


RESEARCH PAPER

Open Access



# A study of feature importance for king salmon health classification with feature selection

Yuye Zhang<sup>1</sup>, Fangfang Zhang<sup>1\*</sup> , Paula Casanovas<sup>2</sup>, Bing Xue<sup>1</sup>, Mengjie Zhang<sup>1</sup> and Jane E. Symonds<sup>2</sup>

## Abstract

King salmon is important for aquaculture in New Zealand, contributing significant economic value. Fish health is a priority for the industry, and the change in the health status of king salmon needs to be accurately detected at the earliest possible stage. Many factors affect the health of king salmon, such as temperature. Identifying the key features that influence health prediction is a crucial step toward achieving this goal. This study utilizes trial data collected by the Cawthron Institute, which includes diverse information on king salmon, such as blood biochemistry and hematology. We explore the data by employing statistical methods and feature selection techniques in machine learning to identify the most relevant features for king salmon health prediction, aiming to classify individuals as healthy or unhealthy with a small number of features. The results show that although the most efficient feature selection techniques on different datasets vary, overall, feature selection approaches can successfully identify relevant and informative features for king salmon health classification. Through the incorporation of a few selected features, the learned classifiers could still achieve statistically equal or better classification performance. This study not only contributes to the understanding of the health indicators of king salmon but also provides crucial insights into health prediction, which will be beneficial to the improvement of the health of king salmon, leading to the development of more effective management strategies for aquaculture.

**Keywords** Feature selection, Machine learning, King salmon, Health classification

## 1 Introduction

Aquaculture is the fastest-growing sector in New Zealand's agricultural production (Camara and Symonds 2014), where king salmon (*Oncorhynchus tshawytscha*) plays an important role and is known for its high nutritional value, e.g., rich in protein and omega-3 fatty acids (NZKS 2020). King salmon (Fig. 1), the main salmon species farmed in New Zealand, provides significant revenue for the country (i.e., approximately \$226 million in

revenue each year (Casanovas et al. 2021)) and the New Zealand Aquaculture Strategy aims to achieve NZ\$3 billion in annual sales by 2035 (New Zealand Government 2019). New Zealand produced 14180 tons of the global 19082 tons (74.3%) in 2019 (Elvy et al. 2023). Understanding the factors influencing the health of king salmon is important for farming, especially in light of climate change and increasing ocean temperatures (Behrens et al. 2022).

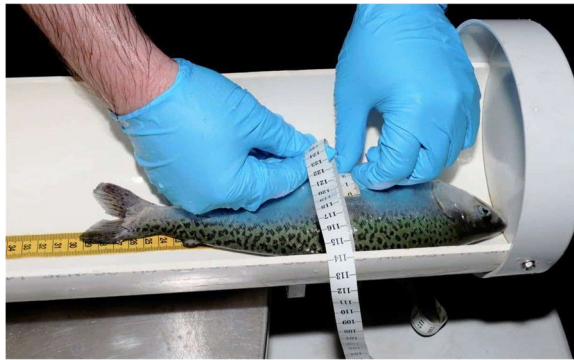
Many factors influence the health of king salmon, such as genetics, husbandry, nutrition, feed management, rearing conditions, and disease prevalence. Genetic variation affects traits such as growth, with research linking specific gene expressions to efficiency in nutrient conversion and susceptibility to physical anomalies (Scholtens et al. 2023). Feeding management, which is vital to

\*Correspondence:

Fangfang Zhang  
[fangfang.zhang@ecs.vuw.ac.nz](mailto:fangfang.zhang@ecs.vuw.ac.nz)

<sup>1</sup> Centre for Data Science and Artificial Intelligence & School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand

<sup>2</sup> Cawthron Institute, Nelson 7042, New Zealand



**Fig. 1** An example of king salmon measured in the Cawthron Institute

physiological development, is influenced by nutritional composition and feeding rations and has an impact on growth rates, feed conversion ratios, and the prevalence of spinal anomalies (Araújo et al. 2023; John et al. 2023). Rearing conditions, such as water temperature, significantly affect growth, feed intake, physiology, and gut microbiota, underscoring their importance in aquaculture settings (Elvy et al. 2022; Steiner et al. 2022). Particularly, global warming has become a crucial concern, as rising temperatures alter aquatic environments, exacerbating disease prevalence and mortality rates in farmed fish (Lane et al. 2022).

Many studies investigating fish health use one or a few factors separately for specific investigations. A comprehensive understanding of the multifaceted factors influencing the health of king salmon could help advance operational efficiency and meet the economic goals of the industry. This study uses datasets from trials conducted in the Finfish Research Centre (FRC) at the Cawthron Institute in New Zealand. The data comprise fish information organized in different datasets, such as blood (i.e., haematology and plasma biochemistry), composition, feeding, growth, external and internal assessments, histology, and biometrics. Each dataset contains a different number of features.

Each fish is labeled as either healthy or unhealthy based on the fish health criteria described in previous studies (Casanovas et al. 2021; Zhang et al. 2024). Variables used for these criteria have not been included as features in this analysis. The king salmon health prediction conducted in this study is a binary classification task. The key to this task is the current lack of clarity regarding which features are important for health classification. The involvement of irrelevant features may bring noise to the identification of classifiers for classification tasks, leading to low classification accuracy. Additionally, this approach could help farmers enhance efficiency and reduce costs

by focusing on fewer features for health testing. The focus of this study is to identify the most likely important features of king salmon health classification tasks.

Statistical methods are integral to biological research, providing tools for generating meaningful insights (Iniesta et al. 2016). Meanwhile, the adoption of machine learning techniques in aquaculture has improved the efficiency of fish breeding in recent years (Zhao et al. 2021). However, studies using machine learning to identify important features for predicting the health of king salmon are still limited.

The goal of this study is to use statistical methods and feature selection approaches to select the most important features for king salmon health prediction. The proposed classification algorithm is expected to have a high classification accuracy but a small number of features. Our key contributions include the following:

- (1) The feature selection methods successfully identified key features for king salmon health classification. These selected features were then used in health classification using the support vector machine (SVM) algorithms, achieving similar classification accuracy to those using all available features, which indicates the effectiveness of the feature selection process.
- (2) When comparing the performance of statistical methods and feature selection approaches, they both achieved similar and statistically better performance than using all available features for king salmon health classification. In addition, feature selection approaches perform better than traditional statistical methods in selecting a small number of features.
- (3) Our findings can help king salmon farmers improve the detection of unhealthy populations in aquaculture by focusing on 28.32% of the features, thereby avoiding the inefficiencies of evaluating all available features.

This paper is organized as follows: Section 2 provides the background of this study. Section 3 provides the details of the data and statistical methods used in this study and the results of the statistical methods. Section 4 describes the applied feature selection algorithms and discusses the performance of learned classifiers by feature selection approaches via a comparison with classical statistical methods. Section 5 concludes the paper.

## 2 Background

### 2.1 Classification

Machine learning, a rapidly evolving area, empowers computational systems with the capability to extract

insights from data (Jordan and Mitchell 2015). Classification is a typical machine learning task that involves the training of classifiers on labeled training data and subsequently applying the learned classifiers to predict labels of the unseen data. Among various classification algorithms, SVMs have exhibited better performance than logistic regression (LR) and random forest (RF) algorithms, particularly on biological data (Abeel et al. 2010). The effectiveness of SVMs in various diagnostic tasks involving biological data highlights their appropriateness for king salmon health prediction. Therefore, the SVM was chosen as the classifier to predict the labels of unseen data.

## 2.2 Statistical methods

Statistical methods are integral to biological research, providing tools for generating meaningful insights and validating hypotheses (Iniesta et al. 2016). Foundational techniques, such as hypothesis testing,  $P$ -values, and confidence intervals, enable the rigorous evaluation of the statistical significance of the data. In the research field of king salmon, these methods have been utilized to investigate whether a specific feature performs differently between unhealthy and healthy groups. Studies typically begin with normality and variance homogeneity tests, employing independent  $t$ -tests or nonparametric tests, such as the Kruskal-Wallis test for non-normal distributed data (Casasnovas et al. 2021). A consistent significance level of  $P < 0.05$  is maintained, and the Wilcoxon rank-sum test is often used for model performance comparisons. More tests, such as one-way analysis of variance and Tukey's honestly significant difference test, are also applied to discern differences between conditions, which are crucial to understanding the different traits of king salmon (Araújo et al. 2023). However, previous studies in this field have often focused on environmental features, and our study diverges by using statistical methods to examine the health classification of king salmon from multiple sources of information.

## 2.3 Feature selection methods

Feature selection, a crucial step in classification, involves identifying and retaining the most significant features for classifier learning. This process can reduce the number of features, which plays an essential role in improving learning performance, preventing overfitting, and reducing computational costs. With fewer features, we can classify the health status of king salmon more cost-effectively in terms of both time and effort. For example, analyzing the fish body's chemical composition requires special equipment and reagents (Araújo et al. 2022). By reducing the number of features, we can also reduce the associated costs of the analysis. Feature selection techniques are

generally categorized into filter, wrapper, and embedded methods (Xue et al. 2015). Feature selection methods are rarely used in king salmon health analyses. However, some works in related domains, such as medical studies, have been reported.

**Filter methods:** These methods, such as Relief, mutual information (referred to as *mutuInfo*), and chi-squared, evaluate and score each feature using specific metrics based on their inherent properties, independent of any machine learning algorithms. Therefore, these methods have been recognized for their computational efficiency and capability to generalize and contribute to robust data handling in complex scenarios, such as processing the medical data used in brain tumor diagnoses (Huda et al. 2016).

**Wrapper methods:** Wrapper methods, such as recursive feature elimination, assess feature subsets by training models using model performance as a criterion. These methods account for feature interactions and often provide better results than filter methods but are more computationally intensive. Although direct studies of king salmon health classification are rare, their success in diagnosing diseases (Senan et al. 2021) in other biological contexts indicates their potential utility.

**Embedded methods:** Embedded methods integrate feature selection into classifier construction. These methods balance the characteristics of the filter and wrapper methods. These methods have also been used in identifying genetic markers and features in cardiovascular studies (Kang et al. 2019). The current literature reflects a predominant reliance on statistical methods for feature importance investigations for aquaculture, revealing a gap in the application of feature selection techniques, which underscores the importance of adopting a comprehensive study of different types of feature selection techniques for king salmon health classification.

## 3 Datasets

All salmon used in this study were sourced from Sanford's commercial hatchery, located in Kaitangata, and subsequently reared in freshwater by Salmon Smolt New Zealand, located in Kaiapoi. Following this phase, all salmon were transferred to the FRC at the Cawthron Aquaculture Park in Nelson, New Zealand. The experimental design is organized into three trials, with each trial comprising a series of experimental events spaced at varied intervals. Different observables and environmental features were assessed during these events, as shown in Fig. 2. Additional details of the methods used to generate the data are described in previous studies (Casasnovas et al. 2021; Young et al. 2023). We treat each aspect across trials as a dataset, and each dataset with several example features assessed is briefly described as follows:

- (1) Blood biochemistry and hematology, e.g., alanine aminotransferase.
- (2) Body chemical composition, e.g., percentage of C20:4n6 arachidonic acid in the whole body measured using fatty acid methyl esters.
- (3) Feeding and feed conversion ratio, e.g., daily feed intake.
- (4) Biometrics, e.g., organ weights.
- (5) Growth: fork length, girth, and weight.
- (6) External and internal assessments, e.g., spinal deformity.
- (7) Histological evaluations, e.g., cellular inflammation of the midgut.
- (8) Trial information: aquaculture tank environment for all sampled fish (e.g., temperature °C) and feeding rations (e.g., 0 satiation ration indicates fasting treatment).
- (9) Health classification: ‘healthy’ or ‘unhealthy’.

### 3.1 Fish health criteria

Table 1 shows the criteria for measuring the health status of king salmon, established by researchers from the Cawthron Institute. Unhealthy fish do not satisfy at least one criterion, whereas healthy fish meet all of the criteria. ‘Event’ in Fig. 2 means the assessment event during the trial. If a single fish has multiple records of health status at different events, then this study only uses the health label at the last event for that fish (i.e., highest event value) because this represents the most recent record and, thus, the most accurate health information available for that fish.

### 3.2 Data preprocessing

Several steps are performed to generate the final datasets. First, converting non-numeric features: Non-numeric features are converted into numeric formats to suit statistical methods and feature selection algorithms. For

example, ‘m’ representing male is converted into 0, and ‘f’ representing female is converted into 1. Labels for unhealthy and healthy king salmon are represented by 0 and 1, respectively. Thus, all available features are numerically represented and categorized as continuous or discrete values. Second, feature restructured: This analysis begins with the creation of extra features in the datasets to make each piece of information represent a unique feature. For example, in the Histology datasets, the feature ‘inflammation’ corresponds to many body parts; then, we combine ‘inflammation’ with body parts to generate new features, such as ‘inflammation\_heart’ to achieve a unique feature representation. Third, data integration: Environmental conditions and feeding rations are integrated from the trial information dataset into other datasets. The labels from the health classification collection are integrated into other datasets. Data integration involves using both the fish ID and the event as a unique identifier because of the multiple records that might exist for a single fish from different events. Fourth, trials integration: The integration of three trials into a single dataset for each collection enables a comprehensive analysis of how different environmental conditions and treatments affect the overall results. Only common features across the three trials are kept for consistency in the analytical process. Finally, handling missing data: The  $k$ -nearest neighbor algorithm with  $k$  set to 5 is used for the imputation technique to fill in the missing values in all datasets.

### 3.3 Datasets information

In this study, our primary attention is centered on the following seven datasets: blood biochemistry and hematology (referred to as blood), body chemical composition (referred to as composition), feeding and feed conversion ratio (referred to as FCR), biometrics, growth, sample assessments (referred to as assessment),

**Table 1** Health criteria for measuring the health status of king salmon

Criteria	Collection	Details
Weight loss/abnormal CF	Growth measurement	Weight loss; low condition factor (exclude if CF < 1.1)
Haematology	Blood analyses	Abnormal appearance of leucocytes, erythrocytes, and thrombocytes
Abnormal white cells	Blood analyses	Reduced: lymphocyte < 87%; increased: neutrophils > 10%; increased: monocytes > 2%
Abnormal stomach or swim bladder	Health assessment	Stomach: abnormal visual assessment; swim bladder: abnormal fluid volume (> 1 mL if < 500 g, > 2 mL if > 500 g); stomach width (> 20 mm if < 500 g, > 35 mm if > 500 g)
Abnormal kidney, or liver, or faeces	Health assessment	Kidney: nephrocalcinosis score ( $\geq 3$ ); high faecal score (3+); low liver index (< 0.75); low CF exclusion
Abnormal histology	Histology analyses	High total histology score (> 12)
High inflammation	Histology analyses	High GI tract inflammation score (> 5); high histology inflammation score (> 10)
Comments	Comments	Based on sampling or assessment comments



Trial	Event	Salinity	Ration(s) at the event	Temperature(s) at the event (°C)	Start date	End date	Comment
1	Arrival in the FRC				Aug 21, 2018	Aug 21, 2018	
	WT2	FW	100	15	Sep 11, 2018	Sep 14, 2018	Assessment before temperature change
	WT4	FW	60, 80, 100	13, 17	Oct 15, 2018	Oct 23, 2018	
	WT7	FW	60, 80, 100	13, 17	Nov 26, 2018	Dec 06, 2018	
	WT10	FW	60, 80, 100	17	Jan 21, 2019	Jan 23, 2019	WT7–WT10 = 17°C only
	WT14	FW	60, 80, 100	17	Mar 12, 2019	Mar 28, 2019	WT10–WT14 = 17°C only
2	Arrival in the FRC				Dec 17, 2018	Dec 18, 2018	
	WT2	SW	100	17	Jan 31, 2019	Feb 01, 2019	
	WT3	SW	100	17	Feb 12, 2019	Feb 13, 2019	
	WT4	SW	100	17	Apr 15, 2019	Apr 18, 2019	
	WT5	SW	100	17	Jun 10, 2019	Jun 27, 2019	
	WT6	SW	100	17	Jul 29, 2019	Aug 12, 2019	
	WT7	SW	100	17	Sep 30, 2019	Oct 22, 2019	
	WT9	SW	100	17	Nov 18, 2019	Dec 03, 2019	
3	Arrival in the FRC		NA		May 06, 2020	May 25, 2020	
	WT2	FW	100	14	Jun 08, 2020	Jun 10, 2020	
	WT3	FW	100	14	Jun 15, 2020	Jun 17, 2020	
	WT4	FW	100	8, 12, 16, 20	Jul 06, 2020	Jul 16, 2020	
	WT5	FW	100	8, 12, 16, 20	Aug 05, 2020	Aug 18, 2020	End of 100% ration
	WT6	FW	25	8, 12, 16, 20	Aug 26, 2020	Sep 08, 2020	WT5–WT6 = 25% ration
	WT7	FW	25	8, 12, 16, 20	Sep 16, 2020	Sep 29, 2020	WT6–WT7 = 25% ration
	WT8	FW	0	8, 12, 16, 20	Oct 14, 2020	Oct 28, 2020	WT7–WT8 = fasting (0% ration)

**Fig. 2** Trial information and details for each event in three trials, where ‘FW’ denotes freshwater, whereas ‘SW’ denotes seawater. The ‘Ration’ value designates the percentage of the satiation ration

**Table 2** Sizes of formed datasets represented by the number of samples and features, and the imbalance ratio of general health classification

Dataset	No. of smpls	No. of features	Imbalance ratio
Blood	923	36	1.02
Composition	754	95	0.91
FCR	338	9	0.74
Growth	924	7	1.02
Assessment	923	8	1.02
Histology	924	38	1.02
Biometrics	923	9	1.02

and histology. After data preprocessing, Table 2 shows the sizes and imbalance ratios of the extracted datasets, where the class imbalance ratio is computed by dividing the number of unhealthy fish by the number of healthy fish. Taking the Blood dataset as an example, the dataset comprises 923 fish samples and 36 features. Most of the datasets have a balanced ratio close to 1, indicating a similar number of healthy and unhealthy fish. The data are split into training (80%) and test (20%), and the feature values are normalized to [0, 1].

### 3.4 Statistical analysis

Statistical tests are employed to identify features with significant differences between healthy and unhealthy king salmon. The process involves using the Shapiro-Wilk test for normality (with a  $P$ -value  $> 0.05$  indicating normal distribution) and Levene’s test for homogeneity of variances. For normally distributed and homogenous data (both Shapiro-Wilk and variance chi-squared tests with  $P > 0.05$ ), a t-test is used. For normally distributed but heterogeneous data ( $P \leq 0.05$  in variance chi-squared test), Welch’s t-test is applied. If data are not normally distributed ( $P \leq 0.05$  in the normality test), then the Wilcoxon rank-sum test is used.

The number of features that are statistically different between the healthy and unhealthy groups varies for each dataset, indicating a certain degree of redundancy or irrelevant information. For the blood (26 features), composition (63 features), and histology (23 features) datasets, approximately two-thirds of the features were selected. Nearly all features in the FCR (7 features), growth (7 features out of 7 features), assessment (7 features), and biometrics (8 features) datasets are statistically different between healthy and unhealthy king salmon. Thus, it is essential to tailor the feature sets for classification to ensure optimal accuracy and efficiency.

## 4 Feature selection approaches

This section focuses on investigating different feature selection approaches for fish analysis. The main goal is to reduce the number of features for health classification without reducing the classification accuracy.

### 4.1 Choice of methods

- (1) Filter methods we used include the ReliefF, chi-squared, and mutual information represented by mutuInfo, which evaluate the relevance of features to target variables. Furthermore, a new intersection method is proposed to select features that are universally recognized as significant across ReliefF, chi-squared, and mutual information. Alternatively, a union of features selected through the three distinct methods is utilized, thereby furnishing a comprehensive set of features.
- (2) Wrapper methods evaluate subsets of features using classifiers on each subset and employ classification accuracy as a criterion for feature selection. One common technique in this approach is recursive feature elimination with cross-validation, which uses the classifier to assign weights to each feature and iteratively removes the least important features based on these weights. This process involves repeatedly constructing a model and calculating its accuracy to ensure robustness across multiple subsets of the data. By integrating feature selection directly into model performance evaluation, wrapper methods provide a more comprehensive understanding of feature relevance than filter methods.
- (3) Embedded methods are similar to wrapper methods but integrate feature selection into the process of classifier training. Features that are retained have an absolute value that is higher than or equal to the medium value. The LR, SVM, and RF algorithms are popularly used as classifiers for the wrapper and embedded methods.

### 4.2 Classification metrics

The *F1* score, a widely acknowledged measure for classification tasks, is chosen as the evaluation metric. The formulas for the evaluation metrics employed are provided. This study mainly focuses on detecting unhealthy king salmon and thus considers the unhealthy class as the positive class. For the *F1* score and *recall* scores, a higher value consistently signifies a better performance.

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN},$$

$$recall = \frac{TP}{TP + FN},$$

where *TP* is the true positive, *FP* is the false positive, *FN* is the false negative, and *TN* is the true negative.

The primary goal of this study is to identify features that are crucial to distinguishing the health status of king salmon. Good results are obtained by models that use fewer features but achieve performance equal to, or better than, those from the full feature set. Each feature selection algorithm is run for 30 independent runs. The effectiveness of each feature selection method is compared using the Wilcoxon rank-sum test, with a significance level of 0.05. Performance indicators are notated using the symbols ‘–’, ‘≈’, and ‘+’ to indicate a decrease, similar, and significantly better performance, respectively, compared with the performance of the algorithm with all available features.

### 4.3 Results of the filter methods

In this section, the values under each column represent the results of the performance of the SVM model employing diverse feature selection techniques. The ‘Allfeatures’ method incorporates all available features present in the dataset without feature selection, which is our baseline. ‘ReliefF’, ‘mutuInfo’, and ‘chi-squared’ are the filter feature selection techniques. The ‘Intersection’ technique adopts a consensus-driven strategy, preserving only those features that are selected by all feature selection methods. The ‘Union’ approach maintains any feature selected by at least one of the individual methods.

Table 3 shows the mean and standard deviations of the *F1* and recall scores obtained by the filter methods on the test datasets. The results show a general trend of comparable classifier performance when the filter methods are employed compared with using all available features. Table 4 shows the average number of selected features across the 30 independent runs by the filter methods. The intersection method selects the smallest number of features for each dataset, as highlighted in bold.

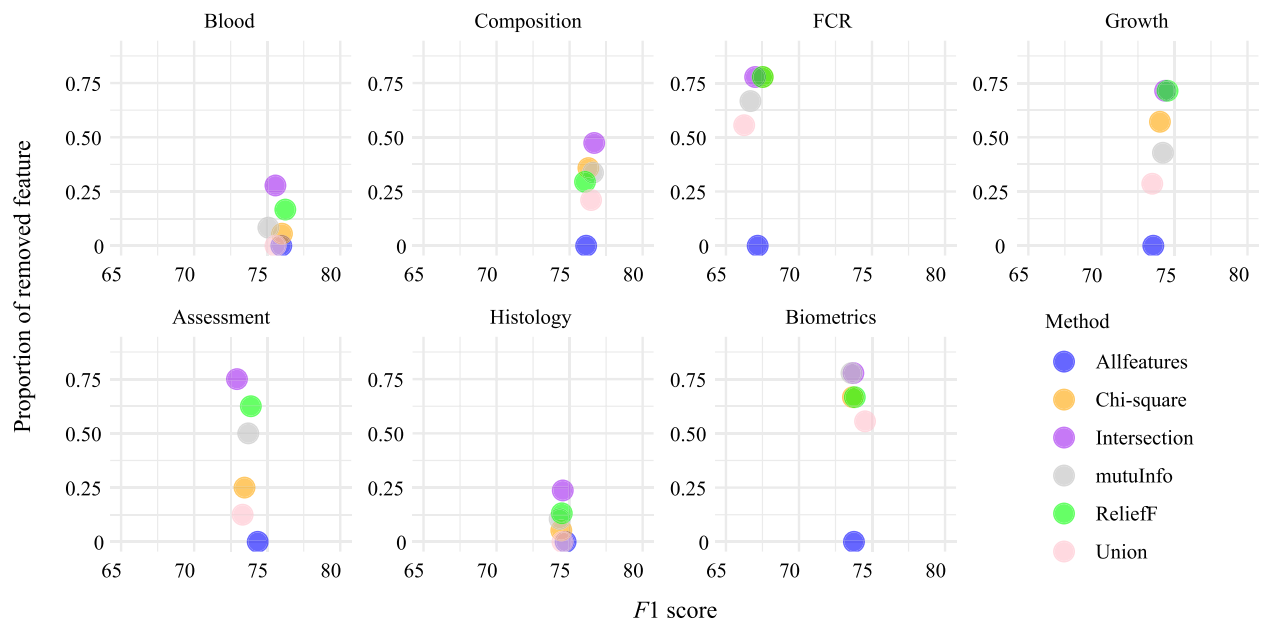
Figure 3 shows the scatter plots of the test *F1* score and the percentages of removed features of the filter algorithms. The union filter is the worst method among the examined filter methods, where the pink dots are always shown close to the blue dots at the bottom. The performance along the vertical axis (*y*-axis) shows that all algorithms had a similar performance. The intersection algorithm represented by the purple dots had comparable classification accuracy for detecting unhealthy king salmon. This finding indicates that the intersection method could be preferable in terms of model simplicity and performance trade-offs.

**Table 3** The mean and standard deviations of *F1* scores and *recall* on test datasets obtained by filter methods

	Dataset	Allfeatures	ReliefF	mutuInfo	Chi-square	Intersection	Union
<i>F1</i> score	Blood	75.71 (3.06)	76.15 (2.86) (≈)	75.40 (2.95) (≈)	75.85 (3.26) (≈)	75.77 (3.08) (≈)	75.71 (3.06) (≈)
	Composition	76.47 (3.08)	76.28 (3.19) (≈)	76.56 (2.99) (≈)	76.42 (3.21) (≈)	76.53 (3.02) (≈)	76.43 (3.11) (≈)
	FCR	67.13 (3.80)	67.49 (3.86) (≈)	66.52 (4.45) (≈)	67.49 (3.86) (≈)	67.00 (4.27) (≈)	66.57 (4.44) (≈)
	Growth	73.48 (2.92)	74.29 (2.11) (≈)	74.25 (1.86) (≈)	73.77 (3.65) (≈)	73.95 (2.21) (≈)	73.41 (4.16) (≈)
	Assessment	74.14 (2.18)	73.94 (2.50) (≈)	73.86 (2.32) (≈)	73.60 (3.12) (≈)	73.20 (3.48) (≈)	73.59 (3.33) (≈)
	Histology	74.49 (2.90)	74.58 (3.00) (≈)	74.40 (3.18) (≈)	74.50 (3.14) (≈)	74.41 (3.00) (≈)	74.49 (2.90) (≈)
	Biometrics	73.60 (2.59)	74.05 (2.52) (≈)	74.07 (2.44) (≈)	73.83 (3.63) (≈)	73.97 (2.18) (≈)	74.26 (2.42) (≈)
<i>recall</i>	Blood	80.72 (4.89)	81.76 (4.82) (≈)	80.36 (4.56) (≈)	80.93 (5.17) (≈)	81.51 (4.79) (≈)	80.72 (4.89) (≈)
	Composition	80.28 (4.53)	79.81 (4.68) (≈)	80.00 (4.58) (≈)	80.00 (4.72) (≈)	80.00 (4.52) (≈)	80.14 (4.64) (≈)
	FCR	85.86 (5.68)	86.67 (5.34) (≈)	84.60 (7.72) (≈)	86.67 (5.34) (≈)	85.63 (6.90) (≈)	84.71 (7.77) (≈)
	Growth	79.07 (5.19)	81.51 (3.16) (≈)	81.08 (3.28) (≈)	79.89 (6.16) (≈)	80.86 (3.09) (≈)	79.14 (7.26) (≈)
	Assessment	78.82 (3.12)	79.89 (4.13) (≈)	79.64 (3.45) (≈)	77.99 (4.83) (≈)	78.57 (5.94) (≈)	77.92 (5.53) (≈)
	Histology	79.61 (4.75)	80.07 (4.95) (≈)	79.64 (5.21) (≈)	79.50 (4.84) (≈)	80.04 (4.81) (≈)	79.61 (4.75) (≈)
	Biometrics	79.18 (4.64)	81.76 (3.54) (+)	82.15 (3.48) (+)	79.93 (6.14) (≈)	80.36 (3.18) (≈)	80.65 (3.87) (≈)

**Table 4** The average number of selected features by the filter methods

Dataset	Allfeatures	ReliefF	mutuInfo	Chi-square	Intersection	Union
Blood	36	30	33	34	<b>26</b>	36
Composition	95	67	63	61	<b>50</b>	75
FCR	9	<b>2</b>	3	<b>2</b>	<b>2</b>	4
Growth	7	<b>2</b>	4	3	<b>2</b>	5
Assessment	8	3	4	6	<b>2</b>	7
Histology	38	33	34	36	<b>29</b>	38
Biometrics	9	3	<b>2</b>	3	<b>2</b>	4



**Fig. 3** Comparison of filter based feature selection methods and their impact on model performance (test *F1* score) across different datasets

Overall, of all of the filter-based feature selection methods, the intersection method has the best performance in terms of classification accuracy and number of selected features.

#### 4.4 Results of the wrapper methods

Table 5 shows the mean and standard deviations of the *F1* and *recall* scores of the wrapper methods. The results show that wrapper-based feature selection methods achieve a similar performance on most datasets. Exceptions are the *F1* score for growth using the LR method and the *recall* results for blood using the SVM and RF methods.

Table 6 shows in detail the average number of features selected by the wrapper method. The results show that LR can select a smaller set of features than SVM and RF.

Figure 4 shows the scatter plots of the test *F1* score and the percentages of removed features of the wrapper algorithms. The results show that LR is the most effective wrapper method, with its purple dots consistently positioned higher than the other dots for most datasets. Therefore, the wrapper method using LR has comparable performance to the other algorithms in terms of classification performance but selects a smaller number of features.

Overall, we determine that, for the wrapper-based feature selection methods, LR performs well in terms of classification accuracy and number of selected features.

#### 4.5 Results of the embedded methods

Table 7 shows the mean and standard deviations of the test *F1* and *recall* scores of embedded methods using all available features, LR, SVM, and RF. Among these, LR

**Table 6** The average number of selected features by the wrapper methods

Dataset	Allfeatures	LR	SVM	RF
Blood	36	<b>17</b>	20	26
Composition	95	<b>23</b>	24	64
FCR	9	<b>4</b>	6	7
Growth	7	<b>4</b>	6	6
Assessment	8	<b>6</b>	<b>6</b>	<b>6</b>
Histology	38	<b>12</b>	15	17
Biometrics	9	<b>5</b>	6	8

achieves similar or even significantly better *F1* and *recall* scores than using all available features across all datasets. The performance of SVM and RF varies, with most results being similar to using all available features, while others are either worse or better.

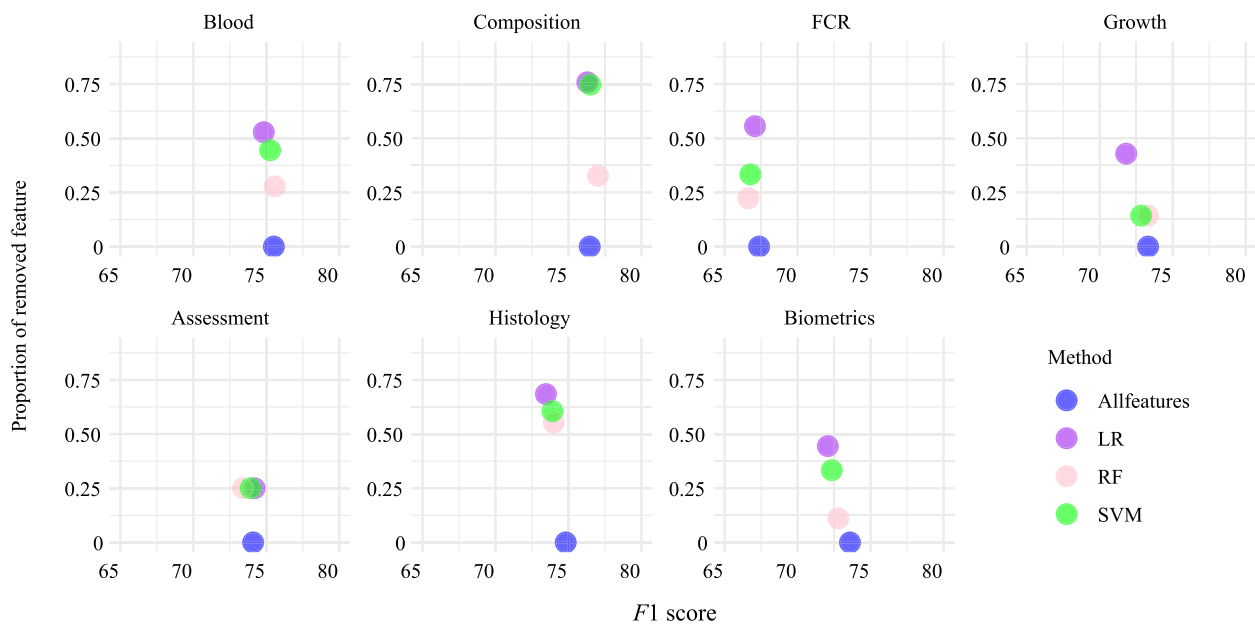
The average number of features selected, shown in Table 8, reveals a consistent reduction in feature counts across datasets when using the embedded methods. RF exhibits the best performance in terms of selecting the number of features because it selects the smallest number of features on average for four out of the seven datasets. However, RF is not preferred as it has significantly worse classification accuracy than using all available features.

Figure 5 visualizes the number of features and *F1* score of the embedded methods. The pink dots denote the RF method in the left region, which means that this method generally selects fewer features but suffers from reduced accuracy, notably in the FCR, growth, and assessment datasets. Meanwhile, the purple and green dots denote more consistent and comparable performance across datasets,

**Table 5** The mean and standard deviations of *F1* scores and *recall* on test datasets obtained by the wrapper methods

	Dataset	Allfeatures	LR	SVM	RF
<i>F1</i> score	Blood	75.71 (3.06)	74.76 (3.92) (≈)	75.23 (3.25) (≈)	75.55 (3.01) (≈)
	Composition	76.47 (3.08)	76.09 (3.34) (≈)	76.19 (3.32) (≈)	76.71 (2.93) (≈)
	FCR	67.13 (3.80)	66.99 (4.01) (≈)	66.66 (4.24) (≈)	66.78 (3.84) (≈)
	Growth	73.48 (2.92)	71.46 (3.80) (–)	72.75 (3.86) (≈)	73.39 (3.01) (≈)
	Assessment	74.14 (2.18)	74.18 (2.14) (≈)	74.15 (2.12) (≈)	73.26 (2.62) (≈)
	Histology	74.49 (2.90)	73.45 (4.69) (≈)	74.04 (3.42) (≈)	73.82 (3.73) (≈)
	Biometrics	73.60 (2.59)	72.05 (3.76) (≈)	72.56 (3.79) (≈)	72.72 (3.52) (≈)
<i>recall</i>	Blood	80.72 (4.89)	79.00 (6.05) (≈)	78.32 (5.06) (–)	78.75 (5.03) (–)
	Composition	80.28 (4.53)	79.17 (5.09) (≈)	79.68 (4.79) (≈)	80.51 (4.37) (≈)
	FCR	85.86 (5.68)	85.52 (6.36) (≈)	84.94 (6.50) (≈)	84.94 (6.24) (≈)
	Growth	79.07 (5.19)	76.63 (6.76) (≈)	77.96 (6.62) (≈)	79.18 (5.33) (≈)
	Assessment	78.82 (3.12)	79.35 (3.05) (≈)	79.46 (3.43) (≈)	78.42 (4.46) (≈)
	Histology	79.61 (4.75)	78.10 (7.66) (≈)	78.71 (5.59) (≈)	79.39 (6.10) (≈)
	Biometrics	79.18 (4.64)	76.13 (6.74) (≈)	76.95 (6.92) (≈)	77.06 (6.30) (≈)





**Fig. 4** Comparison of wrapper feature selection methods and their impact on model performance (test  $F1$  score) across different datasets

**Table 7** The mean and standard deviations of  $F1$  scores and *recall* on test datasets obtained by the embedded methods

	Dataset	Allfeatures	LR	SVM	RF
<i>F1 score</i>	Blood	75.71 (3.06)	75.34 (2.99) (≈)	75.37 (2.85) (≈)	74.77 (3.79) (≈)
	Composition	76.47 (3.08)	76.42 (3.20) (≈)	76.69 (3.20) (≈)	76.56 (3.06) (≈)
	FCR	67.13 (3.80)	65.40 (4.20) (≈)	65.69 (4.35) (≈)	56.97 (5.69) (–)
	Growth	73.48 (2.92)	71.32 (3.69) (–)	71.91 (2.69) (–)	69.28 (2.87) (–)
	Assessment	74.14 (2.18)	74.47 (1.99) (≈)	74.30 (2.01) (≈)	67.18 (3.19) (–)
	Histology	74.49 (2.90)	73.84 (3.16) (≈)	73.80 (2.76) (≈)	74.63 (3.13) (≈)
	Biometrics	73.60 (2.59)	71.94 (3.90) (≈)	72.41 (3.27) (≈)	71.14 (3.51) (–)
<i>recall</i>	Blood	80.72 (4.89)	79.03 (4.28) (≈)	77.45 (4.24) (–)	76.63 (6.60) (–)
	Composition	80.28 (4.53)	80.14 (4.63) (≈)	80.46 (4.50) (≈)	79.95 (4.69) (≈)
	FCR	85.86 (5.68)	83.33 (9.88) (≈)	86.44 (7.01) (≈)	53.56 (8.34) (–)
	Growth	79.07 (5.19)	76.24 (6.55) (≈)	76.38 (4.89) (–)	69.75 (4.00) (–)
	Assessment	78.82 (3.12)	80.97 (3.16) (+)	80.61 (3.28) (+)	65.95 (4.40) (–)
	Histology	79.61 (4.75)	78.49 (5.15) (≈)	78.10 (4.30) (≈)	81.22 (4.88) (≈)
	Biometrics	79.18 (4.64)	76.88 (7.47) (≈)	77.38 (6.22) (≈)	72.65 (6.66) (–)

which means that the LR and SVM methods are better than the RF methods.

Overall, the embedded-based feature selection using the LR or SVM method selects a smaller number of features while maintaining a stable classification performance.

#### 4.6 Discussion

##### 4.6.1 Analysis of the number of selected features

Determining which method of investigation performs the best out of the methods that we investigated is

interesting. The most effective method is selected based on the smallest subset of features. If the number of selected features is the same across methods, then the method that shows improvement in  $F1$  or *recall* score is preferred. Table 9 provides an overview of the best methods for the seven datasets and the number of selected features. The feature selection method is represented by FS\*, and the best statistical method is represented by STAT\*. The selection of the most suitable

**Table 8** The average number of selected features by the embedded methods

Dataset	Allfeatures	LR	SVM	RF
Blood	36	16	<b>13</b>	15
Composition	95	<b>37</b>	40	45
FCR	9	4	4	<b>3</b>
Growth	7	<b>3</b>	4	<b>3</b>
Assessment	8	5	4	<b>3</b>
Histology	38	17	<b>14</b>	<b>14</b>
Biometrics	9	5	<b>4</b>	5

feature selection method for each dataset is based on the *F1* score and the number of selected features.

All feature selection methods mentioned here achieved our goal of selecting fewer features while maintaining similar or better performance compared with using all available features. The results show that the filter methods are the most effective approaches for feature selection in king salmon health classification on datasets that have a small number of features in total. In addition, by comparing the number of features that are statistically different for the two health groups, we determine that feature selection approaches achieve a smaller feature subset than statistical methods ( $28.32\% < 78.58\%$ ).

#### 4.6.2 Analysis of the classification accuracy

Table 10 shows the classification accuracy and recall of using all available features, with the best feature selection

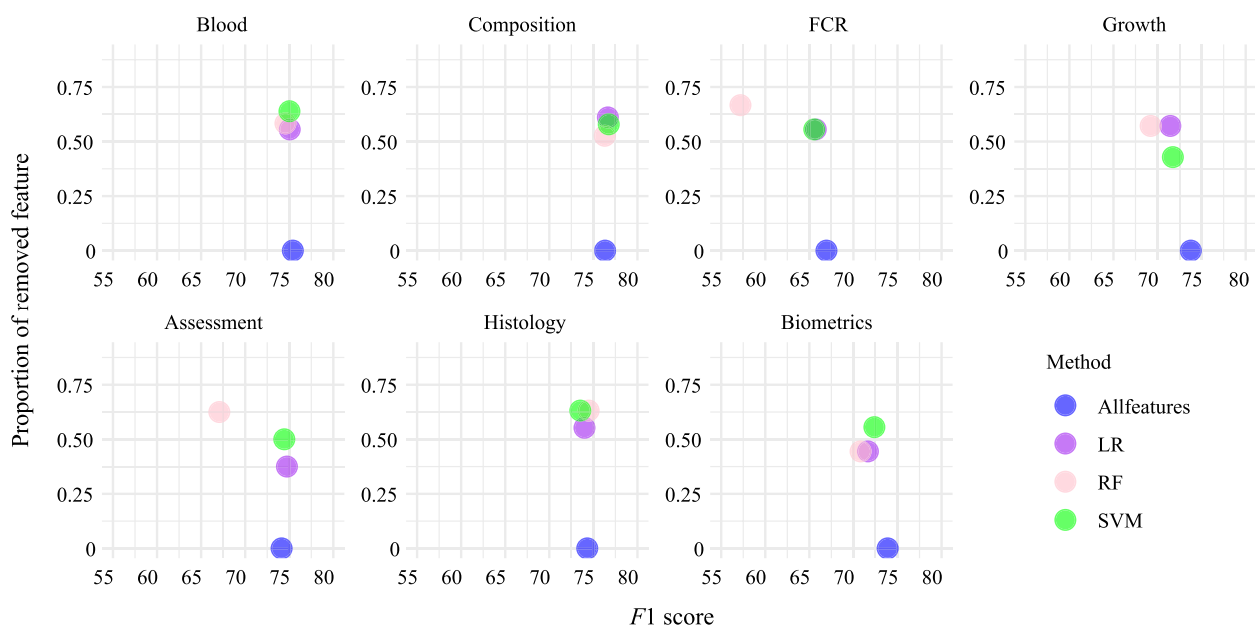
method represented by FS\* and the statistical method represented by STAT\* on different datasets. The best feature selection and statistical methods show comparable classification performance when compared with all available features. If we consider both the classification accuracy and the number of selected features, then the feature selection approaches are significantly better because they select smaller feature sets, as discussed in the previous section.

#### 4.6.3 Analysis of the feature importance

We used the Blood dataset as an example to investigate feature importance. Figure 6 shows the average

**Table 9** The best feature selection methods for seven datasets. The '#Feature' columns represent the number of selected features

Dataset	FS*	FS*/#Feature (%)	STAT*/#Feature (%)
Blood	Embedded (LR)	16/36 (44.44%)	26/36 (72.22%)
Composition	Wrapper (LR)	23/95 (24.21%)	63/95 (66.32%)
FCR	Filter (intersection)	2/9 (22.22%)	7/9 (77.78%)
Growth	Filter (ReliefF)	2/7 (28.57%)	7/7 (100%)
Assessment	Filter (intersection)	2/8 (25.00%)	7/8 (87.50%)
Histology	Wrapper (LR)	12/38 (31.58%)	23/38 (60.53%)
Biometrics	Filter (mutualInfo)	2/9 (22.22%)	8/9 (88.89%)
<b>Average</b>		<b>28.32%</b>	<b>78.58%</b>

**Fig. 5** Comparison of embedded feature selection methods and their impact on model performance (test *F1* score) across different datasets

**Table 10** The test *F1* score and *recall* score of using all features, the best feature selection method, and the statistical method on different datasets

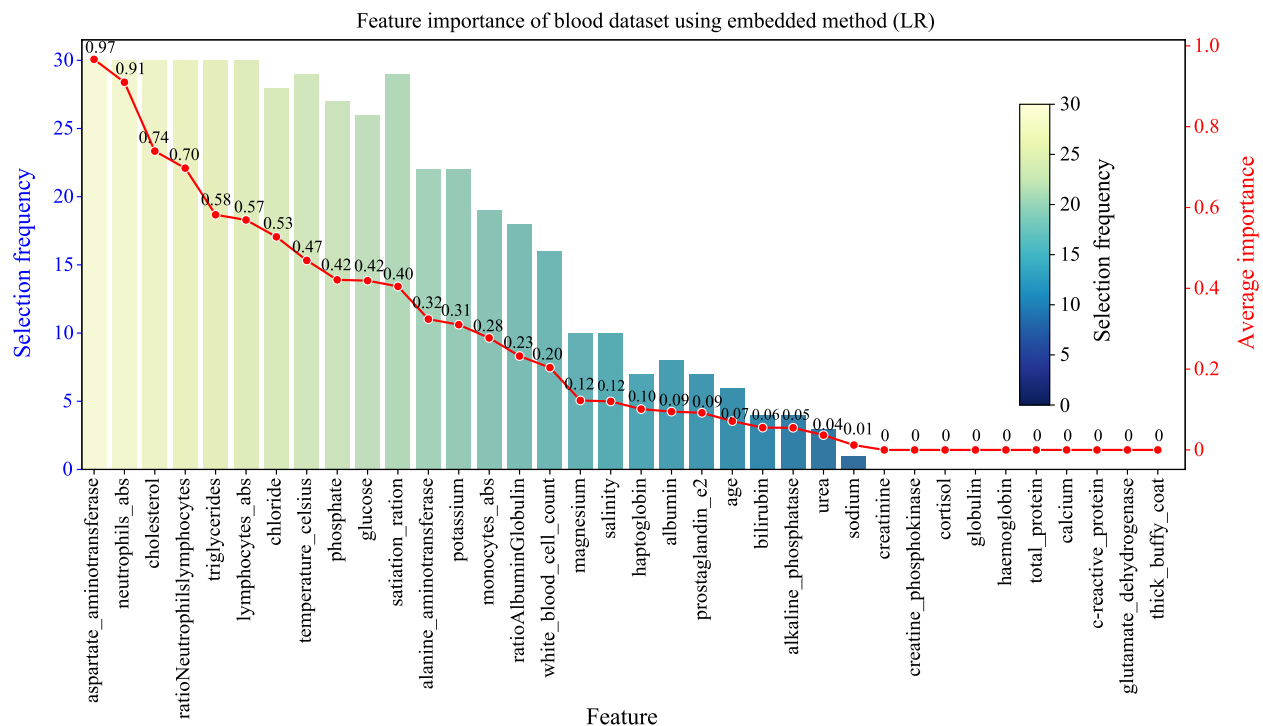
	Dataset	Allfeatures	FS*	STAT*
F1 score	Blood	75.71 (3.06)	75.34 (2.99) (≈)	75.92 (3.20) (≈)
	Composition	76.47 (3.08)	76.09 (3.34) (≈)	76.22 (3.43) (≈)
	FCR	67.13 (3.80)	67.00 (4.27) (≈)	67.34 (3.83) (≈)
	Growth	73.48 (2.92)	74.29 (2.11) (≈)	73.48 (2.92) (≈)
	Assessment	74.14 (2.18)	73.20 (3.48) (≈)	73.20 (3.12) (≈)
	Histology	74.49 (2.90)	73.45 (4.69) (≈)	74.18 (3.36) (≈)
	Biometrics	73.60 (2.59)	73.97 (2.18) (≈)	72.58 (3.73) (≈)
recall	Blood	80.72 (4.89)	79.03 (4.28) (≈)	81.65 (5.04) (≈)
	Composition	80.28 (4.53)	79.17 (5.09) (≈)	79.68 (4.98) (≈)
	FCR	85.86 (5.68)	85.63 (6.90) (≈)	86.21 (5.66) (≈)
	Growth	79.07 (5.19)	81.51 (3.16) (≈)	79.07 (5.19) (≈)
	Assessment	78.82 (3.12)	78.57 (5.94) (≈)	77.17 (4.88) (≈)
	Histology	79.61 (4.75)	78.10 (7.66) (≈)	79.57 (5.66) (≈)
	Biometrics	79.18 (4.64)	82.15 (3.48) (+)	76.81 (6.60) (≈)

frequencies and feature importance of selected features by LR, providing insights into the key factors influencing the health of king salmon. The feature selection frequency is counted as the proportion of features selected by the number of runs among the 30 independent runs.

The dataset prioritizes specific biochemical and physiological markers over others, as evidenced by their higher average importance scores and selection frequencies. The most important features, such as aspartate aminotransferase, not only exhibit a high selection frequency (30 out of the 30 runs) but also achieve high importance scores.

Tables 11 and 12 show features with significant differences between health groups and those identified by statistical methods and feature selection methods, respectively. The overlapping features, such as ‘alanine\_aminotransferase’ and ‘aspartate\_aminotransferase,’ in the enzymes category underscore their pivotal role in discerning health conditions as both statistical and feature selection methods.

Some features shown in Fig. 6, such as creatinine, exhibit zero selection frequency with zero importance scores, indicating no influence on the classifier in this particular dataset. Their absence in feature selection indicates that they are not important for king salmon health prediction in the FRC trials. However, these features are identified by statistical methods based on their distributions, indicating the effectiveness of using machine learning feature selection methods to identify important features for king salmon health prediction.



**Fig. 6** Selection frequency and average feature importance of the features selected by the embedded feature selection method using LR on the blood dataset

**Table 11** Blood data set: features statistically different for healthy and unhealthy king salmon

	Feature
Enzymes	alanine_aminotransferase, alkaline_phosphatase, aspartate_aminotransferase, glutamate_dehydrogenase
Blood biochemistry	bilirubin, c-reactive_protein, sodium, calcium, chloride, magnesium, potassium, cholesterol, creatinine, glucose, haemoglobin, haptoglobin, lymphocytes_abs, monocytes_abs, neutrophils_abs, potassium, triglycerides, urea, ratioNeutrophilslymphocytes
Other	salinity, age, temperature_celsius, satiation_ration

**Table 12** Top 13 most important features (average importance score over 0.3) identified by embedded LR feature selection method using the blood dataset for health classification

	Feature
Enzymes	alanine_aminotransferase, aspartate_aminotransferase
Blood biochemistry	lymphocytes_abs, neutrophils_abs, cholesterol, ratioNeutrophilslymphocytes, triglycerides, chloride, phosphate, glucose, potassium
Other	temperature_celsius, satiation_ration

5 Conclusions and future work

This study investigates statistical methods and feature selection approaches for king salmon health classification. We achieved the goal of using a smaller number of features to build the classifier while maintaining similar prediction performance.

In our study, we developed classification models for predicting the health of king salmon. Through data preprocessing, we effectively cleaned the raw data, resulting in seven datasets. The results show that feature selection approaches and statistical methods have a similar performance on all datasets. However, the feature subsets identified through feature selection methods always provided a smaller feature subset than those selected by traditional statistical methods. This study also provides an example of selected important features for king salmon health classification in terms of blood features. This study can benefit the king salmon industry by providing techniques that help predict fish health to improve farming efficiency.

In the future, we will test our proposed algorithm on the data collected at the farms instead of under controlled conditions in the FRC facility to investigate its effectiveness. In addition, we plan to develop an effective feature selection method by considering all aspects of king salmon information across different datasets together.

Acknowledgements

We wish to thank the staff of Cawthron Institute's Finfish Research Centre (Gareth 514 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 trials. We also want to thank the industry partners who supplied the salmon for the trials and express our gratitude to the staff in the Food Testing Laboratory at Cawthron

Institute for technical assistance and sample analysis. 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.

Additional information

Edited by: Lin Gao.

Authors' contributions

Yuye Zhang: Conceptualization, Investigation, Methodology, Formal analysis, Writing-Original Draft. Fangfang Zhang: Conceptualization, Investigation, Methodology, Formal Analysis, Supervision, Writing-Original Draft, Review & Editing. Paula Casanovas: Writing-Review & Editing. Bing Xue: Supervision, Writing-Review & Editing. Mengjie Zhang: Supervision, Writing-Review & Editing. Jane E. Symonds: Resources, Writing-Review & Editing.

Data availability

The data that have been used are confidential. The data that support the findings of this study are available from the authors upon reasonable request.

Declarations

Ethics approval and consent to participate

Animal ethics approval for the salmon sampling was provided by the Nelson Marlborough Institute of Technology Animal Ethics Committee.

Consent for publication

Not applicable.

Competing interests

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

Received: 17 July 2024   Revised: 12 August 2024   Accepted: 16 October 2024  
Published online: 07 November 2024

## References

- Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26(3):392–398
- Araújo BC, Lovett B, Preece MA, Burdass M, Symonds JE, Miller M et al (2023) Effects of different rations on production performance, spinal anomalies, and composition of Chinook salmon (*Oncorhynchus tshawytscha*) at different life stages. *Aquaculture* 562:738759
- 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. *Comp Biochem Physiol A-Mol Integr Physiol* 264:111113
- Behrens E, Rickard G, Rosier S, Williams J, Morgenstern O, Stone D (2022) Projections of future marine heatwaves for the oceans around New Zealand using New Zealand's Earth System Model. *Front Clim* 4:798287
- Camara MD, Symonds JE (2014) Genetic improvement of New Zealand aquaculture species: programmes, progress and prospects. *N Z J Mar Freshw Res* 48(3):466–491
- 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
- Elvy JE, Symonds JE, Hilton Z, Walker SP, Tremblay LA, Casanovas P et al (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
- Elvy JE, Symonds JE, Hilton Z, Walker SP, Tremblay LA, Herbert NA (2023) The relationships between specific dynamic action, nutrient retention and feed conversion ratio in farmed freshwater Chinook salmon (*Oncorhynchus tshawytscha*). *J Fish Biol* 102(3):605–618
- Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M (2016) A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. *IEEE Access* 4:9145–9154
- Iniesta R, Stahl D, McGuffin P (2016) Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med* 46(12):2455–2465
- Johne AS, Carter CG, Wotherspoon S, Hadley S, Symonds JE, Walker SP et al (2023) Modeling the effects of ration on individual growth of *Oncorhynchus tshawytscha* under controlled conditions. *J Fish Biol* 103(5):1003–1014
- Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. *Science* 349(6245):255–260
- Kang CZ, Huo YH, Xin LH, Tian BG, Yu B (2019) Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J Theor Biol* 463:77–91
- Lane HS, Brosnahan CL, Poulin R (2022) Aquatic disease in New Zealand: synthesis and future directions. *N Z J Mar Freshw Res* 56(1):1–42
- New Zealand Government MfPI (2019) Aquaculture strategy for New Zealand. <https://www.mpi.govt.nz/fishing-aquaculture/aquaculture-fish-and-shell-fish-farming/aquaculture-strategy-for-new-zealand/>. Accessed 13 Nov 2023
- NZKS (2020) New Zealand king salmon. <https://www.kingsalmon.co.nz/freshwater/>. Accessed 12 Dec 2023
- Scholtens M, Dodds K, Walker S, Clarke S, Tate M, Slattey T et al (2023) Opportunities for improving feed efficiency and spinal health in New Zealand farmed Chinook salmon (*Oncorhynchus tshawytscha*) using genomic information. *Aquaculture* 563:738936
- Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N et al (2021) Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *J Healthc Eng* 2021:1004767
- Steiner K, Laroche O, Walker SP, Symonds JE (2022) Effects of water temperature on the gut microbiome and physiology of Chinook salmon (*Oncorhynchus tshawytscha*) reared in a freshwater recirculating system. *Aquaculture* 560:738529
- Xue B, Zhang MJ, Browne WN, Yao X (2015) A survey on evolutionary computation approaches to feature selection. *IEEE Trans Evol Comput* 20(4):606–626
- Young T, Laroche O, Walker SP, Miller MR, Casanovas P, Steiner K et al (2023) Prediction of feed efficiency and performance-based traits in fish via integration of multiple omics and clinical covariates. *Biology* 12(8):1135
- Zhang FF, Zhang YY, Casanovas P, Schattschneider J, Walker SP, Xue B et al (2024) Health prediction for king salmon via evolutionary machine learning with genetic programming. *J R Soc N Z* 1–26. <https://doi.org/10.1080/03036758.2024.2329228>
- Zhao SL, Zhang S, Liu JC, Wang H, Zhu J, Li DL et al (2021) Application of machine learning in intelligent fish aquaculture: a review. *Aquaculture* 540:736724

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.