure 18 shows one of the best classifiers learned by GP on the blood dataset in trial 1. Regarding the top five features for the blood dataset in trial 1 as shown in Figure 16, F30 appears three times, and F10 and F11 appear twice in the classifier. This shows the importance of these three features, which is consistent with the observations in Figure 16. The classifier shown in Figure 18 can be simplified as Equation (5). This classifier can be used to make a king salmon health prediction in

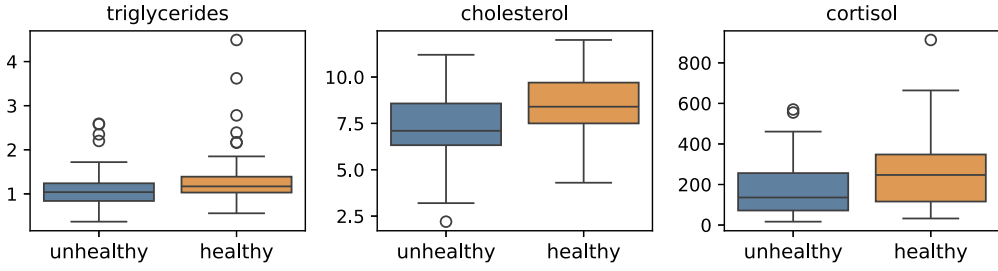


Figure 17. Boxplots of feature values of the top important features, i.e. triglycerides (F30), cholesterol (F10), and cortisol (F11) grouped by healthy and unhealthy fish of the blood dataset in trial 1.

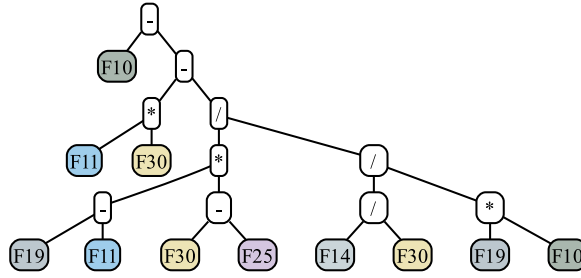


Figure 18. One of the best learned classifiers for the blood dataset in trial 1.

less than a second, which is important for large scale real-world applications.

$$\text{classifier} = F10 - F11 * F30 - \frac{F10 * F19 * F30 * (F19 - F11) * (F30 - F25)}{F14} \quad (5)$$

Conclusions

The goal of this paper was to investigate if GP would be a good approach to predict king salmon health, i.e. determining if a fish is healthy or unhealthy. The goal has been successfully achieved by a designed GP algorithm and its comparison with other machine learning techniques.

The results show that GP is a promising algorithm to learn classifiers for predicting the king salmon health tasks. GP achieves the best overall performance in most trials. This paper observes that different trial data have various difficulties to handle, and the achieved classification performance varies. Specifically, high classification accuracy is achieved in trial 2 followed by trial 1, and the achieved accuracy in trial 3 is the lowest. A further investigation shows that this is caused by different characteristics of the datasets such as the instances distributions. Regarding the robustness of GP for different folds on each dataset, GP has stable training performance across folds with low variance. In addition, this paper investigated the feature importance for fish health prediction, which can provide a future guidance for farming. The learned classifiers learned by GP also show that GP can successfully learn effective classifiers for king salmon health classification tasks automatically. This study builds the foundation of using GP for king salmon health prediction in terms of dataset building and GP

algorithm design for king salmon health prediction. The provided feature importance information is an important step forward in designing effective tools for king salmon farms to improve farming effectiveness.

Some interesting directions can be further studied in future. We plan to design effective feature selection algorithms to only use important features for king salmon health classification. We will start working on how the health information of different organs such as heart and liver can affect the overall health status of king salmon. In addition, we plan to design effective sampling methods for classification tasks with high imbalance ratios.

Acknowledgments

We wish to thank the staff of the Cawthron Institute's Finfish Research Centre (Gareth Nicholson, Chaya Bandaranayake, Michael Scott, Chris Chamberlain, Chris Ensor, Nick Hearn, Liam van den Heuvel and Jordan Elvy) for operating the facility and for sampling and assessing fish during the three trials. We also want to thank industry partners who supplied the salmon for the trials and express our gratitude to staff in the Food Testing Laboratory at Cawthron Institute for technical assistance and sample analysis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by the Science for Technological Innovation Challenge 564 (SfTI) fund under contract 2019-S7-CRS and MBIE Data Science SSIF Fund under 565 the contract RTVU1914. The Cawthron FRC trials were funded by MBIE, contract CAWX1606.

ORCID

Fangfang Zhang  <http://orcid.org/0000-0001-5516-3972>

References

- Ahsan MM, Siddique Z. 2022. Machine learning-based heart disease diagnosis: a systematic literature review. *Artificial Intelligence in Medicine*. 128:102289. doi: [10.1016/j.artmed.2022.102289](https://doi.org/10.1016/j.artmed.2022.102289).
- Ain QU, Al-Sahaf H, Xue B, Zhang M. 2022. Automatically diagnosing skin cancers from multi-modality images using two-stage genetic programming. *IEEE Transactions on Cybernetics*. 53 (5):2727–2740. doi: [10.1109/TCYB.2022.3182474](https://doi.org/10.1109/TCYB.2022.3182474).
- Alianso AS, Syafaah L, Faruq A. 2022. K-nearest neighbor imputation for missing value in hepatitis data. In: *AIP Conference Proceedings*. Vol. 2453. Malang: AIP Publishing.
- Araújo BC, Symonds JE, Walker SP, Miller MR. 2022. Effects of fasting and temperature on the biological parameters, proximal composition, and fatty acid profile of Chinook salmon (*Oncorhynchus tshawytscha*) at different life stages. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*. 264:111113. doi: [10.1016/j.cbpa.2021.111113](https://doi.org/10.1016/j.cbpa.2021.111113).
- Banzhaf W, Nordin P, Keller RE, Francone FD. 1998. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. San Francisco, CA: Morgan Kaufmann.

- Bhowan U, Johnston M, Zhang M, Yao X. 2012. Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*. 17(3):368–386. doi: [10.1109/TEVC.2012.2199119](https://doi.org/10.1109/TEVC.2012.2199119).
- Buschmann AH, Muñoz JL. 2019. Challenges for future salmonid farming. In: Cochran JK, Bokuniewicz H, Yager P, editors. *Encyclopedia of ocean sciences*. 3rd ed. Amsterdam: Elsevier.
- Casanovas P, Walker SP, Johnston H, Johnston C, Symonds JE. 2021. Comparative assessment of blood biochemistry and haematology normal ranges between Chinook salmon (*Oncorhynchus tshawytscha*) from seawater and freshwater farms. *Aquaculture*. 537:736464. doi: [10.1016/j.aquaculture.2021.736464](https://doi.org/10.1016/j.aquaculture.2021.736464).
- Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing*. 408:189–215. doi: [10.1016/j.neucom.2019.10.118](https://doi.org/10.1016/j.neucom.2019.10.118).
- Charbuty B, Abdulazeez A. 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*. 2(01):20–28. doi: [10.38094/jastt20165](https://doi.org/10.38094/jastt20165).
- Devarriya D, Gulati C, Mansharamani V, Sakalle A, Bhardwaj A. 2020. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications*. 140:112866. doi: [10.1016/j.eswa.2019.112866](https://doi.org/10.1016/j.eswa.2019.112866).
- Dahri H, Al Maghayreh E, Mahmood A, Elkilani W, Nagi MF. 2019. Automated breast cancer diagnosis based on machine learning algorithms. *Journal of Healthcare Engineering*. 2019: Article ID 4253641, 11 pages. doi: [10.1155/2019/4253641](https://doi.org/10.1155/2019/4253641).
- Elvy JE, Symonds JE, Hilton Z, Walker SP, Tremblay LA, Casanovas P, Herbert NA. 2022. The relationship of feed intake, growth, nutrient retention, and oxygen consumption to feed conversion ratio of farmed saltwater Chinook salmon (*Oncorhynchus tshawytscha*). *Aquaculture*. 554:738184. doi: [10.1016/j.aquaculture.2022.738184](https://doi.org/10.1016/j.aquaculture.2022.738184).
- Esmaili M, Carter CG, Wilson R, Walker SP, Miller MR, Bridle AR, Symonds JE. 2022. Proteomic investigation of brain, liver and intestine in high feed intake and low feed intake Chinook salmon (*Oncorhynchus tshawytscha*). *Aquaculture*. 551:737915. doi: [10.1016/j.aquaculture.2022.737915](https://doi.org/10.1016/j.aquaculture.2022.737915).
- Esmaili N, Carter CG, Wilson R, Walker SP, Miller MR, Bridle A, Symonds JE. 2021. Proteomic investigation of liver and white muscle in efficient and inefficient Chinook salmon (*Oncorhynchus tshawytscha*): fatty acid metabolism and protein turnover drive feed efficiency. *Aquaculture*. 542:736855. doi: [10.1016/j.aquaculture.2021.736855](https://doi.org/10.1016/j.aquaculture.2021.736855).
- Espejo PG, Ventura S, Herrera F. 2009. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 40(2):121–144. doi: [10.1109/TSMCC.2009.2033566](https://doi.org/10.1109/TSMCC.2009.2033566).
- Feddern ML, Schoen ER, Shaftel R, Cunningham CJ, Chythlook C, Connors BM, Murdoch AD, VR von Biela, Woods B. 2023. Kings of the north: bridging disciplines to understand the effects of changing climate on Chinook salmon in the Arctic-Yukon-Kuskokwim region. *Fisheries*. 48:317–356.
- Fushiki T. 2011. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*. 21:137–146. doi: [10.1007/s11222-009-9153-8](https://doi.org/10.1007/s11222-009-9153-8).
- Kim J, Jeong J, Shin J. 2020. M2m: imbalanced classification via major-to-minor translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA. p. 13896–13905.
- Krawczyk B, Galar M, Jeleń Ł, Herrera F. 2016. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*. 38:714–726. doi: [10.1016/j.asoc.2015.08.060](https://doi.org/10.1016/j.asoc.2015.08.060).
- Kumar A, Sinha N, Bhardwaj A. 2020. A novel fitness function in genetic programming for medical data classification. *Journal of Biomedical Informatics*. 112:103623. doi: [10.1016/j.jbi.2020.103623](https://doi.org/10.1016/j.jbi.2020.103623).
- Liu ZW, Chen G, Fan Dong C, Qiu W-R, Hua Zhang S. 2023. Intelligent assistant diagnosis for pediatric inguinal hernia based on a multilayer and unbalanced classification model. *Frontiers in Physiology*. 14:384.

- Lulijwa R, Young T, Symonds JE, Walker SP, Delorme NJ, Alfaro AC. 2021. Uncoupling thermo-tolerance and growth performance in Chinook salmon: blood biochemistry and immune capacity. *Metabolites*. 11(8):547. doi: [10.3390/metabo11080547](https://doi.org/10.3390/metabo11080547).
- NZKS. 2020. New Zealand king salmon. [accessed 2023 Sept 24]. <https://www.kingsalmon.co.nz/our-salmon/>.
- Pei W, Xue B, Shang L, Zhang M. 2021. Genetic programming for development of cost-sensitive classifiers for binary high-dimensional unbalanced classification. *Applied Soft Computing*. 101:106989. doi: [10.1016/j.asoc.2020.106989](https://doi.org/10.1016/j.asoc.2020.106989).
- Pei W, Xue B, Shang L, Zhang M. 2022. High-dimensional unbalanced binary classification by genetic programming with multi-criterion fitness evaluation and selection. *Evolutionary Computation*. 30(1):99–129. doi: [10.1162/evco_a_00304](https://doi.org/10.1162/evco_a_00304).
- Poli R, Langdon WB, McPhee NF. 2008. A field guide to genetic programming. San Francisco, CA: Springer.
- Qi C, Tang X, Dong X, Chen Q, Fourie A, Liu E. 2019. Towards intelligent mining for backfill: a genetic programming-based method for strength forecasting of cemented paste backfill. *Minerals Engineering*. 133:69–79. doi: [10.1016/j.mineng.2019.01.004](https://doi.org/10.1016/j.mineng.2019.01.004).
- Reddy GT, Reddy MPK, Lakshman K, Rajput DS, Kaluri R, Srivastava G. 2020. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*. 13:185–196. doi: [10.1007/s12065-019-00327-1](https://doi.org/10.1007/s12065-019-00327-1).
- Salmi N, Rustam Z. 2019. Naïve bayes classifier models for predicting the colon cancer. In: IOP Conference Series: Materials Science and Engineering. Vol. 546. Bristol: IOP Publishing. p. 052068.
- Santoso LW, Singh B, Rajest SS, Regin R, Hameed Kadhim K. 2021. A genetic programming approach to binary classification problem. *EAI Endorsed Transactions on Energy Web*. 8(31):e11–e11.
- Speiser JL, Miller ME, Tooze J, Ip E. 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*. 134:93–101. doi: [10.1016/j.eswa.2019.05.028](https://doi.org/10.1016/j.eswa.2019.05.028).
- Stead SM, Laird L. 2002. The handbook of salmon farming. Berlin: Springer Science & Business Media.
- Theobald O. 2017. Machine learning for absolute beginners: a plain English introduction. Vol. 157. London: Scatterplot Press.
- Winston PH. 1984. Artificial intelligence. Boston: Addison-Wesley Longman.
- Zhang M-L, Zhou ZH. 2007. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition*. 40(7):2038–2048. doi: [10.1016/j.patcog.2006.12.019](https://doi.org/10.1016/j.patcog.2006.12.019).
- Zhao R, Symonds JE, Walker SP, Steiner K, Carter CG, Bowman JP, Nowak BF. 2021. Effects of feed ration and temperature on Chinook salmon (*Oncorhynchus tshawytscha*) microbiota in freshwater recirculating aquaculture systems. *Aquaculture*. 543:736965. doi: [10.1016/j.aquaculture.2021.736965](https://doi.org/10.1016/j.aquaculture.2021.736965).
- Zhou ZH. 2021. Machine learning. Singapore: Springer Nature.