School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Internet: office@ecs.vuw.ac.nz

Machine Learning for King Salmon Health Prediction

Yuye Zhang

Supervisors: Fangfang Zhang, Bing Xue, Mengjie Zhang

Submitted in partial fulfilment of the requirements for
Bachelor of Science with Honours in Artificial intelligence.

Abstract

The health of King Salmon (Oncorhynchus tshawytscha) holds significant economic value for New Zealand but has not been thoroughly investigated. This study using machine learning-based prediction methods leverages datasets sourced from the Cawthron Institute, encompassing variables such as blood biochemistry and hematology, body composition, feeding ratios, biometrics, growth measurements, sample assessments, histological data, and environmental parameters. An exhaustive data preprocessing and exploratory analysis are conducted on the collected raw data. Feature selection is imperative in our project as it can help to identify the most relevant predictors thereby enhancing model performance. Moreover, it improves interpretability by emphasizing the most significant predictors. Subsequently, we employ machine learning techniques with feature selection methods to facilitate the health prediction of King Salmon. Through this approach, this research identifies and underscores the pivotal variables influencing the health of King Salmon, offering significant insights into aquaculture health management.

Acknowledgments

First of all, I wish to extend my deepest appreciation to my supervisors: Dr Fangfang Zhang, Prof Bing Xue, and Prof Mengjie Zhang. The foundation of this research rests firmly upon their unparalleled knowledge and experience in the field. Their persistent guidance has illuminated my path, ensuring that I remained focused and resolute in my pursuit of academic excellence. Their invaluable feedback has often been the beacon, steering me away from potential pitfalls and towards innovative solutions. Beyond their academic insights, their continuous encouragement and unwavering faith in my abilities fortified my determination, enabling me to tackle the inherent challenges of this study head-on. For all of these and more, I am deeply thankful.

Furthermore, I must convey my heartfelt appreciation to Jane Symonds, Paula Casanovas, and Jessica Schattschneider from the Cawthron Institute, who have been instrumental pillars of support throughout this research. Their unwavering dedication, profound expertise, and a keen eye for detail have significantly contributed to the advancement and quality of this project. Furthermore, I am also grateful to our collaborators from the Cawthron Institute. Their invaluable partnership, providing both experimental support and data, is indispensable to the fruition of this research. Their expert insights from a biological standpoint added depth and contextual relevance to this work. The synergy between their biological knowledge and our research objectives enabled an exploration that transcended disciplinary boundaries. To all of them, I offer my sincere thanks for their enduring support and for making this journey both enlightening and fulfilling.

Contents

Chapter 1

Introduction

1.1 Motivation

King Salmon, scientifically known as *Oncorhynchus tshawytscha*, predominantly farmed in New Zealand, holds paramount significance from both ecological and economic perspectives. The health and well-being of this species are intricately influenced by a blend of genetic, environmental, and anthropogenic factors [30, 72, 79, 90]. With the burgeoning growth of the aquaculture industry, it becomes imperative to comprehend the complex dynamics underpinning the health of King Salmon. Such understanding is crucial not only for optimizing production but also for safeguarding sustainability.

Machine learning, particularly classification models, has transformed various fields by enabling intricate pattern recognition that often surpasses traditional statistical methods[34]. For King Salmon, such classification models can classify health status by analyzing complex interrelations among variables. While traditional methods may be limited in extracting patterns from a large number of variables, machine learning can potentially assimilate information from a plethora of parameters[71], such as blood biochemistry, and more, to make accurate health predictions.

Despite King Salmon's critical importance to the global aquaculture sector, the application of machine learning in aquaculture, particularly for King Salmon, remains an underexplored terrain in contemporary research. In particular, there is a noticeable dearth of research employing machine learning techniques to predict their health. This gap is particularly evident in the fact that there have been no studies investigating king salmon using a combination of parameters extracted from blood biochemistry and hematology, body composition, feeding and feed conversion rates, biometric information, growth, sample assessment, histology, and environmental data. Furthermore, the identification of key factors influencing health predictions for these fish remains an unresolved question. Meanwhile, the presence of numerous features poses a challenge: not all of them are equally important for classification. This is where the concept of feature selection becomes indispensable [60]. By pinpointing the most relevant features, we can potentially achieve several objectives: (1) enhance the accuracy and efficiency of the classification model; (2) reduce the computational cost; and (3) provide clear insights into the specific factors that significantly influence King Salmon's health that may potentially reveal the mechanisms that influence fish health. Understanding the key determinants for King Salmon health is not just an academic work; it has real-world implications, aiding breeders in optimizing conditions, and aiding researchers in targeted studies.

1.2 Goal

For this study, we aim to apply machine learning techniques to predict the health status of King Salmon by collecting comprehensive information on King Salmon, with a particular emphasis on feature selection. By successfully identifying these paramount features, we hope to unravel deeper insights into King Salmon biology. Such findings are poised to revolutionize salmon farming and conservation strategies, emphasizing the real-world relevance of this study. Our primary objectives are listed below.

(1) To comprehensively understand and describe the data collection methods, clarify the criteria for determining King Salmon health, implement preprocessing techniques, and provide a clear overview of the dataset's structure. This will set the foundation for the subsequent stages of analysis, ensuring the data is accurate and primed for detailed examination.

(2) To employ statistical methods to investigate the data distributions and identify the features with statistical differences between health conditions. Find correlations through Phi-K analysis[7], and visually represent the data using PCA[84] and t-SNE[16]. This step aims to uncover hidden patterns, relationships, or anomalies that can guide subsequent feature selection and model training processes.

(3) To explore, evaluate, and implement a range of feature selection methods including filter, wrapper, and embedded methods. The objective here is to identify the most relevant features that contribute significantly to model performance, optimizing the predictive models for both accuracy and efficiency.

(4) To compare the features deemed statistically different for healthy and unhealthy conditions with those identified as informative for classifying health conditions by machine learning methods. This comparison aims to discern the alignment or divergence between traditional statistical assessments and machine learning evaluations.

1.3 Contributions

This project makes the following contributions:

(1) This study presents a rigorous methodology for King Salmon data collection, setting a standard that guarantees the precision and trustworthiness of the acquired data. We have provided an exhaustive description of the data collection methods, elucidated the criteria for determining fish health, and implemented advanced preprocessing techniques. By establishing a clear structure of the dataset, this research ensures that all subsequent analyses are grounded on a consistent and solid foundation, primed for further investigation.

(2) This study delves into the King Salmon dataset using various statistical methods. This is the first large-scale investigation of statistically significant differences in the features of king salmon when it comes to their different health conditions. Beyond simply exploring data distributions, we've identified pivotal features exhibiting statistical disparities among health conditions. By employing the advanced Phi-K analysis, we offer a credible view of data interrelationships. Additionally, our illustrative use of PCA and t-SNE plots paints a comprehensive picture of data trends, paving the way for an enlightened feature selection process.

(3) This project addresses the gap in the existing research on the health prediction of King Salmon using machine learning techniques. Our research innovatively investigates data extracted from blood biochemistry and hematology, body composition, feeding metrics, biometric details, growth patterns, sample assessments, histology, and environmental factors. By doing so, we have ventured into this field, laying the groundwork for future research in this domain.

(4) This study shows that we have systematically explored and evaluated a comprehensive range of feature selection methods, including the filter, wrapper, and embedded techniques. This exhaustive approach ensures that our model incorporates only the most pivotal features, while maintaining its predictive performance. Our findings in this domain hold the potential to revolutionize the way machine learning is applied to marine biology datasets, especially on King Salmon.

(5) This research makes a contribution by juxtaposing features identified through traditional statistical tests with those deemed critical by machine learning algorithms. Such a comparison provides invaluable insights, shedding light on the alignment or discrepancies between conventional methods and advanced machine learning techniques.

Please note that this project is an applied research that aims to leverage both statistical methods and machine learning techniques for solving real-world application tasks of King Salmon fish health issues. The King Salmon under study come from the Marlborough Sound, with trials conducted by the Cawthron Institute in New Zealand. The project faced challenges due to delayed data arrivals and multiple times of data modifications. Although navigating these uncertainties is demanding, we manage to surmount these obstacles and achieve commendable results through diligent efforts.

1.4 Structure

This chapter provides a brief overview of the project. The remaining chapters of the report are organized as follows.

- Chapter 2 summarizes existing and related research, with different sections demonstrating different aspects, like statistical methods, machine learning and classification, and feature selection.
- Chapter 3 contains a detailed description of the data and the preprocessing steps.
- Chapter 4 presents the results of the investigation of the datasets using statistical methods and visualization.
- Chapter 5 discusses the model performance of the feature selection methods used. As well as comparisons between statistical methods, and machine learning with feature selection methods.
- Chapter 6 concludes the project and identifies future work directions.

Chapter 2

Background and Related Work

This chapter delves into the foundational elements and previous studies underpinning the broader scope of our research. Initially, we sketch the backdrop of New Zealand's aquaculture landscape, emphasizing the prominence of the King Salmon industry. As we transition from general overviews, the chapter illuminates the specific statistical methodologies previously employed for King Salmon analysis. Furthermore, an exploration of machine learning's role, specifically in classification and feature selection, is undertaken. Lastly, the related work section serves to bridge the gap between existing knowledge and current research avenues, detailing influential factors on King Salmon's health, statistical techniques, and the interplay of machine learning in this domain.

2.1 Background

2.1.1 New Zealand Aquaculture and King Salmon Industry

The aquaculture in New Zealand primarily revolves around the production of mussels, Chinook salmon, and Pacific oysters. With an eye on expansion, the industry aspires to augment its export revenues to 3 billion New Zealand Dollars by the year 2035 [78]. In order to realize this ambition and uphold its reputation for being 'clean and green', it is imperative for the industry to ensure the productivity and the sustainability of its practices. Among them, the Chinook salmon(*Oncorhynchus tschawytscha*), commonly known as King salmon, holds a distinguished position in New Zealand as being the only type of salmon cultivated within the New Zealand region, contributing to more than half of the worldwide production of this species [58]. King salmon embodies an unparalleled source of high-grade protein and long-chain omega-3 fatty acids play a crucial role in health enhancement [59]. Besides, New Zealand farmed salmon has a lower carbon footprint compared to the global average carbon footprints published in other animal protein studies, which also be an environmentally friendly product [65].

However, research indicates that the health of King Salmon has been considerably compromised, with instances of fish mortality reported globally [90]. Recent findings reveal that the New Zealand King Salmon suffered a post-tax net loss of 73 million, a consequence of farm mortalities [72]. The substantial economic, environmental, and health ramifications inherent to King salmon aquaculture accentuate the necessity for rigorous, multifaceted research exploring the factors influencing the health, and consequently, the productivity of this species. This comprehensive understanding will potentially advance the operational efficiency and yield of King salmon farming, thereby serving the economic objectives tied to this industry.

2.1.2 Statistical Methods for King Salmon

In the field of biology, the application of statistical methods plays a pivotal role in the generation of meaningful insights and the validation of hypotheses[34]. Foundational statistical techniques, such as hypothesis testing and the calculation of p-values and confidence intervals, have long been at the core of biological research, allowing researchers to rigorously evaluate whether observed differences or associations are statistically significant[82].

In previous studies, they have employed statistical tests to examine variables concerning the biology area, including King Salmon[5]. A common methodology is observed across these studies[45]. Initially, all analyzed parameters undergo tests for normality and homogeneity of variance. If the data conforms to the prerequisites for parametric analysis, comparisons between different treatments are conducted using the independent t-test [44]. A significance level of $P < 0.05$ is consistently applied across all statistical tests. In the situation where the data violates the normality assumption, an appropriate non-parametric test, such as the Kruskal-Wallis test[63], is employed. If variances are found to be heterogeneous, results are derived from outputs where the assumption of equal variances is not made. Furthermore, when comparing the model performance of multiple runs, the Wilcoxon rank-sum test is a common method to use [10].

Principal Component Analysis (PCA) [84] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [16] stand as two leading techniques designed to transform high-dimensional data into a two-dimensional format. PCA, a linear approach, works by pinpointing orthogonal axes, known as principal components, that capture the maximum variance in the data. This allows for the projection of data into a lower-dimensional space without significant loss of variance. In contrast, t-SNE, a non-linear technique, prioritizes the maintenance of local data structures from the high-dimensional space during its transition to low dimensions. This makes t-SNE especially proficient at visualizing intricate data clusters.

2.1.3 Classification in Machine Learning

Machine learning, a rapidly evolving area, empowers computational systems with the capability to extract insights from data [39]. This field typically involves the training of models on data with labels, enabling them to predict future outcomes[29]. Engaging with data that possesses labels falls under the domain of supervised learning. In this paradigm, given a new input, the model predicts the most likely class or label for it, a process termed classification. The intricacies arise from the challenges of managing real-world data complexities, the nuances of preprocessing, and the distinct attributes of the problem.

Prior to any machine learning application, it is important to process and prepare the data. This involves integrating disparate data sources, cleaning noise, normalizing features to a standard scale, and transforming categorical variables into a format suitable for algorithms [23]. Data reduction tasks, including instance selection, are pivotal in refining the datasets [46]. Real-world datasets often come with their inherent complexities such as imbalance class distribution, noise, missing values, outliers, and non-standardized features[42]. The issue of imbalanced class distribution, where the number of instances for each class varies significantly, is a prevalent challenge in real-world datasets. For instance, in medical diagnosis datasets, there is often a stark disparity between the number of healthy and unhealthy samples[53]. Addressing these inconsistencies demands a meticulous preprocessing strategy.

Central to the classification application is the classification algorithm, which learns from a training dataset and uses the learned model to categorize new data points into specific classes. The simplest algorithm is k-Nearest Neighbor(KNN) [35]. It is a very simple but highly efficient and effective algorithm for pattern recognition. The instances are classified

based on the class of their nearest neighbor [18]. The Support Vector Machine(SVM), is a well-regarded algorithm for classification tasks that has been widely used in biological application[76]. SVM uses kernel functions, such as the Radial Basis Function, to transform intricate problems into simpler ones so that the similarity between pairs of data points can be computed using a kernel function without the need to explicitly transform the entire data[28].

2.1.4 Feature Selection Methods

Feature selection refers to the method of selecting a subset of features that are most important for specific tasks. In the field of health classification, the identification of important features holds great significance. It serves to identify and retain only features that significantly contribute to the prediction performance of a model. It not only optimizes model training time and reduces the complexity of a model but also assists in preventing the curse of dimensionality and overfitting. The fundamental principle of any feature selection method for classification are same: to identify the most critical and predictive features from the available data, thereby enhancing the accuracy and interpretability of the resulting models. Feature selection techniques can be broadly categorized into three main methods: filter, wrapper, and embedded.

Filter Methods Among the various techniques adopted for feature selection, filter methods such as Relief, Mutual Information, and Chi-square have gained prominence due to their capability to pinpoint significant features related to labels for classification tasks. The filter methods mainly focus on the properties of data and are independent of any learning methodology. Thus, they are generally not computationally expensive and have a good generalization ability according to Chandrashekar and Sahin(2014)[13].

- ReliefF Method: The ReliefF[80] fundamentally ranks features based on their ability to differentiate between instances in close proximity. It estimates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instances of the same and different classes. The score it assigns to each feature is a measure of how well the feature distinguishes between classes.
- Chi-Squared Method: In the chi-squared method[36], the score for each feature is calculated based on the chi-squared statistic. This score measures the independence of each feature from the output class. High scores mean the feature and the output are highly dependent, which typically indicates a valuable feature.
- Mutual Information Method: Mutual information measures the information that features and labels share[47]. It measures how much knowing one of these variables reduces uncertainty about the other. High mutual information between a feature and the output class suggests that the feature is significant.

Wrapper Methods A wrapper method adopts a more comprehensive approach to feature selection that assesses subsets of features by training specific models during the feature selection process, using the model's performance as a selection criterion. The recursive feature elimination(RFE) [27] is a typical wrapper feature initially used for gene selection. The primary advantage of wrapper methods is their ability to account for feature interactions, often leading to better performance. They are more computationally expensive than filter methods but often provide better results since they consider the interaction between features. However, they also run the risk of overfitting to the chosen model. The Recursive Feature

Elimination technique, when paired with algorithms such as Logistic Regression(LR)[51], Linear Support Vector Classification(LSVC)[32], and Random Forest (RF)[75], exemplifies this approach.

Embedded Methods Embedded techniques integrate feature subset selection within classifier construction. Similar to wrapper methods, embedded methods are also specific to a particular learning algorithm. It considers interactions with the classification model and is, therefore, less computationally demanding than wrapper methods[70]. Besides, embedded methods determine the importance of each feature while the model is being trained[48]. They often strike a balance between filter and wrapper methods, considering both feature relevance and model performance, and are often more efficient than wrapper methods.

A prominent example of this approach can be observed in certain algorithms in the wrapper method: LR, LSVC, and RF. These algorithms inherently possess mechanisms to rank or weigh features based on their importance or contribution to the model. For instance[49], LR and LSVC, when regularized using L1 regularization, tend to push coefficients of less important features toward zero, effectively selecting a subset of features. Similarly, RF[40] provides a feature importance score based on the average decrease in impurity brought by a feature across all trees. However, the features selected by one algorithm might not necessarily be optimal for another. This specificity stems from the fact that each algorithm has its own criteria for evaluating feature importance.

2.2 Related Work

2.2.1 Influential Factors on the Health of King Salmon

The health of King Salmon is a multifaceted parameter, influenced by numerous factors that may present tiny or pronounced changes in the fish. The choice of which factors to include in the dataset, and thus included in a machine learning model, especially when making predictions about the health of the fish, becomes crucial. In this section, we provide an overview of the research specifically conducted on King Salmon and delve into a discussion on several aspects of the features. Drawing from the existing literature shown below that utilizes King Salmon data from New Zealand, we highlight the identified factors that either influence or are influenced by the health status of King Salmon.

Genetic factors significantly influence the phenotypic traits of king salmon, encompassing variations in growth rates, size, and overall health among individuals. For example, the pursuit of breeding for a shorter and deeper body shape in organisms could heighten the vulnerability to spinal curvature [74]. The abnormal shape of the spine is included in the Sample assessments collection in the datasets used. Moreover, the administration of antibiotics and probiotics exerts an influence on the composition of the gut microbiome, consequently inducing alterations in gene expression and potentially impacting the health status of Chinook Salmon[69]. Certain genes have demonstrated the capability to enhance the conversion of protein from feed into body tissue with heightened efficiency, which was achieved through the augmented expression of the proteasome, lipid, and carbon metabolic pathways in the liver[21].

Feeding habits significantly impact the physiological development and growth of King Salmon, making it vital to evaluate the effects of factors like the nutritional composition and feeding rations on growth rates to optimize feeding practices. In the context of the current investigation[5], the estimated maintenance demands for protein in King salmon were found to be nearly two times higher than the corresponding estimate reported for Atlantic salmon, both species being subjected to the same water temperature conditions.

Nevertheless, the fish food production process has inherent limitations, necessitating the addition of starch, which adversely impacts protein utilization in carnivorous fish such as King Salmon[26].

Moreover, a recent study[5] demonstrates that Chinook salmon reared under **satiation feeding conditions** exhibited a significantly higher final body weight at harvest, whereas those subjected to a restricted ration showed a potential reduction in the incidence of spinal anomalies among farmed individuals. The feed conversion ratio (FCR) is the ratio of feed intake (FI) to body weight gain, which worsened as the fish increased in size[74]. That is the undesired outcome to avoid as feed-efficient fish consume less share of meals and maintain superior growth rates[20]. Specifically, feeding juvenile King Salmonto satiation did not yield the anticipated outcomes of efficient, rapid, and consistent growth, both at individual and cohort levels[37].

The **rearing conditions** employed in aquaculture settings can substantially affect King Salmon development. Factors such as water temperature can influence overall health. In the studies [20, 77], it was observed that both growth and feed intake increased with temperature and exhibited notable correlations with FCR efficiency. Besides, the study also showed that the correlation and alteration of fish physiology, health, composition, and gut microbiota were notably influenced by water temperature[77]. The related information can be found in multiple datasets such as the Growth collection.

Furthermore, **fish diseases** represent a significant threat to global aquaculture, resulting in considerable production losses annually [6]. A recent study conducted on King Salmon has revealed a noteworthy prevalence of Aeromonas spp. in individuals exhibiting fluid accumulation in their swim bladders. This finding is of particular concern due to the well-known association of various Aeromonas species with fatal outcomes, leading to substantial financial losses and increased expenses in the fish production sector within the aquaculture industry [91]. Furthermore, For instance, chronic elevation of corticosteroids in King Salmon has been suggested to increase the presence of pathogens, consequently impacting their health condition [17].

2.2.2 Statistical Methods for King Salmon

In recent research, a variety of statistical tests have been utilized to analyze the data of King Salmon. During data collection, prior studies have examined the presence of significant differences between two capture methods for sampling. This was achieved using either an analysis of variance (ANOVA) approach [67] or the Kruskal-Wallis rank sum test [11]. When analyzing experimental results, a study investigating the impact of varying rations on the spinal anomalies of King Salmon employed a similar methodology as described above. This encompassed preliminary assumption tests followed by tests to discern group differences [5]. Additionally, the ANOVA and Tukey's HSD test [56] were harnessed to contrast phenotypes across diverse traits of King Salmon. From these studies, it becomes evident that employing statistical tests to discern differences between conditions is a standard practice in the field. However, it is noteworthy that none of the prior research has endeavored to comprehensively apply statistical tests to King Salmon measurements for the purpose of determining key features. Furthermore, there has been a lack of in-depth analysis concerning potential health effects, indicating a gap in the existing literature that warrants further exploration.

The area of biology has long recognized the invaluable contributions of visualization techniques such as PCA and t-SNE in deciphering intricate datasets. In the context of Nutritional Epidemiology, PCA stands as a cornerstone, elucidating patterns and reducing dimensionality while preserving as much variance as possible. The comprehensive method-

ology documented by Santos et al. [73] is a testament to the robustness and applicability of PCA in this domain. On the other hand, t-SNE has carved a niche for itself in transcriptional analyses. As elucidated by Cieslak et al. [16], this technique adeptly maps high-dimensional transcriptional states into a lower-dimensional space, highlighting clusters and underlying structures. Such insights gleaned from t-SNE can be pivotal, offering a clear lens to discern relationships and aiding in subsequent analytical processes.

2.2.3 Classification in Machine Learning for King Salmon

Under the machine learning context, the core problem we aim to address in this study is a binary classification question: distinguishing King Salmon between healthy and unhealthy conditions. However, this task is quite challenging as the intricacies of our problem arise from the challenges of managing real-world data complexities, the nuances of preprocessing, and the distinct biological attributes of the species.A recent study employed a combination of King Salmon features to predict feed efficiency[88]. However, this study did not focus on health prediction and did not consider the factors and intricacies associated with the health status of the species, underscoring the novelty and significance of our research.

Biological datasets, especially those derived from real-world systems, are inherently complex. They often exhibit characteristics such as noise, missing values, outliers, and non-standardized features. Addressing these challenges necessitates a rigorous preprocessing approach. For instance, taking into account the phenomenon of reproductive death is crucial during data cleaning [25]. Regarding imputation techniques, a study involving 12 medical datasets indicated that KNN interpolation might be the most effective method for handling missing values, outperforming deletion, mean interpolation, and median interpolation in terms of error rate [2].

In terms of classification algorithms, the SVM, particularly with a radial-based kernel, has demonstrated superior performance. It emerged as the best classifier when compared to Random Forest, Logistic Regression, and KNN in terms of correct classification rate[68]. Besides, SVM has consistently exhibited state-of-the-art results in various diagnostic tasks involving biological data[1]. Moreover, previous research has concluded that the SVM classifier is predominantly employed across the majority of the studies they analyzed[3]. Additionally, the F-measure like the F1 score is recognized as a more robust evaluation metric compared to accuracy for considering both precision and recall, especially when dealing with imbalanced datasets [38].

2.2.4 Feature Selection

Due to the lack of research specifically pertaining to King salmon, for this section, we mainly reviewed studies related to biology and medical application, aligning with the research strategies adopted in previous scholarly works [25]. This approach is predicated on the assumption that not only the physiological and biological processes in King Salmon may exhibit similarities with other species, but also similar data structures, thereby rendering the application of these feature selection methods both relevant and appropriate in the context of the present study.

Filter Methods Within the discourse on feature selection for health-focused investigations, filter methods such as Relief, Mutual Information, and Chi-square are frequently employed by researchers. These methodologies are widely used for their capability to identify features that are instrumental for classification tasks. Urbanowicz et al. (2018)[80] proffered an

in-depth analysis of Relief-based algorithms, emphasizing the increasing significance of feature selection in biomedical data mining. This attention stems from the mounting challenges posed by escalating feature dimensionality, coupled with the demand for computationally streamlined methodologies capable of modeling intricate associations. On a similar note, Dissanayake and Johar (2021) [19] harnessed an array of feature selection techniques, such as Chi-square, mutual information, and ReliefF, to process the Cleveland heart disease dataset. Their results affirmed the efficacy of these methods in enhancing heart disease prediction. Building on this paradigm, Huda et al. (2017) [33] unveiled a novel approach, utilizing the filter method, tailored to tackle the intricate dynamics of imbalanced medical data, particularly in brain tumor diagnoses. This underscores the adaptability and robustness of filter methods.

Wrapper Methods To date, rare studies have employed the REF wrapper feature selection method specifically on King Salmon health analysis. Although the REF method has been recognized as a potent tool in the field of biology. Its efficacy has been demonstrated in diagnosing diseases such as Alzheimer's and cancer, as evidenced by recent studies [66, 85]. In a notable instance concerning the prediction of Parkinson's Disease Depression, the REFCV method reduced the feature number from 35 to 10, achieving even greater accuracy in the process [61]. The growing utilization of the REF method in pivotal medical research domains underscores its credibility and precision.

Embedded Methods While there hasn't been any previous research specifically employing embedded feature selection methods on King Salmon data, the use of embedded methods is pervasive in the broader field of biology. Numerous studies have highlighted the efficacy and robustness of embedded techniques in extracting relevant features from biological datasets. For instance, Kang et al. [41] demonstrated the utility of embedded methods in identifying significant genetic markers, while Khanji et al. [43] employed LASSO, an embedded method, to effectively select features in cardiovascular studies. The success of these methods in diverse biological contexts suggests their potential applicability and value in analyzing King Salmon data.

2.3 Chapter Summary

Existing literature prominently demonstrates a reliance on conventional statistical tests to understand relationships between specific features within King Salmon health datasets. While these approaches have their advantages, there's a marked deficit in studies that exploit comprehensive data preprocessing methods, exploratory data analysis, and advanced feature selection techniques tailored to fish health. This shortcoming emphasizes a significant void in contemporary research, underscoring the need for studies that utilize the full scope of data-driven methodologies to address the intricate challenges of fish health.

Chapter 3

Data and Preprocessing

In this chapter, the content is segmented into four primary sections. Initially, the materials and methods deployed for data collection and analysis are elucidated, laying the groundwork for the research. This is followed by a detailed exploration of the criteria used to evaluate the health of fish, ensuring a comprehensive understanding of the health metrics. The subsequent section delves deep into the preprocessing techniques applied, emphasizing the importance of refining raw data to derive meaningful insights. Lastly, detailed information on the datasets used in the study, including their source, composition, and relevance, is presented. Throughout the chapter, the meticulous processes and considerations adopted to ensure data reliability and relevance in assessing fish health are emphasized. Of particular note is the initial state of the data, which is primarily raw and sourced directly recorded in the laboratory. This real-world data is characterized by its inherent inconsistencies and rich, unfiltered information that needs a detailed process.

3.1 Materials and Data Collection

All King Salmon used in this study were procured from the commercial hatchery, Sanford's Kaitangata, and subsequently reared in freshwater by Salmon Smolt New Zealand, located in Kaiapoi. Following this phase, the King Salmon were transferred to the Finfish Research Centre (FRC) at Cawthron Aquaculture Park (CAP) in New Zealand.

The primary objective of this research, conducted by the Cawthron Institute, was to gather comprehensive data on various aspects of the King Salmon. These aspects include:

- Blood biochemistry and hematology
- Body chemical composition
- Feeding and feed conversion ratio (FCR)
- Biometrics
- Growth
- Sample assessments
- Histological evaluations
- Trial conditions
- Health classifications

Trial	Event	Salinity	Ration(s) at the event	Temperature(s) at the event (°C)	Start date	End Date	Comments
1	Arrival in the FRC				21-Aug-18	21-Aug-18	
	WT2	FW	100	15	11-Sep-18	14-Sep-18	Assessment before temperature change
	WT4	FW	60,80,100	13,17	15-Oct-16	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	WT7-WT10 = 17 °C only
	WT14	FW	60,80,100	17	12-Mar-19	28-Mar-19	WT10-WT14 = 17 °C only
2	Arrival in the FRC				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	WT9-WT11 = temperature challenge and controls
3	Arrival in the FRC		NA		06-May-20	25-May-20	
	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	End of 100 % ration
	WT6	FW	25	8,12,16,20	26-Aug-20	08-Sep-20	WT5-WT6 = 25 % ration
	WT7	FW	25	8,12,16,20	16-Sep-20	29-Sep-20	WT6-WT7 = 25 % ration
	WT8	FW	0	8,12,16,20	14-Oct-20	28-Oct-20	WT7-WT8 = fasting (0 % ration)

Figure 3.1: FRC trial information and details for each event in three trials.

Each aspect represents a distinct data collection, with several observables assessing different facets within each collection. To illustrate, the growth collection includes three observables: fork length, girth, and weight. Meanwhile, the trial information collection provides insights into the aquaculture tank environment for all sampled fish, detailing parameters like temperature (in °C) and feeding satiation ration (e.g., a 0 satiation ration indicates fasting treatment). Observables in the health classification category denote the binary health status of the fish, either in its entirety body or of specific body parts, classified as ”healthy” or ”unhealthy”. It's worth noting that the comments collection, comprised mainly of experimenter remarks, was not utilized here. The comments collection predominantly comprises textual annotations provided by various experimenters, and the majority of information embedded within these comments is already incorporated and represented in the primary dataset discussed earlier.

The experimental design was structured into three distinct trials. Each trial consisted of a series of experimental events spaced at varied intervals, during which different observables and environmental parameters were assessed. Figure 3.1 provides a detailed overview of the conditions for each event across all trials. In terms of salinity references, 'FW' indicates freshwater, while 'SW' stands for seawater. Moreover, the 'Ration' value designates the percentage of the satiation ration.

3.2 Fish Health Criteria

The overarching health status of the fish serves as the primary indicator in this study, grounded in criteria established by researchers from the Cawthron Institute. As mentioned above, the observable `general_health_classification` in health classification collection informs the final health classification for a given fish (healthy or unhealthy). This is the final health classification for the fish whose health was accessed, which was determined based on factors encompassing growth performance, general health assessments conducted during the tri-

als, necropsy observations, and haematology appearance. For fish deemed unhealthy, they did not satisfy one or more of the criteria delineated in Table 3.1.

Over time, the Cawthron Institute has undergone several modifications in this criteria. In collaboration with these modifications, we collectively reviewed and adjusted these criteria, ultimately opting to emphasize those anchored on blood-related features. If one fish has multiple records, we choose to label the fish using the record having blood data. This decision was made considering such criteria furnished the most comprehensive and illuminating information pertaining to the fish's health. This cooperative approach ensured the research's robustness and alignment with contemporary standards. However, this decision might change as more investigation progresses.

The Condition Factor(CF) in the Table 3.1 was calculated with the following equation:

$$CF = \frac{w \times 100,000}{L}$$

where CF is Fulton's condition factor (mm) [22], w is the weight (g) and L is the fork length.

The observables described in Table 3.1, such as Condition Factor and Leucocyte appearance, have been omitted from the datasets when training the machine learning models as they directly classify fish as healthy or not. However, derived features like 'Weight loss', as well as the foundational features used in their computation, such as the weight utilized to determine weight loss, are retained.

3.3 Preprocessing

Each trial is organized into distinct collections, with each collection forming an individual dataset. In this research, primary attention is centered on the following datasets: blood biochemistry and hematology (referred to as "blood"), body chemical composition ("composition"), feeding and feed conversion ratio ("FCR"), biometrics, growth, sample assessments ("assessments"), and histology.

1. Feature Engineering One of the critical stages in this analysis is the meticulous formation of explicit features for the datasets. Certain datasets contain not just observables but also specific information like which body part is analyzed. Take the histology dataset as an example: it would be an oversimplification to merely consider 'inflammation' as a distinct feature. Rather, a more nuanced approach would recognize 'inflammation of the heart' as a unique feature. This is particularly pertinent because the dataset evaluates the inflammation scores of various other organs such as the liver, stomach, and so forth. Hence, combining the observable with its associated body part offers a more granular and accurate representation of the data, ensuring more robust analytical outcomes. Upon completion, we obtain structured tabular datasets where each row signifies an individual fish record, and every column corresponds to a distinct feature or label.

2. Data Integration When endeavoring to amalgamate columns from one collection to another, it is imperative to utilize both the **fish ID and the event** concurrently, treating this combination as a unique identifier. The event could be considered as a timestamp that represents different experiments. This procedural necessity arises due to the existence of scenarios where a single fish might be assessed during different events. Consequently, multiple records might exist for a singular fish, each documenting varying environmental conditions and feeding practices to which the fish was subjected. Thus, it may have different values for the same feature in different events. Hence, the amalgamation process must be executed

Criteria	Collection	Details
Weight loss or abnormal CF	growth measurement	(1) Weight loss; (2) Low Condition Factor, exclude if Condition Factor < 1.1 (except at tagging when there is no lower limit)
haematology appearance	blood analyses	Leucocyte appearance, Erythrocyte appearance, Thrombocyte appearance not normal
Abnormal white cells	blood analyses	(1) Reduced % lymphocyte $< 87\%$; (2) Increased % neutrophils $> 10\%$; (3) Increased % of monocytes $> 2\%$
Abnormal stomach, swim bladder	health assessment	(1) Abnormal stomach = Y – based on visual assessment; (2) Abnormal volume of swim bladder fluid: • weight $<$ 500 g, abnormal volume $>$ 1 mL; • weight $>$ 500 g, abnormal volume $>$ 2 mL (3) Abnormal stomach width: • weight $<$ 500 g, stomach width $>$ 20 mm; • weight$>$ 500 g, stomach width $>$ 35 mm
Abnormal kidney, liver, faeces	health assessment	(1) Kidney - nephrocalcinosis score ≥ 3; (2) High faecal appearance score, 3 and above; (3) Low liver index: < 0.75 Low condition factor, exclude if CF < 1.1 (except at tagging when there is no lower limit)
Abnormal histology scores	histology analyses	High total histology score based on sum of all individual tissues > 12
High inflammation	histology analyses	(1) High GI tract inflammation score > 5; (2) High histology inflammation score > 10
Abnormal spinal curvature	health assessment	Presence of spinal curvature (moderate or severe) or present (if visual) Note: We only consider this a health sign if the fish has other issues.
Abnormalities(comments)	comments	Based on comments at sampling or during assessments

Table 3.1: Health Criteria for King Salmon

Trial	Tag	WT2	WT3	WT4	WT5	WT6	WT7	WT8	WT9	WT10	WT11	WT14
Trial 1	0	-	-	2	-	-	3	-	-	4	-	5
Trial 2	6	7	8	9	10	11	12	-	13	-	14	-
Trial 3	15	16	17	18	19	20	21	22	-	-	-	-

Table 3.2: Event values across different trials.

with meticulous attention to these nuances to ensure the integrity and accuracy of the consolidated data.

Considering that both the environmental condition and feeding satiation ration influence every sample at all events, it is imperative to amalgamate these two factors from the trial conditions collection into the other collection datasets, with the exception of the health classification dataset. Furthermore, the integration of overall health status into the preceding datasets is requisite to create a new dataset. This newly constructed dataset will only preserve fish with a health condition label.

3. Data Cleaning The next step in the pre-processing of the constructed data is to discard any columns in the data that contain no values, thereby reducing redundancy and saving computational resources. Additionally, the ID column is dropped from the features as it probably represents identifiers that do not contribute to the model's learning.

Upon conducting a meticulous examination of the data, an anomaly was identified within the biometrics collection: three fish were labeled as male. This finding is anomalous, particularly considering the unique reproductive characteristics of king salmon, which undergo a phenomenon known as 'Reproductive Suicide" as mentioned in the previous chapter. King Salmon, belonging to the Pacific salmon genus, exhibit a biological phenomenon wherein their immune defenses are notably compromised upon reaching reproductive maturity, resulting in post-reproductive mortality [25, 81]. Consequently, given that male fish die shortly after reaching maturity and females do not die unless they mate, any instances labeled as 'male' within the datasets should be regarded as noise. Given this biological reality, the presence of male fish within the dataset that does not align with expected post-reproductive mortality patterns raises questions about the accuracy and reliability of these instances. Thus, these records are removed from the datasets.

4. Converting Non-Numerics The next step is to convert non-numeric feature values into numeric counterparts as machine learning algorithms predominantly operate on numeric values. A recurring non-numeric feature across datasets is the "event" feature. The mappings for this feature are delineated in Table 3.2. Specifically, the "tag" denotes king salmon cultivated on the farm, while events prefixed with "WT" signify the sampling periods, indicating that the king salmon were raised at the Cawthron Aquaculture Park.

In terms of labeling, unhealthy king salmon are represented by the value 0, whereas healthy king salmon are denoted by the value 1. Post-transformation, every feature is numerically represented. The categorization of a feature as either continuous or discrete is predicated upon its intrinsic significance.

5. Incorporating New Blood Features As advised by the Cawthron Institute, three new features are incorporated into the Blood collection for all three trials. However, due to the absence of the functional feature, Hematocrit, in the Trial 3 dataset, only the first two features were integrated. The three new features are computed as follows:

- **Albumin:globulin ratio:**

$$\frac{Blood\_plasma\_Albumin(g/L)}{Blood\_plasma\_Globulin(g/L)}$$

- **Neutrophil:lymphocyte ratio:**

$$\frac{Blood\_Haematology\_NeutrophilsAbs(10^9/L)}{Blood\_Haematology\_LymphocytesAbs(10^9/L)}$$

- **Mean corpuscular haemoglobin concentration(meanCorHaeCon):**

$$\left(\frac{Blood\_Haemoglobin(g/L)}{10} \times 100\right) \div Hematocrit\%$$

When calculating new values that are derived from the ratio of two variables, the presence of missing data in either the numerator or the denominator necessitates careful handling to ensure the validity of the resulting value. Specifically, if a variable in the numerator or denominator is missing, the calculated value should be designated as missing or empty as well. Furthermore, in cases where the denominator is zero, the calculated value will be designated as missing.

6. Data Splitting and Normalization Then data is split into a training set and a test set. The split is stratified, ensuring that each set contains a representative distribution of the target variable classes. The training set contained 80% of instances and the test set contained 20%. This approach ensures that the test set is sufficiently representative of the minority class, thereby making the evaluation metrics more reliable. After splitting the data, the train and test data are then scaled, which is done by scaling each feature to a given range [0, 1]. It achieves this by subtracting the minimum value in the feature and then dividing by the range of the feature like Equation 3.1. This step is to bring all features to a similar scale. This improves the performance of some machine learning models and ensures fair weight across all features.

$$X_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

7. Handling Missing Data Addressing missing values is imperative for robust analysis. The presence of these gaps can be attributed, in part, to the intricacies of the eigenvalues. For instance, the Feed Conversion Ratio (FCR) is an interval measure necessitating two columns for capturing associated events - the start event and the end event. Conversely, the rest features demand an extra column to document their measurement events. This discrepancy causes a null value in the FCR's dual-event columns whenever another feature is populated. To mitigate such disparities, we segregated the instances with FCR into an independent dataset named feeding-FCR. The rest instances formed another dataset named feeding-meal for the main feature is `share_of_meal`, which is the daily food intake of an individual divided by the sum of the whole tank.

Additionally, the disparity in event timings for different measurements further exacerbates the missing data challenge. Specifically, some measurements, which entail fish dissection, render subsequent feature measurements infeasible. Compounding this, feature reconstruction, as detailed earlier, also contributes to missing data. A salient example can be found in the "assessments" collection of Trial 1, where the fluid volume in the swim bladder was cataloged using two distinct units: percentage and milliliters (mL). Consequently, if

	Trial1			Trial2			Trial3		
Dataset	**Rows**	**Columns**	**Ratio**	**Rows**	**Columns**	**Ratio**	**Rows**	**Columns**	**Ratio**
blood	213	39	0.82	459	40	3.54	392	38	0.77
composition	104	117	1.54	277	117	4.65	392	128	0.77
Feeding-FCR	208	6	0.81	126	6	1.74	118	6	0.69
Feeding-ShareMeals	212	5	0.83	209	5	2.03	317	5	0.7
growth	213	6	0.52	459	5	3.54	392	6	0.77
assessment	213	8	0.82	459	15	3.54	392	12	0.77
histology	213	37	0.82	459	37	3.54	387	37	0.74
biometrics	213	17	0.82	458	16	3.58	392	10	0.77

Table 3.3: Sizes of datasets represented by the number of samples and features, and the imbalance ratio of unhealthy class and healthy class for different trials.

the volume is recorded as a percentage, its equivalent in mL remains unrecorded. To manage these missing values and ensure data integrity, we employed the K-nearest neighbor algorithm, setting k at 5, to extrapolate and fill the missing values effectively. The K-Nearest Neighbors imputation strategy is then applied to fill any remaining missing values in the data based on the 5 nearest neighbors, measured using Euclidean distance as the distance metric. This process helps to retain valuable information that would otherwise be lost if rows with missing data were dropped. We have 8 different datasets and some of the features have a high percentage of missing values. If we delete the features or all instances contain missing values, there is a high likelihood that it will compromise the integrity of the entire dataset.

Lastly, any feature in the dataset that has zero variance (i.e., all instances have the same value) is dropped. These features provide no discriminating information for the learning model and hence can be disregarded.

3.4 Datasets Information

The basic information of the datasets after pre-processing is shown in Table 3.3. It's evident that the datasets differ notably across the three trials in terms of size and imbalance ratios(represented by the "Ratio" column). Taking the blood dataset in Trial 1 as an example, the dataset comprises 213 fish samples and 39 features. The class imbalance ratio is computed by dividing the number of unhealthy fish by the number of healthy fish for each trial. In trials 1 and 3, the number of unhealthy king salmon is generally fewer than the healthy ones, with the exception of the composition dataset in Trial 1. In contrast, Trial 2 predominantly features a greater number of unhealthy king salmon compared to their healthy counterparts.

3.5 Chapter Summary

In Chapter 3, we delved deeply into the data and its preprocessing, beginning with the materials and methods adopted to collect the data, followed by a comprehensive overview of the fish health criteria. The chapter further emphasized the preprocessing techniques implemented to refine the data and presented detailed information on the datasets' structure. These foundational steps ensure that the data is primed for rigorous analysis, which will be extensively covered in the subsequent Chapter 4, encompassing exploratory data analysis methodologies and visualizations.

Chapter 4

Exploratory Data Analysis

In this chapter, we embark on a systematic exploration to sift through the King Salmon dataset. Our goal is to reveal underlying structures, patterns, and potential anomalies. This initial deep dive serves a dual purpose: first, to provide a qualitative grasp of the data's characteristics, and second, to inform subsequent analytical stages by highlighting any pertinent trends or disparities. EDA employs a suite of visual and quantitative techniques, ranging from descriptive statistics to intricate data visualizations.

4.1 Data Distributions

As discussed in Chapter 2, we use statistical methods to find the features that have differences between the two health groups. The Shapiro-Wilk test is used to check the normality of the distribution. When the p-value > 0.05, it implies that the distribution of the data is not significantly different from the normal distribution. In other words, we can assume the normality. Levene's test is an inferential statistic used to check if the variances of a variable obtained for two groups are equal or not when data comes from a non-normal distribution. It tests the null hypothesis that the population variances are equal or not, It is known as homoscedasticity.

In this section, we use the blood collection dataset as a representative example to showcase the results. Table 4.1 elucidates the results of the initial blood collection trial. This table systematically presents the variables, their respective p-values for normality and homogeneity, and a categorical determination of whether they meet the criteria for normality and homogeneity. When the data adhered to a normal distribution (as evidenced by the Shapiro-Wilk test of normality with $p > 0.05$) and exhibited homogeneity of variance (confirmed by a variance chi-squared test with $p > 0.05$), a t-test was employed for data analysis. Like the *alkaline_phosphatase* in the Table 4.1 below. Conversely, when the data were normally distributed (as indicated by a normality test with $p > 0.05$) but demonstrated heterogeneous variance (as shown by a variance chi-squared test with $p \leq 0.05$), Welch's t-test, a non-isotropic variant, was utilized. For data that did not conform to a normal distribution (as determined by a normality test with $p \leq 0.05$), the Wilcoxon rank-sum test was applied as a non-parametric alternative.

The significance of these features underscores their potential relevance in differentiating between health groups in the trial. There are 18 features that show statistically significant differences between the two health conditions in the blood collection of Trial 1. The same process was reiterated to evaluate the features across all datasets. The objective was to discern the number of features that exhibited statistically significant differences between health groups across different trials. Such an analysis is pivotal in understanding the consistency

variable	normality_pvalue	is_normal	homogeneity_pvalue	is_homogeneous
alanine_aminotransferase	0	No	0.1583	Yes
albumin	0	No	0.73	Yes
alkaline_phosphatase	0.3078	Yes	0.916	Yes
aspartate_aminotransferase	0	No	0.0259	No
c-reactive_protein	0	No	0.9043	Yes
calcium	0	No	0.8361	Yes
chloride	0	No	0.0189	No
cholesterol	0	No	0.0029	No
colour	0	No	0	No
cortisol	0	No	0.1475	Yes

Table 4.1: The assumptions check of the blood collection for Trial 1(part).

Collection	Trial 1	Trial 2	Trial 3
blood	18	21	10
composition	40	11	8
FCR	3	0	0
SharedMeals	1	2	1
growth	4	1	2
assessment	3	2	2
histology	7	13	14
biometrics	6	4	2

Table 4.2: The number of features that are statistically different for two health condition groups on all datasets.

and variability of the data across different collection points and trials.

Table 4.2 provides a comprehensive overview of the number of significant features across three trials for various data collections. From the table, it is evident that the 'composition' collection in Trial 1 has the highest number of significant features, amounting to 40. In contrast, the 'FCR' collection in Trial 2 did not yield any significant features. Such variations across trials and collections underscore the complex nature of the data and the potential influences of external factors on the results.

Furthermore, a comparative analysis of the trials reveals that Trial 1 generally has a higher number of significant features across most collections, with the exception of histology, where Trial 3 leads with 14 significant features. This observation suggests that while Trial 1 might have had conditions conducive to yielding a higher number of significant results for most collections, the histology data in Trial 3 is distinct.

Based on the tests conducted, some features have emerged as significant. These features can serve as a foundation for deeper analysis. For instance, in the blood collection of Trial 1, the significant features can be classified under categories such as Enzymes, Blood Biochemistry, Blood Cell Related, and Others. These classifications are detailed in Table 4.3. The prominence of these features may highlight their potential utility in distinguishing between the health groups within the trial.

We take a deeper analysis of the enzymes and blood biochemistry categories. From Figure 4.1, the first two subplots are the alkaline phosphatase and creatine phosphokinase, which were relatively downregulated in the Unhealthy group. Alkaline phosphatase is widely found in the liver, bones, and bile ducts, and its downregulation could mean low liver or bone function [62]. Creatine kinase is primarily associated with muscle damage, and its downregulation may imply less muscle activity. In on latest study on fish, creatine has been shown to stimulate muscle growth and increase body mass [86]. Additionally, it has the potential to improve feed utilization, especially in the context of plant-based diets.

Enzymes	alkaline_phosphatase, creatine_phosphokinase
Blood Biochemistry	chloride, potassium, cholesterol, urea, c-reactive_protein, cortisol, prostaglandin_e2
Blood Cell Related	haematocrit, lymphocytes_abs, monocytes_abs, white_blood_cell_count, neutrophil_lymphocyte_ratio, thick_buffy_coat
Other	event, temperature_celsius, satiation_ration

Table 4.3: Features with significant differences between health groups in Trial 1 for health classification.

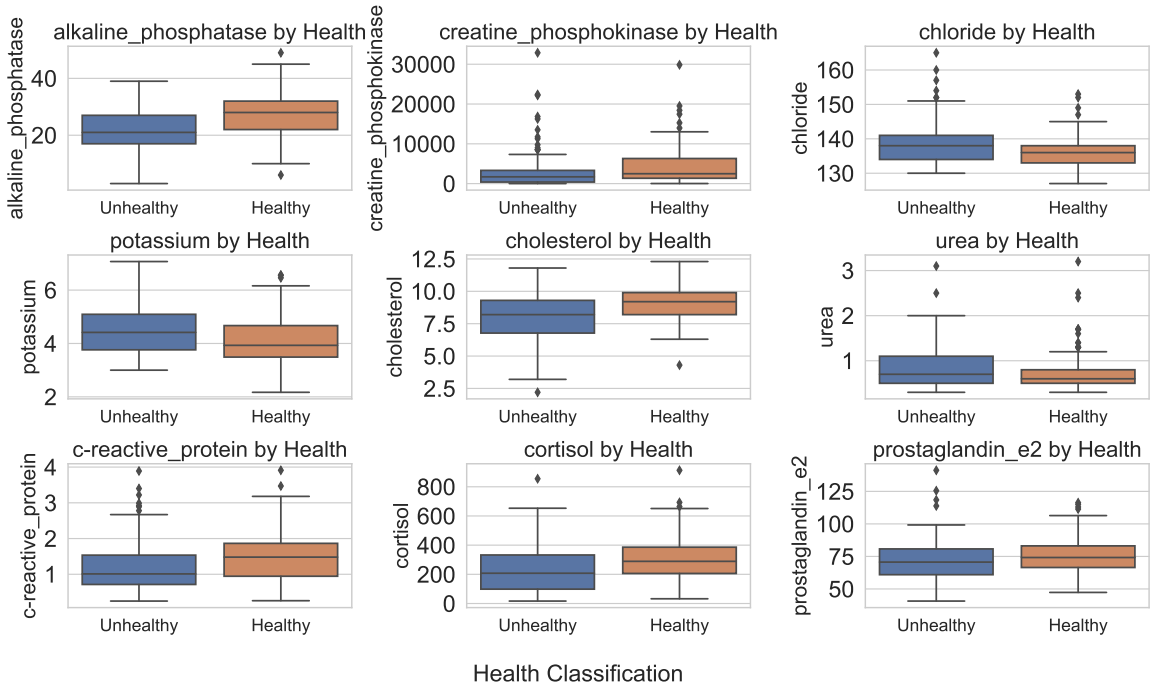

Figure 4.1: The boxplot of Enzymes and Blood Biochemistry for blood dataset of the Trial 1.

Furthermore, upon examining Figure 4.2, it's evident that the downregulation of alkaline phosphatase persists across all three trials. This consistent pattern suggests that it might be a crucial indicator.

The rest subplots in Figure 4.1 are features related to blood biochemistry. In the Unhealthy group, there was a relative downregulation of C-reactive protein, cholesterol, and cortisol when compared to the Healthy group. Conversely, chloride and urea showed upregulation. These biomarkers are typically linked to inflammation and stress[31, 83, 50]. The observed downregulation might suggest a subdued inflammatory response and diminished metabolic activity in the Unhealthy group. Meanwhile, the elevated levels of chloride and urea could hint at potential kidney issues or an electrolyte imbalance[64, 52]. Prostaglandin E2 was lower than in the Healthy group. Prostaglandin E2 is commonly associated with inflammation and pain, and its down-regulation may suggest decreased immune function[8].

4.2 Correlation Heatmaps: Phi-K (Phik) Analysis

Figure4.3 shows the heatmap, which provides a comprehensive visualization of the correlation matrix using the phik correlation coefficient[7], elucidating the relationships between

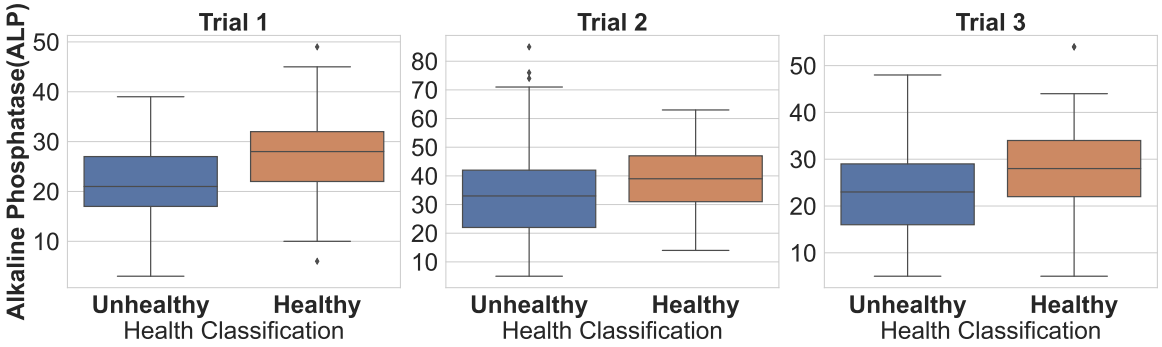

Figure 4.2: The boxplots of alkaline phosphatase on three trials.

different fish features for the biometrics collection on Trial 3. The visualization presented is a heatmap that represents a correlation matrix of various fish attributes. Each cell in the heatmap corresponds to the correlation coefficient between two attributes. The color intensity of each cell is indicative of the strength of the correlation, with darker red indicating stronger relationships. The diagonal elements, naturally, have a correlation coefficient of 1, signifying a perfect correlation of an attribute with itself. To enhance interpretability, the heatmap is annotated with specific symbols based on the magnitude of the correlation. Cells with correlation values having a value greater than 0.7 are annotated with ***, those with values between 0.3 and 0.7 are marked with **, and cells with values below 0.3 are denoted by *. The lower half of the heatmap displays the actual correlation coefficients, while the upper half provides these symbolic annotations. The term `empty` within a feature name indicates the absence of any specific body part during measurement, signifying that the observable is distinct or measured on the whole body.

It is evident from the data that only two features, `event` and `weight_liver`, are significantly associated with health status, as indicated by the ** notation. This suggests that these features might play a more pivotal role in health classification compared to others. As presented in Table 4.2, two features exhibit statistically significant differences between health groups. However, the features pinpointed by the tests are `girth_empty` and `temperature_celsius`, which deviate from the previously mentioned features. It's crucial to mention here that correlation does not imply causation. This discrepancy necessitates a more in-depth analysis to determine the most influential features for health classification.

Besides, the attribute `girth_empty` emerges as a significant predictor, manifesting strong correlations with several other attributes, implying potential multicollinearity. Specifically, its correlation coefficients with `weight_fillet`, `weight_viscera`, `weight_heart`, and `weight_liver` are three stars. This implies that the gonad's weight might be intrinsically linked to these attributes, possibly due to shared biological processes or developmental stages.

Additionally, the attribute `satiation_ration` exhibits a perfect correlation of 1 with the event, denoting a direct and unwavering relationship. A glance at Figure 3.1 reveals that as events progress, the feeding satiation ration diminishes from 100 to 0, an expected trend. In stark contrast, its correlation with `weight_viscera` stands at 0, signifying a complete lack of association between the two attributes. This might suggest that fasting does not exert a significant impact on the viscera. Such a pronounced discrepancy necessitates a deeper exploration to understand the underlying causes driving this phenomenon.

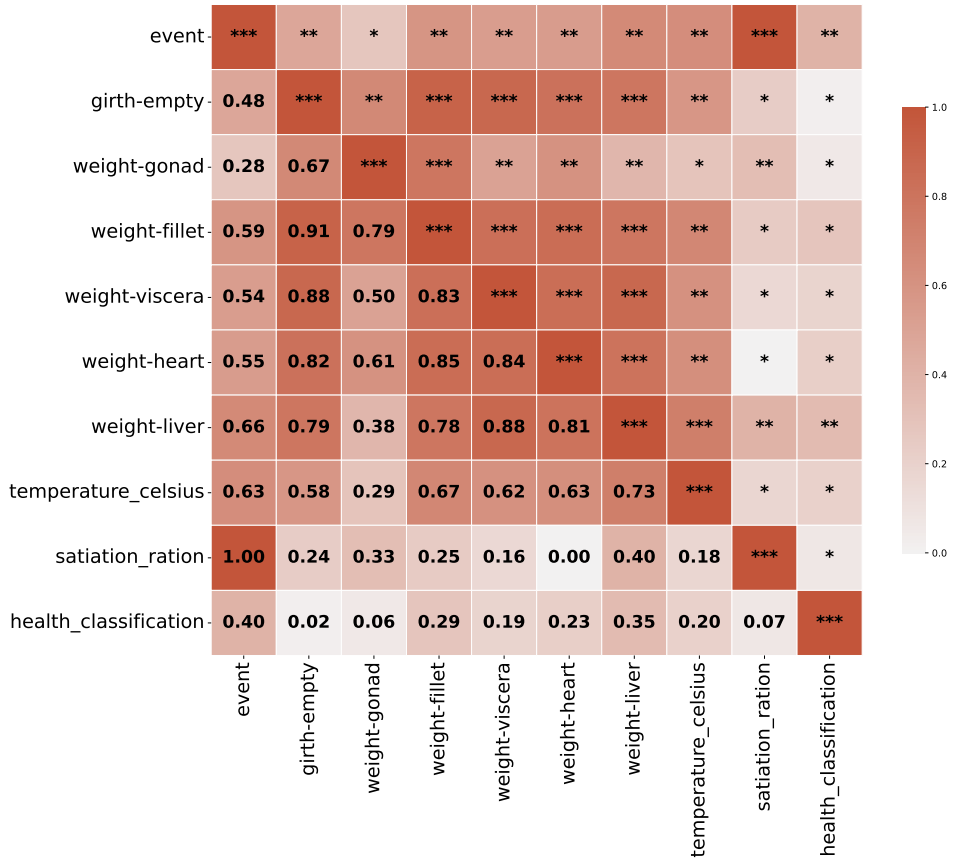

Figure 4.3: The heatmap of biometrics dataset on trial 3, which visualizes the correlation matrix of various fish attributes. The color intensity and annotations represent the strength and direction of the correlation between the attributes

4.3 PCA and t-SNE for Dataset Visualization

As mentioned in Chapter 2, PCA and t-SNE are powerful tools for dimension reduction. We utilize their ability to visualize the data. In the context of classification, we generally look for clear separations between different classes. A dataset with well-separated clusters in the t-SNE plot often suggests that a classifier would have an easier time distinguishing between classes, leading to better performance. The purple dots always refer to the unhealthy group and the yellow dots refer to the healthy group.

4.3.1 PCA Plots and Analysis

Figure 4.4 provides a visual representation of the data distribution across the three trials using PCA. For Trial 1, the data is segmented into roughly four distinct clusters along dimension 1. Multiple separate clusters hint at possible sub-groups within the healthy and unhealthy categories. While the rightmost cluster seems challenging to differentiate due to the close proximity of the data points, the other clusters have regions that predominantly contain only one class or a few noise dots. This suggests that these clusters might represent distinct groups within the data. The presence of a few data points of a different color within these predominantly single-colored clusters might indicate outliers or instances that are harder to classify.

The data distribution in Trial 2 is characterized by clusters where the right region is

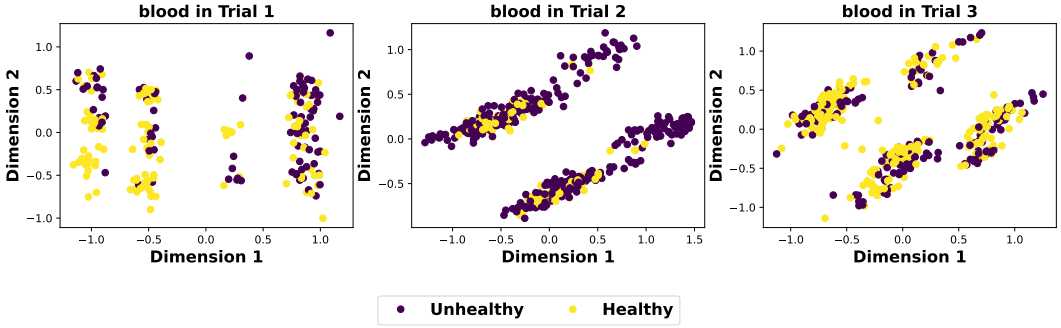

Figure 4.4: The PCA plots for the blood collection across the three trials.

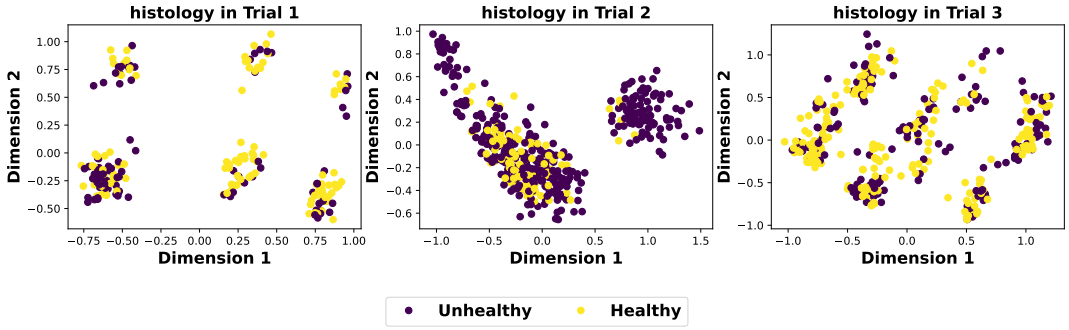

Figure 4.5: The PCA plots for the histology collection across the three trials

predominantly populated by unhealthy instances. This clear demarcation on one side of the clusters suggests a pattern or trend specific to the unhealthy class in this trial. The left side of these clusters, however, seems to be more mixed, indicating that while the unhealthy instances have a distinct pattern, the healthy ones might be more dispersed.

Trial 3 presents the most challenging scenario. The data points, irrespective of their class, are intermingled, making it difficult to discern any clear clusters or groups. The lack of clear separation between the two classes suggests that the features might not be as discriminative in this trial, or there might be other underlying factors causing this overlap. This might imply a low accuracy of classification.

In summary, the PCA plots for the three trials highlight the variability in data distribution and class separability across trials. While Trial 1 and Trial 2 offer some degree of separability (with Trial 1 being more distinct), Trial 3 underscores the challenges in distinguishing between the two classes. The nature of the data or the underlying biology under different conditions of three trials might contribute to these differences.

The comparison between the PCA plots of the blood collection (4.4) and the histology collection (4.5) reveals distinct differences in the data structure and distribution of the two collections. The histology PCA plot exhibits different cluster shapes compared to the blood PCA plot. This change in shape indicates that the underlying features and their relationships in the histology dataset are different from those in the blood dataset. The distinct cluster shapes in the histology plot suggest that the data points in this collection might be influenced by different sets of features or interactions between features, leading to unique data distributions.

Despite the differences in cluster shapes between the two collections, a consistent pattern is observed across the three trials for both collections. Specifically, Trial 3 consistently ap-

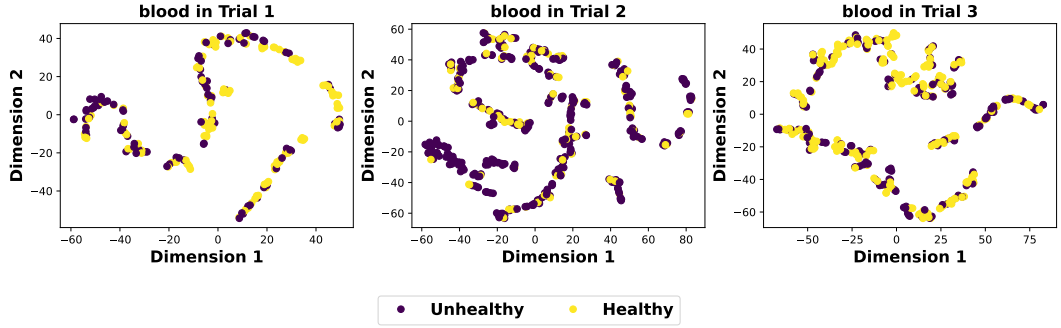

Figure 4.6: The t-SNE plots for the blood collection across the three trials

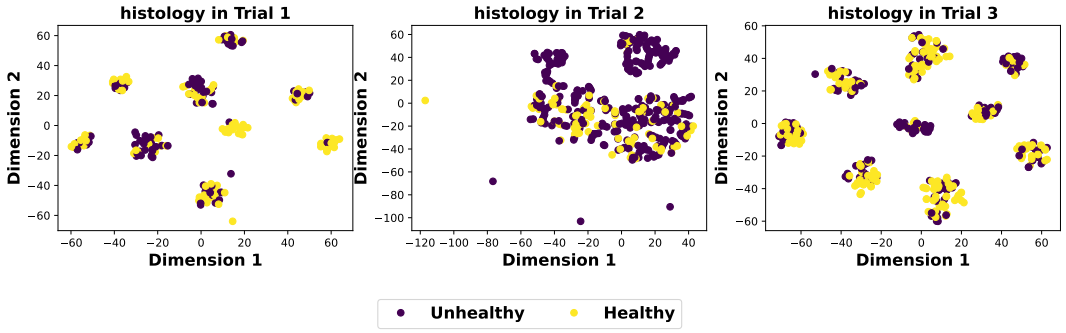

Figure 4.7: The t-SNE plots for the histology collection across the three trials

pears to be the most challenging in terms of class separability for both blood and histology datasets. This consistency suggests that while the specific features and their relationships might differ between the two collections, certain overarching trends or factors affect the data in a similar manner across trials.

The distinct data structures in the histology collection, as visualized in the PCA plot, imply that machine learning models might perform differently on this dataset compared to the blood dataset. The lack of clear class separability in Trial 3 for both collections indicates that models might face challenges in achieving high classification accuracy for this trial, irrespective of the collection.

4.3.2 t-SNE Plots and Analysis

The t-SNE visualizations provide a more intricate view of the data distribution compared to PCA, focusing on preserving local structures.

The t-SNE plots for both blood (shown in Figure 4.6) and histology (shown in Figure 4.7) collections show a higher number of clusters compared to their PCA counterparts. This is expected, as t-SNE is designed to capture local structures, leading to the formation of smaller, more defined clusters. The presence of more clusters in the histology collection for Trial 1 suggests that there might be sub-groups within the main classes, which could be indicative of finer-grained patterns or sub-types within the data.

In both t-SNE plots, Trial 1 and Trial 2 exhibit regions where clusters are relatively pure, meaning they predominantly contain data points from one class. This suggests that there are regions in the data where instances of one class are densely packed. Trial 3, however, shows more mixed clusters, indicating a higher degree of overlap between classes. The spread is relatively even, with many clusters where both categories intermingle. There's no clear

separation between the two categories. This aligns with previous observations that Trial 3 is more challenging in terms of class separability. However, the presence of almost pure clusters(regions) in Trial 1 and Trial 2 indicates that there are distinct sub-groups within the data that are well-separated.

4.4 Chapter Summary

In this chapter, we delve deep into the King Salmon dataset to identify underlying patterns and trends. By employing a mix of visualization tools and quantitative methods, the data's complex nature is understood, and key trends are highlighted. Techniques like the Shapiro-Wilk test evaluate data normality, guiding the choice of further tests. Results indicate variations between health groups, with certain biochemical markers playing pivotal roles. However, discrepancies between trials underscore the dataset's complexity. Visualization tools like PCA shed light on the data distribution, revealing challenges and insights for potential health classification strategies.

Chapter 5

Feature Selection

In this chapter, we delve deeply into the realm of feature selection, a critical process in refining machine learning models. The chapter commences by shedding light on the chosen methodologies, setting the stage for further discussions. Subsequent sections pivot to the nuances of classification and its evaluation, ensuring a holistic understanding of the methods employed. The key of this chapter is the detailed exploration of the results yielded by different approaches: filter methods, wrapper methods, and embedded methods. The main goal is to reduce the number of features without reducing or even improving the performance. Finally, we present and compare the results obtained from machine learning with feature selection against those derived from statistical methods.

5.1 Choice of Methods

Feature selection methodologies can be broadly classified into three categories: filter methods, wrapper methods, and embedded methods.

- Filter methods we used here include the ReliefF, Chi-Squared, and Mutual Information, which evaluate the relevance of features to target variables independently of any learning algorithms. These methods are the common methods worked effectively as discussed in Chapter 2. Additionally, an intersection method is employed to eliminate redundancy, selecting features that are universally significant across different criteria. Alternatively, a union of features selected through three distinct methods is utilized, thereby furnishing a comprehensive set of features for the construction of predictive models. Such advancements could further enhance the effectiveness of feature selection to identify the important features, potentially leading to more accurate and efficient health classifications for king salmon.

- Wrapper methods, represented by techniques like RFECV, evaluate subsets of features by training models on each subset and using the resultant performance as a criterion for feature selection. The main difference between REF and RFECV is that the REF performs cross-validation on the reduced feature set to evaluate the model's performance.

- Embedded methods amalgamate the strengths of both filter and wrapper methods by integrating feature selection into the process of model training. Algorithms such as LR, Linear Support Vector Classification[32](named SVM in the tables below), and RF algorithms were used for the wrapper and embedded methods. However, it is crucial to note that the features selected through these algorithms may not be universally optimal due to the different criteria used for evaluation across various algorithms.

5.2 Classification and Evaluation

To evaluate the efficacy of feature selection, we employ SVM as the classifier for classification purposes. Each SVM that is trained on a distinct dataset is regarded as an individual model. In order to evaluate the performance of models on the test set, the F1 score is chosen as the evaluation metric, a widely acknowledged measure for unbalanced classification tasks. This score is computed using the confusion matrix, which tabulates the instances of predicted versus actual labels for both positive and negative classes. This process yields four distinct outcomes: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Here we mainly focus on detecting unhealthy king salmon and thus consider the unhealthy class as the positive class. The real meaning of the four outcomes is listed below:

- True Positives (TP): These are the unhealthy instances correctly classified as unhealthy.
- False Positives (FP): These are the healthy instances incorrectly classified as unhealthy.
- True Negatives (TN): These are the healthy instances correctly classified as healthy.
- False Negatives (FN): These are the unhealthy instances incorrectly classified as healthy.

The formulas for the evaluation metrics employed are provided below. Regardless of whether it's the F1 score, precision, or recall, a higher value consistently signifies better performance.

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

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

When shown in tables, the average F1 scores are expressed in percentages and the standard deviation, which is encapsulated in parentheses. In our experiments, each feature selection method was executed to 30 independent runs, each corresponding to a distinct data split. For each run, the data is split into training and test sets with a consistent ratio of 8:2. The observed variability in the results across these runs can be attributed to the inherent characteristics of our dataset. Specifically, the limited size of the dataset, coupled with the presence of outliers, can introduce significant variability in the training and test sets during each split. Such variability can lead to different feature subsets being deemed important in different runs, thereby affecting the consistency of the feature selection process.

The performance of the feature selection methods based on the 30 independent runs is compared using the Wilcoxon rank-sum test with a significance level of 0.05. Performance indicators were symbolically represented to facilitate interpretation. Specifically, the notation **(–)** was used to denote scenarios where the performance diminished in comparison to models utilizing the full feature set. Conversely, the symbol ($\approx$) is indicative of performance metrics that remained comparable to those achieved with the complete feature repertoire.

Our primary objective in this chapter is to identify salient features instrumental in discerning health conditions. The ideal outcome would be models that leverage a reduced number of features yet deliver performance metrics either equivalent to or even surpassing those of models trained on the entire feature set. Given this goal, our analysis placed a heightened emphasis on the statistical difference in performance(i.e. significantly similar is notated using the symbol ($\approx$)) rather than the performance scores, as these symbols provide a more intuitive understanding of how feature selection impacts model efficacy.

Datasets	Allfeatures	Relief	chisquare	mutualinfo	intersection	union
blood	71.86(2.60)	75.52(1.54)**(+)**	75.12(1.95)**(+)**	75.19(1.78)**(+)**	74.22(2.05)**(+)**	73.82(1.87)**(+)**
composition	80.88(3.29)	83.61(2.80)**(+)**	84.28(2.81)**(+)**	84.14(2.85)**(+)**	80.30(4.85)(≈)	81.91(2.87)(≈)
FCR	67.54(1.41)	66.82(2.93)(≈)	67.99(1.53)(≈)	67.56(2.44)(≈)	66.85(2.97)(≈)	68.12(1.48)(≈)
SharedMeal	69.41(1.40)	70.28(1.72)(≈)	67.80(3.38)(≈)	69.44(1.45)(≈)	65.59(5.76)**(–)**	69.84(1.48)(≈)
growth	69.09(1.71)	69.74(1.97)(≈)	68.81(2.56)(≈)	69.35(1.59)(≈)	66.98(5.51)(≈)	69.34(1.74)(≈)
assessment	69.97(2.18)	70.48(1.73)(≈)	67.99(3.82)(≈)	70.32(1.85)(≈)	67.73(4.00)(≈)	70.06(2.05)(≈)
histology	78.26(1.91)	78.99(1.79)(≈)	78.95(1.97)(≈)	79.10(1.79)(≈)	77.50(2.09)(≈)	78.47(1.79)(≈)
biometrics	71.68(1.86)	73.56(1.56)**(+)**	75.07(2.59)**(+)**	72.19(1.75)(≈)	72.17(3.79)(≈)	71.82(1.79)(≈)

Table 5.1: Filter method - Training F1 score - Trial 1

Datasets	Allfeatures	RelifF	mutualinfo	chisquare	intersection	union
blood	61.01(6.39)	62.32(5.82)(≈)	61.75(7.33)(≈)	63.65(6.67)(≈)	63.08(6.96)(≈)	61.96(6.55)(≈)
composition	66.36(9.79)	67.50(9.64)(≈)	65.69(9.50)(≈)	66.20(8.62)(≈)	67.11(8.14)(≈)	66.10(9.87)(≈)
FCR	64.47(6.92)	64.61(6.94)(≈)	65.03(6.15)(≈)	65.00(6.65)(≈)	65.00(6.24)(≈)	65.37(5.54)(≈)
SharedMeal	63.40(6.11)	64.21(6.77)(≈)	62.68(7.43)(≈)	65.11(6.26)(≈)	61.21(8.85)(≈)	63.33(6.59)(≈)
growth	64.57(6.04)	64.73(5.39)(≈)	63.77(5.54)(≈)	65.14(5.87)(≈)	62.55(7.43)(≈)	64.19(6.07)(≈)
assessment	62.99(7.94)	63.55(7.43)(≈)	63.19(6.55)(≈)	64.17(6.96)(≈)	63.58(6.03)(≈)	63.26(7.79)(≈)
histology	59.88(7.19)	60.09(7.45)(≈)	60.07(6.23)(≈)	60.31(7.84)(≈)	60.96(6.64)(≈)	60.08(7.28)(≈)
biometrics	62.98(5.90)	64.65(7.65)(≈)	64.59(7.85)(≈)	62.79(6.70)(≈)	63.60(9.21)(≈)	63.01(6.34)(≈)

Table 5.2: Filter method - Test F1 score - Trial 1

5.3 Results of Filter Methods

For each trial, we will mainly utilize the results presented in three tables for analysis. For the tables below, the term "Collections" refers to different datasets that have been used in this study. Each collection, such as blood, composition, FCR, etc., represents a unique dataset with its specific features and instances. The values under each column present the results of the SVM model's performance employing diverse feature selection techniques. The "Allfeatures" method incorporates every feature present in the dataset without resorting to any form of selection or reduction, which is our baseline. In contrast, "RelifF," "chisquare," and "mutualinfo" are filter feature selection techniques correspondingly, each harnessing its unique criteria to cherry-pick features. The "intersection" technique adopts a consensus-driven strategy, preserving only those features that multiple individual methods simultaneously deem pivotal. On the other hand, the union approach maintains any feature highlighted by at least one of the individual methods.

Filter Methods Results on Trial 1 Observing the training F1 scores in Table 5.1, it is evident that for most datasets, employing filter feature selection methods generally leads to an improvement in performance compared to using all features. Notably, the "blood" and "biometrics" datasets exhibit a marked enhancement in performance with the application of these methods. On the other hand, the test F1 scores in Table 5.2 tend to be slightly lower, indicating that the performance improvements in the training phase do not always translate to the test phase and a potential overfitting scenario during training. However, it is worth noting that the disparities between training and test scores are not extremely high, suggesting that the model has a reasonable generalization capability. The increase in the standard deviation may caused by the limited test size due to the data splits, which could be seen that the composition dataset has the largest standard deviation.

Besides, the average count of features selected is depicted in Table 5.3. It enumerates the mean tally of features that remain after the application of each feature selection technique across diverse collections. This table is instrumental in underlining the extent of dimension-

number of features	All features	ReliefF	chisquare	mutualinfo	intersection	union
blood	39	25	23	26	16	33
composition	117	50	65	50	29	82
FCR	6	4	4	4	3	5
SharedMeal	5	4	3	4	2	5
growth	6	5	3	5	3	6
assessment	8	5	4	5	3	6
histology	37	29	32	31	22	36
biometrics	17	11	11	10	6	14

Table 5.3: Results comparison table - number of feature selected(AVG) -Trial 1

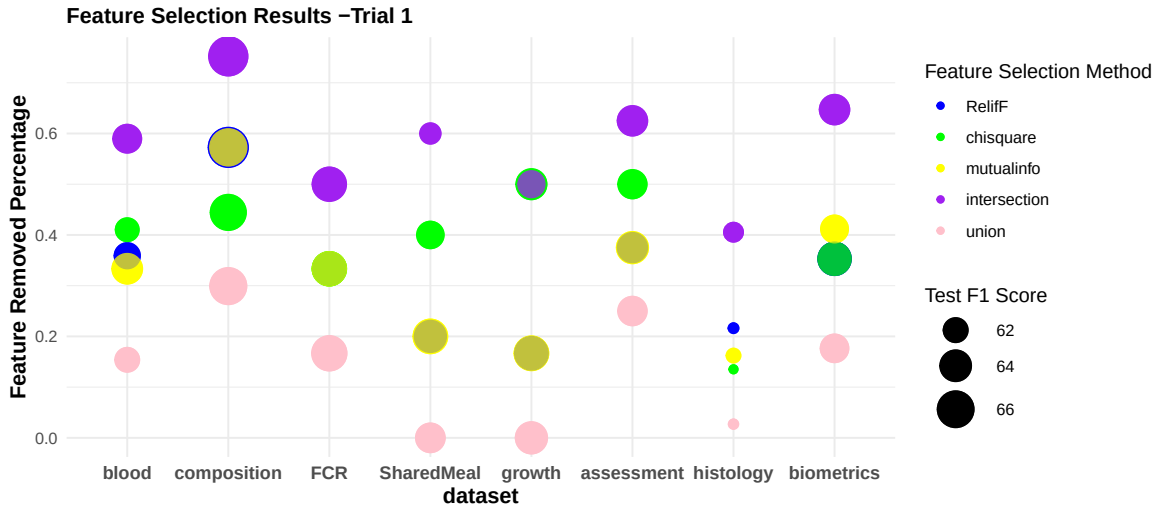

Figure 5.1: Comparison of filter feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 1.

ality reduction attained by every method. What is worth noticing is that the intersection usually selects the smallest number of features. A cursory glance at Tables 5.1 and 5.2 reveals that the intersection method outperforms or is on par with the models using a full feature set. This may indicate the intersection method potentially could be a preferable method in terms of model simplicity and performance trade-offs.

Lastly, the bubble plot, Figure 5.1, offers a visual representation of the outcomes of applying diverse feature selection techniques across multiple datasets in discerning the efficacy of each method and its implications on model performance. The x-axis of the plot delineates eight distinct datasets: blood, composition, FCR, SharedMeal, growth, assessment, histology, and biometrics. The y-axis represents the percentage of features excluded during the feature selection phase, formulated as 1-(number of selected features/total initial features). An integral aspect of the plot is the portrayal of five feature selection methodologies, each denoted by a unique color: ReliefF, chi-square, mutual information method, intersection, and union. The dispersion of colors across the plot intimates that the efficiency of a feature selection technique is not universally consistent but is contingent upon the specific dataset under consideration. Another pivotal element of the visualization is the use of bubble sizes to signify the Test F1 Score, a metric indicative of model performance. This dimensionality allows for an intricate understanding of the trade-off between the extent of feature reduction and the subsequent model performance. In certain scenarios, a substantial reduction in features does not invariably precipitate a decline in the test F1 score. This phenomenon is indicative of the removed features being superfluous or non-contributory to model performance.

Datasets	Allfeatures	Relief	chisquare	mutualinfo	intersection	union
blood	79.68(2.15)	80.75(2.48)(≈)	82.28(1.84)**(+)**	82.72(1.85)**(+)**	81.36(2.56)**(+)**	81.03(2.13)(≈)
composition	80.27(3.34)	83.58(2.58)**(+)**	82.79(2.54)**(+)**	80.79(3.04)(≈)	81.75(3.64)**(+)**	80.68(3.18)(≈)
FCR	57.19(5.54)	60.71(5.21)**(+)**	60.75(3.93)**(+)**	60.88(5.02)**(+)**	61.46(4.98)**(+)**	59.01(5.23)(≈)
SharedMeal	73.32(5.66)	77.35(1.73)**(+)**	74.70(6.62)**(+)**	70.26(7.08)(≈)	77.30(1.68)**(+)**	70.01(7.18)(≈)
growth	61.12(5.74)	71.66(3.22)**(+)**	71.30(3.27)**(+)**	62.57(7.90)(≈)	70.51(5.25)**(+)**	62.13(7.08)(≈)
assessment	61.07(3.24)	75.58(6.40)**(+)**	72.70(6.05)**(+)**	67.05(3.20)**(+)**	75.75(8.32)**(+)**	61.65(4.72)(≈)
histology	78.43(1.87)	80.84(1.53)**(+)**	79.51(2.33)**(+)**	79.14(1.90)(≈)	79.32(2.58)(≈)	78.40(1.89)(≈)
biometrics	74.03(5.08)	78.73(4.56)**(+)**	75.30(5.12)(≈)	76.89(6.47)**(+)**	74.43(7.68)(≈)	75.19(5.24)(≈)

Table 5.4: Filter method - Training F1 score - Trial 2

Datasets	Allfeatures	RelifF	mutualinfo	chisquare	intersection	union
blood	72.91(3.77)	73.61(4.07)(≈)	75.75(3.56)**(+)**	76.39(3.91)**(+)**	75.32(4.35)**(+)**	74.28(4.22)(≈)
composition	70.48(7.77)	74.71(6.11)**(+)**	72.60(7.33)(≈)	71.19(5.92)(≈)	74.27(5.97)(≈)	70.55(7.18)(≈)
FCR	52.30(9.37)	54.91(9.81)(≈)	55.29(11.46)(≈)	55.85(9.53)(≈)	55.28(11.48)(≈)	53.68(9.11)(≈)
SharedMeal	69.63(9.38)	73.56(6.11)(≈)	70.88(10.70)(≈)	65.62(10.56)**(–)**	73.28(6.66)(≈)	65.06(9.96)**(–)**
growth	59.37(7.06)	70.64(4.87)**(+)**	69.70(4.63)**(+)**	60.43(9.54)(≈)	69.57(7.22)**(+)**	60.17(8.77)(≈)
assessment	57.97(5.58)	72.93(9.64)**(+)**	69.97(8.74)**(+)**	63.18(6.55)**(+)**	74.27(10.66)**(+)**	58.18(6.90)(≈)
histology	71.76(4.26)	72.97(4.40)(≈)	72.44(5.02)(≈)	71.80(5.23)(≈)	73.71(4.48)(≈)	71.85(4.63)(≈)
biometrics	68.20(5.49)	74.98(4.80)**(+)**	70.15(5.68)(≈)	71.33(6.99)**(+)**	71.28(7.22)**(+)**	69.58(6.23)(≈)

Table 5.5: Filter method - Test F1 score - Trial 2

As illustrated in the bubble plot Figure 5.1, the purple dots, symbolizing the intersection method, consistently occupy the uppermost positions, denoting the maximum percentage of feature elimination. Furthermore, within the same column, the size of the purple bubbles remains relatively consistent across the majority of datasets. This uniformity in bubble size suggests that the excised features are likely redundant or do not significantly enhance the model's performance. Conversely, the union of the filter methods is the worst one that sometimes takes the whole feature sets.

The intersection method, which takes the common features selected by the three aforementioned filter methods, generally demonstrates competitive performance. The intersection method's strength lies in its ability to harness the collective wisdom of multiple filter methods. Moreover, real-world datasets, such as the ones used in this study, often contain a plethora of features, some of which might be noisy, redundant, or irrelevant. By focusing on the commonalities among the three filter methods, the intersection method effectively filters out such extraneous features, leading to a more robust and accurate model. This approach inherently reduces the risk of including noisy or irrelevant features, which can degrade model performance. This reduction in dimensionality can lead to more efficient models that are less prone to overfitting. However, it's worth noting that the intersection method's performance is dataset-dependent, especially for our datasets that use a real-world dataset that has more complex characteristics.

When comparing the same method across various datasets, a consistent observation emerges: the size of the dots corresponding to the 'histology' dataset is invariably the smallest. This suggests that the model's performance on this dataset is suboptimal, likely attributable to its inherent data structure.

Filter Methods Results on Trial 2 In the second trial, as illustrated in Table 5.4, the intersection results reveal a notable pattern. For most datasets in Trial 2, there appears to be a statistically significant improvement with fewer features. Overall, the training F1 scores exhibit a similar trend as observed in the first trial. The application of filter feature selection methods, in general, leads to an enhancement in performance compared to the utilization of

number of features	All features	ReliefF	chisquare	mutualinfo	intersection	union
blood	40	31	28	29	19	37
composition	117	48	75	34	18	84
fcr	6	2	2	2	2	3
SharedMeal	5	2	2	2	2	3
growth	6	2	4	3	2	5
assessment	15	3	8	3	2	10
histology	37	19	27	26	12	33
biometrics	16	5	9	9	3	13

Table 5.6: Results comparison table - number of feature selected(AVG) - Trial 2.

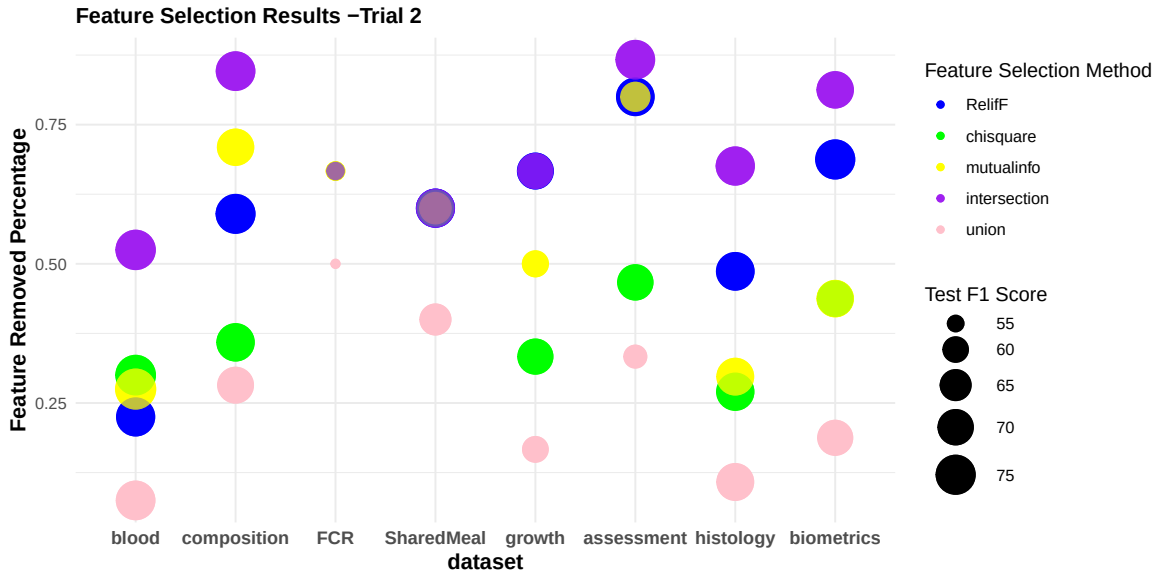

Figure 5.2: Comparison of filter feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 2.

all features. For example, when filter methods are applied to the "growth" dataset, there is a notable enhancement in performance, registering an increase of approximately 10.44% at its peak.

When we transition to the test phase, as illustrated in Table 5.5,the intersection method also improved 4 datasets performance with less features. Overall, while most datasets maintain similar performance levels, more datasets in Trial 2 exhibit improved performance compared to those in the same collection in Trial 1. These results echo findings from the first trial, hinting at potential overfitting during training. However, the modest discrepancies between training and test scores suggest that the model retains a commendable ability to generalize. The increased standard deviation, especially in datasets like "composition", could be attributed to the limited test size due to data splits. However, there are more models that showed statistically significant improvement.

The average count of features selected for the second trial is presented in Table 5.6. This table accentuates the dimensionality reduction achieved by each method. Similar to the first trial, the intersection method typically selects the fewest features. A closer inspection of Tables 5.4 and 5.5 reveals that the intersection method, in many cases, either outperforms or is competitive with models using the full feature set. This underscores the potential of the intersection method as a viable choice, striking a balance between model simplicity and performance. As evidenced by 5.2, the purple dots, representing the intersection method,

Datasets	Allfeatures	Relief	chisquare	mutualinfo	intersection	union
blood	66.70(2.39)	67.73(2.08)**(+)**	68.63(2.06)**(+)**	68.12(2.07)**(+)**	67.00(3.41)(≈)	67.24(2.36)(≈)
composition	66.50(2.62)	68.96(2.53)**(+)**	67.82(2.46)(≈)	68.46(2.09)**(+)**	67.49(2.92)(≈)	66.72(2.40)(≈)
FCR	66.28(2.58)	56.59(3.06)**(–)**	55.72(5.00)**(–)**	56.57(3.10)**(–)**	56.90(3.10)**(–)**	59.69(5.34)**(–)**
SharedMeal	54.31(4.17)	55.87(2.26)(≈)	55.96(2.74)(≈)	54.30(3.05)(≈)	55.29(4.43)(≈)	54.04(4.35)(≈)
growth	59.85(2.76)	60.77(1.73)(≈)	61.76(1.69)**(+)**	60.92(1.46)(≈)	58.98(2.74)(≈)	59.75(2.81)(≈)
assessment	60.38(2.17)	61.37(1.93)(≈)	61.18(1.94)(≈)	61.23(2.21)(≈)	59.14(4.13)(≈)	60.36(2.39)(≈)
histology	75.95(1.95)	76.87(1.90)(≈)	76.17(1.54)(≈)	77.51(1.52)**(+)**	75.58(2.09)(≈)	75.95(1.95)(≈)
biometrics	62.89(2.35)	63.88(1.64)(≈)	63.72(2.16)(≈)	63.40(2.21)(≈)	56.34(6.50)**(–)**	62.92(2.29)(≈)

Table 5.7: Filter method - Training F1 score - Trial 3.

Datasets	Allfeatures	RelifF	mutualinfo	chisquare	intersection	union
blood	51.92(8.19)	52.55(7.56)(≈)	51.74(7.98)(≈)	52.04(8.30)(≈)	52.26(6.98)(≈)	51.88(7.91)(≈)
composition	52.21(7.40)	52.91(7.11)(≈)	51.74(6.77)(≈)	51.94(8.39)(≈)	52.95(7.16)(≈)	52.68(7.53)(≈)
FCR	54.98(11.14)	53.17(12.43)(≈)	50.45(13.37)(≈)	51.94(12.02)(≈)	52.45(12.87)(≈)	53.51(12.58)(≈)
SharedMeal	47.81(9.37)	54.97(6.74)**(+)**	50.22(7.17)(≈)	47.81(7.16)(≈)	53.38(6.88)**(+)**	47.47(8.51)(≈)
growth	54.87(6.13)	55.99(5.62)(≈)	56.73(5.94)(≈)	55.23(6.12)(≈)	55.08(6.24)(≈)	54.71(6.14)(≈)
assessment	54.81(7.53)	56.02(5.65)(≈)	55.29(6.39)(≈)	54.97(5.88)(≈)	54.99(6.30)(≈)	54.89(7.36)(≈)
histology	55.89(4.52)	55.68(4.47)(≈)	55.42(5.18)(≈)	56.24(4.90)(≈)	54.01(4.82)(≈)	55.86(4.56)(≈)
biometrics	57.48(6.60)	57.41(6.03)(≈)	57.40(6.93)(≈)	56.93(6.67)(≈)	51.19(7.22)**(–)**	57.59(6.54)(≈)

Table 5.8: Filter method - Test F1 score - Trial 3

consistently maintain their position at the top of the plot. Moreover, their size remains comparable to the dots situated below them. This observation reinforces the notion that the intersection method surpasses other methods in performance, effectively eliminating redundant features.

The unbalanced ratio in Trial 2 is notably higher than in Trial 1, with datasets such as "composition" having a ratio as high as 4.65. This could introduce challenges in model training, as the model might be biased towards the majority class. However, the filter feature selection methods, especially the intersection method, still manage to deliver competitive performance, underscoring their resilience and effectiveness even in the face of class imbalances. In conclusion, the second trial reaffirms the observations from the first trial regarding the potential benefits of filter feature selection methods, especially the intersection method. The real-world datasets used in this study, with their inherent complexities, highlight the robustness and adaptability of these methods.

Filter Methods Results on Trial 3 In the third trial, as delineated in Table 5.7, the training F1 scores for most datasets are noticeably lower than those observed in the first two trials. This is particularly evident in datasets such as "FCR", where the F1 score has seen a significant decline. The application of filter feature selection methods still tends to enhance performance compared to using all features, but the magnitude of this enhancement is less pronounced than in previous trials.

Transitioning to the testing phase, as depicted in Table 5.8, the enhancements in performance that were evident during the training phase do not translate as robustly. The majority of the datasets exhibit a marked decline in performance when juxtaposed against the outcomes from the first two trials. For the "SharedMeal" collection dataset, the performance with all features was subpar, even falling below random guessing. This downturn in performance during the testing phase for Trial 3 is anticipated. As visualized in the PCA plot 4.4, it becomes evident that the data from Trial 3 presents a more challenging landscape for classification. The overlap or lack of clear distinction between data points in the PCA plot for Trial 3 suggests that the underlying patterns or features that differentiate the classes are less pro-

number of features	All features	ReliefF	chisquare	mutualinfo	intersection	union
blood	38	34	32	30	22	38
composition	128	78	63	87	40	103
fcr	6	1	1	1	1	2
SharedMeal	5	3	3	3	2	5
growth	6	4	5	4	2	6
assessment	12	7	8	8	4	11
histology	37	29	32	35	24	37
biometrics	10	6	6	7	3	9

Table 5.9: Results comparison table - number of feature selected(AVG) - Trial 3.

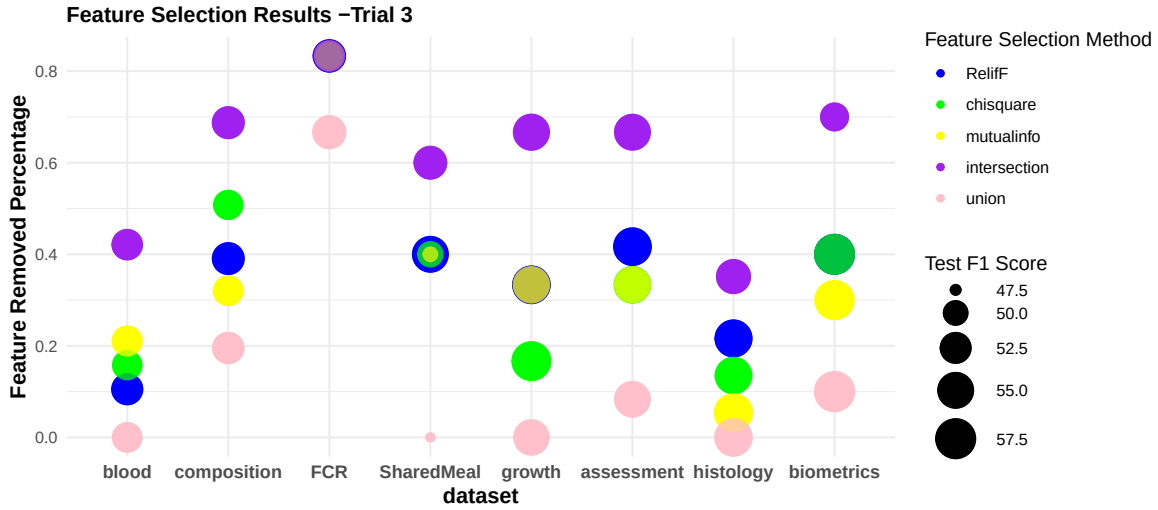

Figure 5.3: Comparison of filter feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 3

nounced in this trial. This could be attributed to various factors, including changes in data collection, inherent variability in the samples, or other external influences that affected Trial 3 differently than the previous trials. However, after eliminating certain features, there was a marked improvement in the "SharedMeals" dataset, with results now exceeding those of random predictions. This indicates that the intersection method can effectively handle and enhance results for intricate datasets.

The average count of features selected for the third trial, presented in Table 5.9, indicates a varied dimensionality reduction across datasets. As with the previous trials, the intersection method typically selects the fewest features. Additionally, for the 'FCR' collection, nearly every method opted for a singular feature. However, only two models, including the intersection method, exhibited an improvement in performance. A closer inspection of Tables 5.7 and 5.8 reveals that the intersection method, except the 'biometrics', either outperforms or is competitive with models using the full feature set.

Figure 5.3 reveals the distinct pattern associated with the intersection method, represented by the purple dots. Irrespective of the test F1 score, the intersection method consistently achieves the highest percentage of feature removal. This is further accentuated by a pronounced vertical gap distinguishing it from other methods. This reiterates that the intersection method's ability to select a reduced set of features while maintaining competitive performance remains evident.

The unbalanced ratio in Trial 3 stands at approximately 0.77 for all datasets, which is a stark contrast to the previous trials. This lower ratio indicates a less balanced distribution

Datasets	Allfeatures	LR	SVM	RF
blood	71.86(2.60)	69.99(5.20)(≈)	68.60(6.32)(≈)	69.42(5.10)(≈)
composition	80.88(3.29)	83.49(5.81)**(+)**	82.68(6.36)(≈)	80.27(4.59)(≈)
FCR	67.54(1.41)	67.35(2.25)(≈)	67.07(2.90)(≈)	58.53(13.62)(≈)
SharedMeal	69.41(1.40)	68.92(2.99)(≈)	69.06(2.78)(≈)	68.50(3.89)(≈)
growth	69.09(1.71)	67.05(3.46)(≈)	67.49(2.98)(≈)	65.99(4.05)**(–)**
assessment	69.97(2.18)	68.64(2.57)(≈)	69.48(1.93)(≈)	66.01(14.39)(≈)
histology	78.26(1.91)	71.16(2.68)**(–)**	71.53(5.44)**(–)**	67.18(5.22)**(–)**
biometrics	71.68(1.86)	73.61(1.77)**(+)**	73.82(2.59)**(+)**	72.30(5.25)**(+)**

Table 5.10: Wrapper method - Training F1 score - Trial 1.

Datasets	Allfeatures	LR	SVM	RF
blood	61.01(6.39)	59.39(7.36)(≈)	58.91(7.17)(≈)	55.11(8.50)**(–)**
composition	66.36(9.79)	68.97(10.00)(≈)	66.03(10.80)(≈)	64.87(11.07)(≈)
FCR	64.47(6.92)	64.20(7.13)(≈)	63.75(8.08)(≈)	53.11(15.85)**(–)**
SharedMeal	63.40(6.11)	62.46(7.83)(≈)	63.15(6.79)(≈)	62.61(5.95)(≈)
growth	64.57(6.04)	62.56(6.93)(≈)	62.48(7.47)(≈)	60.91(7.06)**(–)**
assessment	62.99(7.94)	62.16(7.90)(≈)	62.50(8.39)(≈)	58.61(17.35)(≈)
histology	59.88(7.19)	60.84(7.12)(≈)	58.62(4.85)(≈)	59.20(7.84)(≈)
biometrics	62.98(5.90)	66.36(6.80)(≈)	66.14(6.93)(≈)	62.78(7.97)(≈)

Table 5.11: Wrapper method - Test F1 score - Trial 1.

between unhealthy and healthy instances compared with Trial 1. Thus, we are not surprised that the F1 scores in Trial 3 are generally lower. This could be attributed to the inherent complexities and characteristics of the real-world datasets used in this study. Real-world datasets often come with their unique challenges, and the results from Trial 3 underscore the importance of robust feature selection and model training techniques to handle such complexities.

Summary By analysing the results obtained by of filter feature selection methods, it is observed that these methods often enhance the training performance across almost all datasets. However, these training benefits do not always extend to the testing phase, hinting at potential overfitting. Notably, the disparity between training and test scores is not large, suggesting the model's satisfactory generalization. Despite its simple approach, the intersection method often performs competitively, if not better than models with all features, indicating its potential preference for balancing simplicity and performance. Overall, the intersection method's ability to get feature reduction and maintain performance stands out across trials.

5.4 Results of Wrapper Methods

Wrapper Methods Results on Trial 1 In Trial 1, utilizing the wrapper method, as depicted in Table 5.10, the training F1 scores for datasets such as "histology" show an improvement when the LR algorithm is employed, as compared to using all features. However, for datasets like "biometrics", there is a noticeable decline in performance when using the RF algorithm. This suggests that while the wrapper method can enhance performance for some datasets, it might not be universally beneficial across all datasets and algorithms.

Moving to the testing phase, as detailed in Table 5.11, there is a discernible decline in performance across all datasets when the RF algorithm is employed, even if some of these reductions are not statistically significant. In contrast, the performance remains statistically consistent across all datasets when utilizing the SVM and LR algorithms. This observation

Datasets	Allfeatures	LR	SVM	RF
blood	39	11	15	20
composition	117	24	20	63
FCR	6	5	5	4
SharedMeal	5	4	4	5
growth	6	4	5	4
assessment	8	5	5	5
histology	37	7	12	4
biometrics	17	7	10	9

Table 5.12: Wrapper method - number of feature selected(AVG) - Trial 1.

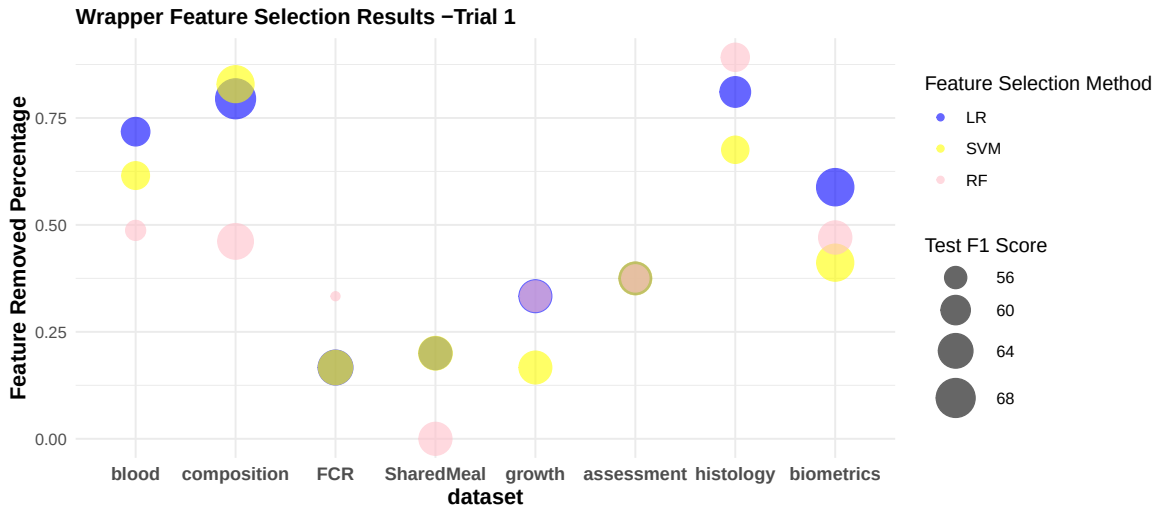

Figure 5.4: Comparison of wrapper feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 1

highlights the critical interplay between algorithm selection and the wrapper method. Comparing the results of the wrapper method in Trial 1 with the intersection filter method in the previous trial, it is evident that the wrapper method, especially when paired with the LR algorithm, tends to select fewer features. This suggests that the wrapper method might offer a more aggressive dimensionality reduction. The chosen algorithm can profoundly impact the resultant performance, emphasizing the need for judicious algorithm selection in the context of the wrapper method.

The average count of features selected using the wrapper method for Trial 1 is presented in Table 5.12. This table accentuates the dimensionality reduction achieved by each algorithm. For instance, for the "blood" dataset, the LR algorithm selects only 11 features, a significant reduction from the original 39 features. This reduction in feature dimensionality can be advantageous in terms of computational efficiency and model interpretability. What stands out is that the wrapper method employing RF not only exhibits a decline in the Test F1 score but also retains over 50 percent of the features. This suggests that this approach may not be a good choice for these datasets.

Upon examining the bubble plot Figure 5.4, a distinct pattern emerges, contrasting with what is observed in 5.1. Notably, none of the three methods consistently dominates in performance. The size of the top two bubbles is generally comparable, with the exception of the 'FCR' dataset. In this particular dataset, while the wrapper method employing the RF algorithm exhibits a larger bubble size than the other two methods, its Test F1 score is markedly low, rendering it less significant for consideration. Additionally, it is imperative to high-

Datasets	Allfeatures	LR	SVM	RF
blood	79.68(2.15)	77.09(5.23)(≈)	80.25(2.89)(≈)	81.45(2.50)**(+)**
composition	80.27(3.34)	80.12(5.53)(≈)	81.09(4.29)(≈)	82.00(3.65)**(+)**
FCR	57.19(5.54)	59.28(7.98)(≈)	58.89(7.78)(≈)	55.09(7.47)(≈)
SharedMeal	73.32(5.66)	68.10(7.02)**(–)**	69.30(7.92)(≈)	71.74(5.40)**(–)**
growth	61.12(5.74)	58.78(5.48)(≈)	59.55(7.16)(≈)	62.73(6.74)(≈)
assessment	61.07(3.24)	59.07(5.46)**(–)**	58.75(3.09)**(–)**	67.54(5.75)**(+)**
histology	78.43(1.87)	76.02(2.74)**(–)**	72.92(8.09)**(–)**	79.27(2.32)(≈)
biometrics	74.03(5.08)	76.02(5.87)(≈)	74.53(5.46)(≈)	75.43(4.55)(≈)

Table 5.13: Wrapper method - Training F1 score - Trial 2.

Datasets	Allfeatures	LR	SVM	RF
blood	72.91(3.77)	73.54(4.77)(≈)	74.68(3.79)(≈)	75.19(5.06)**(+)**
composition	70.48(7.77)	71.16(8.44)(≈)	72.32(7.04)(≈)	72.30(8.98)(≈)
FCR	52.30(9.37)	49.10(13.56)(≈)	48.16(13.80)(≈)	50.66(13.03)(≈)
SharedMeal	69.63(9.38)	63.48(9.08)**(–)**	64.91(10.09)**(–)**	67.56(8.96)(≈)
growth	59.37(7.06)	57.38(7.47)(≈)	58.41(8.02)(≈)	60.79(8.09)(≈)
assessment	57.97(5.58)	56.92(7.06)(≈)	57.12(4.35)(≈)	64.85(7.34)**(+)**
histology	71.76(4.26)	70.50(4.33)(≈)	67.74(7.40)**(–)**	71.94(4.37)(≈)
biometrics	68.20(5.49)	71.34(6.43)**(+)**	68.92(5.79)(≈)	68.54(5.36)(≈)

Table 5.14: Wrapper method - Test F1 score - Trial 2.

light certain collections, namely 'fcr', 'SharedMeal', and 'growth'. These collections exhibit a notably low feature removal percentage, often well below the 0.5 mark, even for the top-performing methods. This is in stark contrast to the intersection of filter methods, which consistently achieves a removal percentage above 0.5 for these datasets. Intrinsically, these collections start with a limited feature set. When subjected to wrapper methods, the feature reduction is minimal, with the process either eliminating a mere single feature or, in certain cases, making no eliminations at all. Given that these datasets yield comparable test F1 scores, the wrapper method using LR or SVM might be a good option.

Wrapper Methods Results on Trial 2 An examination of Table 5.13, detailing the training F1 scores for Trial 2, reveals a departure from the patterns observed in Trial 1. Notably, the wrapper method employing RF enhances performance on several datasets. In contrast, the LR and SVM algorithms either maintain their performance or exhibit a decline on certain datasets. This underscores the variability in data structure across trials for the same collection. Furthermore, it emphasizes that the choice of algorithm can profoundly impact performance outcomes, contingent on the specific characteristics of the data.

Moving to the testing phase, as depicted in Table 5.14, both the "blood" and "assessment" datasets demonstrate enhanced performance when subjected to the RF algorithm, marking a distinct improvement over their performance in Trial 1. However, it's worth noting that the test F1 scores of the three models lag behind the scores achieved by the intersection methods in Trial 2, as presented in Table 5.5.

Table 5.15 delineates the average count of features chosen when employing the wrapper method for Trial 2. Intriguingly, for certain datasets, the number of features selected surpasses that of the intersection filter method, as observed with the "composition" dataset. Moreover, despite the inclusion of a greater number of features, the performance of models on these datasets, when trained using the wrapper method, is diminished. This suggests that, for the data in Trial 2, the wrapper method may not be the best selection strategy.

Upon examining the plot 5.5, a consistent pattern emerges where no single method dominates across all datasets. Notably, the bubbles corresponding to the "FCR" dataset are

Datasets	Allfeatures	LR	SVM	RF
blood	40	13	17	34
composition	117	37	29	55
FCR	6	2	2	2
SharedMeal	5	3	2	4
growth	6	4	4	5
assessment	15	7	3	7
histology	37	22	19	26
biometrics	16	7	9	13

Table 5.15: Wrapper method - number of feature selected(AVG) - Trial 2

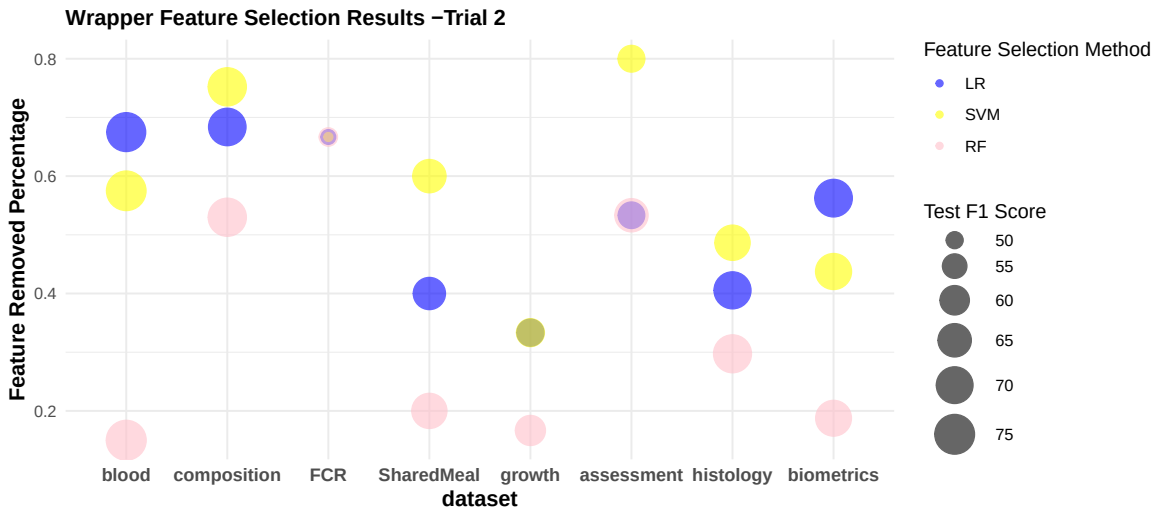

Figure 5.5: Comparison of wrapper feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 2.

markedly smaller than those of other datasets. This suggests that the FCR dataset may possess limited informative value for health classification. This observation aligns with the trends seen in 5.2, where the bubbles for the FCR dataset are also notably smaller compared to others.

Wrapper Methods Results on Trial 3 From Table 5.16, which showcases the training F1 scores for Trial 3, several observations emerge. Only for the "SharedMeal" dataset, the performance of all algorithms, when compared to using all features, indicates a similar performance. This suggests that the wrapper method's aggressive feature selection might not always yield optimal results. Similarly, for the "composition" and "FCR" datasets, all three algorithms—LR, SVM, and RF—demonstrate a decrease in performance compared to using all features. This trend is consistent across multiple datasets, such as "FCR" and "histology", highlighting the challenges of feature selection in real-world scenarios.

Transitioning to the test phase, as depicted in Table 5.17, the "SharedMeal" dataset presents an interesting outcome. The SVM algorithm, despite its suboptimal performance in the training phase, outperforms the other algorithms in the test phase. This underscores the importance of validating model performance on unseen data, as training performance might not always be indicative of real-world applicability.

In Table 5.18, the average count of features chosen using the wrapper method for Trial 3 is detailed. For the "composition" dataset, the LR algorithm selects a mere 22 features, marking a substantial reduction from the 40 features chosen by the intersection filter method.

Datasets	Allfeatures	LR	SVM	RF
blood	66.70(2.39)	59.00(4.82)**(–)**	61.67(3.30)**(–)**	62.80(8.17)(≈)
composition	66.50(2.62)	65.36(3.34)(≈)	63.22(5.19)**(–)**	60.00(6.37)**(–)**
FCR	66.28(2.58)	60.80(5.71)**(–)**	61.89(5.65)**(–)**	63.15(4.62)**(–)**
SharedMeal	54.31(4.17)	55.31(3.70)(≈)	55.65(2.21)(≈)	52.56(4.66)(≈)
growth	59.85(2.76)	57.07(3.13)**(–)**	57.84(3.19)**(–)**	59.74(2.75)(≈)
assessment	60.38(2.17)	58.63(3.21)**(–)**	58.38(3.44)**(–)**	60.77(2.83)(≈)
histology	75.95(1.95)	71.14(3.97)**(–)**	70.51(4.25)**(–)**	71.99(7.84)(≈)
biometrics	62.89(2.35)	58.14(6.18)**(–)**	59.39(4.44)**(–)**	61.75(3.22)(≈)

Table 5.16: Wrapper method - Training F1 score - Trial 3.

Datasets	Allfeatures	LR	SVM	RF
blood	51.92(8.19)	53.61(6.40)(≈)	55.99(6.05)(≈)	51.50(8.15)(≈)
composition	52.21(7.40)	52.05(6.47)(≈)	51.22(7.79)(≈)	48.77(8.67)(≈)
FCR	54.98(11.14)	51.74(12.28)(≈)	51.39(11.79)(≈)	52.12(15.43)(≈)
SharedMeal	47.81(9.37)	50.88(8.15)(≈)	55.57(5.96)**(+)**	47.44(6.59)(≈)
growth	54.87(6.13)	51.50(6.75)**(–)**	52.50(6.16)(≈)	55.75(6.33)(≈)
assessment	54.81(7.53)	52.98(6.20)(≈)	52.12(7.02)(≈)	55.21(6.48)(≈)
histology	55.89(4.52)	54.98(6.12)(≈)	55.53(5.02)(≈)	53.37(5.44)(≈)
biometrics	57.48(6.60)	53.45(8.47)(≈)	53.54(7.09)**(–)**	56.82(7.54)(≈)

Table 5.17: Wrapper method - Test F1 score - Trial 3.

Such a decrease in feature dimensionality can be beneficial, enhancing computational efficiency and facilitating model interpretability. However, it's essential to note that a pronounced reduction in features does not invariably lead to superior performance. For instance, with the "biometrics" dataset, while the intersection filter method opted for only 3 features, it yielded a performance that is statistically inferior. In contrast, the wrapper method employing LR achieved comparable results, albeit with 5 features. This suggests that the filter method might overlook feature interactions. It underscores the notion that there isn't a universally optimal method, emphasizing the need for a more comprehensive consideration of features.

An analysis of 5.6 reveals intriguing patterns. Specifically, for the 'blood' dataset, all bubbles are positioned higher than those in 5.3. Yet, their sizes remain comparable. This suggests that the wrapper method, for this dataset, manages to select fewer features while achieving enhanced performance. In contrast, for the majority of the other datasets, the bubbles corresponding to the filter method tend to occupy higher positions, indicating their relative superiority in those contexts.

Summary Across three trials, the efficacy of the wrapper method for feature selection varied depending on the dataset and algorithm used. In Trial 1, the wrapper method, especially with the LR algorithm, demonstrated the potential for aggressive dimensionality reduction. However, its performance benefits were not universally observed. Trial 2 highlighted the impact of data structure differences on results. While the RF algorithm showed promise in some datasets, LR and SVM often remained consistent or declined. Interestingly, the wrapper method sometimes surpassed the intersection filter method in feature selection but did not consistently improve performance. In Trial 3, the challenges of real-world feature selection became evident, with the wrapper method not always ensuring optimal results despite aggressive feature selection. Overall, these trials underscore the need for careful algorithm selection, the intricacies of individual datasets, and the absence of a universally optimal feature selection method.

Datasets	Allfeatures	LR	SVM	RF
blood	38	4	6	12
composition	128	22	18	20
FCR	6	2	2	4
SharedMeal	5	4	4	3
growth	6	4	5	5
assessment	12	6	7	7
histology	37	20	25	19
biometrics	10	5	6	8

Table 5.18: Wrapper method - number of feature selected(AVG) - Trial 3

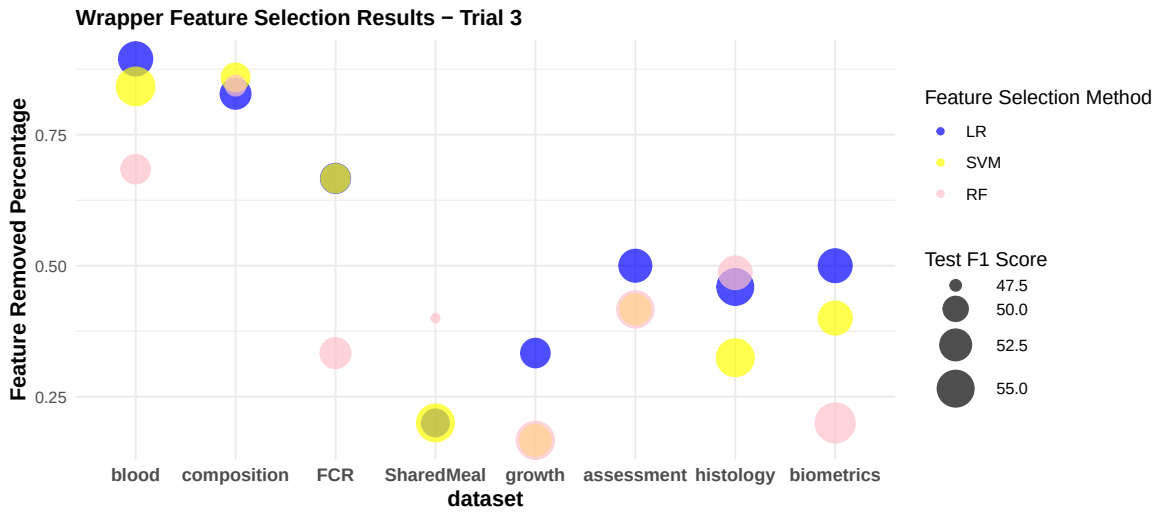

Figure 5.6: Comparison of wrapper feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 3.

5.5 Results of Embedded Methods

Embedded Methods Results on Trial 1 By making feature selection an intrinsic part of the model training process, embedded methods provide a more holistic view of feature importance than the filter, and more generality to different classifiers.

From Table 5.19, which presents the training F1 scores for Trial 1 using the embedded method, several observations can be discerned. The "composition" dataset, exhibits a significant improvement in performance compared to using all features. This underscores the capability of embedded methods, particularly when regularized using L1 regularization, to effectively select a subset of features that contribute most to the model's performance. A similar trend is observed for the "biometrics" dataset, where both LR and SVM algorithms outperform the all-features baseline. However, it's worth noting that for some datasets, such as "FCR", the embedded method results in a decline in performance, suggesting that the method's efficacy is contingent on the nature of the dataset and the underlying relationships between features.

Transitioning to the test phase, as depicted in Table 5.20, the results are more varied. Overall, the embedded method using LR or SVM could get comparable performance to the all-features baseline, while the RF algorithm exhibits a slight decline. This suggests that while embedded methods can be effective in feature selection during the training phase, their performance on unseen data can be influenced by various factors, including the algorithm's inherent criteria for evaluating feature importance.

Datasets	Allfeatures	LR	SVM	RF
blood	71.86(2.60)	73.96(1.71)**(+)**	70.48(3.68)(≈)	71.38(3.16)(≈)
composition	80.88(3.29)	86.52(2.69)**(+)**	85.45(2.90)**(+)**	83.40(4.11)**(+)**
FCR	67.54(1.41)	53.96(5.19)**(–)**	54.46(3.88)**(–)**	41.09(5.24)**(–)**
SharedMeal	69.41(1.40)	69.31(2.26)(≈)	67.01(3.69)(≈)	61.03(2.54)**(–)**
growth	69.09(1.71)	69.19(1.68)(≈)	63.40(1.94)**(–)**	63.47(3.11)**(–)**
assessment	69.97(2.18)	69.69(1.40)(≈)	68.42(2.94)(≈)	43.38(6.45)**(–)**
histology	78.26(1.91)	75.27(1.93)**(–)**	74.32(1.73)**(–)**	77.14(2.38)**(–)**
biometrics	71.68(1.86)	73.72(1.66)**(+)**	75.23(2.42)**(+)**	71.43(4.43)(≈)

Table 5.19: Embedded method - Training F1 score - Trial 1.

Datasets	Allfeatures	LR	SVM	RF
blood	61.01(6.39)	60.92(7.77)(≈)	59.06(7.36)(≈)	57.16(6.10)**(–)**
composition	66.36(9.79)	68.24(9.72)(≈)	65.85(10.77)(≈)	64.98(9.17)(≈)
FCR	64.47(6.92)	51.99(9.48)**(–)**	52.67(8.57)**(–)**	34.82(10.09)**(–)**
SharedMeal	63.40(6.11)	64.78(6.40)(≈)	61.80(8.30)(≈)	55.79(7.97)**(–)**
growth	64.57(6.04)	65.25(5.90)(≈)	61.47(7.12)(≈)	57.75(7.70)**(–)**
assessment	62.99(7.94)	64.62(5.74)(≈)	63.47(7.39)(≈)	38.31(10.27)**(–)**
histology	59.88(7.19)	61.35(6.36)(≈)	61.49(6.55)(≈)	61.32(7.03)(≈)
biometrics	62.98(5.90)	66.93(7.19)**(+)**	66.91(6.94)**(+)**	62.13(7.37)(≈)

Table 5.20: Embedded method - Test F1 score - Trial 1.

In Table 5.21, the average count of features chosen using the embedded method for Trial 1 is detailed. It becomes apparent that the embedded method often opts for a more concise feature set compared to the intersection of the filter method, as observed in datasets like 'histology'. However, the reduction is not as pronounced as that achieved by the wrapper methods. Given the comparable test performance, the wrapper method seems to be the most favorable choice for this dataset. In terms of feature selection, the wrapper method offers an edge in model interpretability over the intersection filter method, though it doesn't surpass the embedded method in this regard.

Upon analyzing Figure 5.7, it's evident that the embedded method, when employing the RF algorithm for the 'assessment' collection, occupies a higher position compared to both Figure 5.1 and Figure 5.4. However, given its test F1 score of approximately 38.32, this method is not the most suitable choice. For datasets characterized by a limited number of features, the intersection of the filter method tends to be positioned higher. Conversely, for more extensive datasets (i.e., those with a greater number of features), the wrapper method demonstrates its efficacy by removing a larger proportion of features while maintaining comparable test performance.

Embedded Methods Results on Trial 2 From Table 5.22, which delineates the training F1 scores for Trial 2. It's noteworthy that certain datasets, such as "SharedMeal", witness a decline in performance with the embedded method, particularly with the SVM algorithm. Besides, the embedded method using the RF algorithm performs better than the baseline on 6 out of 8 datasets. This underscores the method's sensitivity to the nature of the dataset and the relationships between features.

Transitioning to the test phase, as depicted in Table 5.23, the results are nuanced. The "blood" dataset, across all three algorithms (LR, SVM, and RF), demonstrates an improvement over the all-features baseline. This suggests that the embedded method's feature selection during the training phase translates effectively to unseen data. However, for datasets like "SharedMeal", the performance decline observed in the training phase is also mirrored in the test phase, particularly with the SVM algorithm. This reiterates the importance of al-

Datasets	Allfeatures	LR	SVM	RF
blood	29	18	17	20
composition	117	46	47	53
FCR	6	2	2	1
SharedMeal	5	3	3	2
growth	6	4	3	3
assessment	8	3	4	2
histology	37	14	15	17
biometrics	17	6	7	5

Table 5.21: Embedded method - number of feature selected(AVG) - Trial 1.

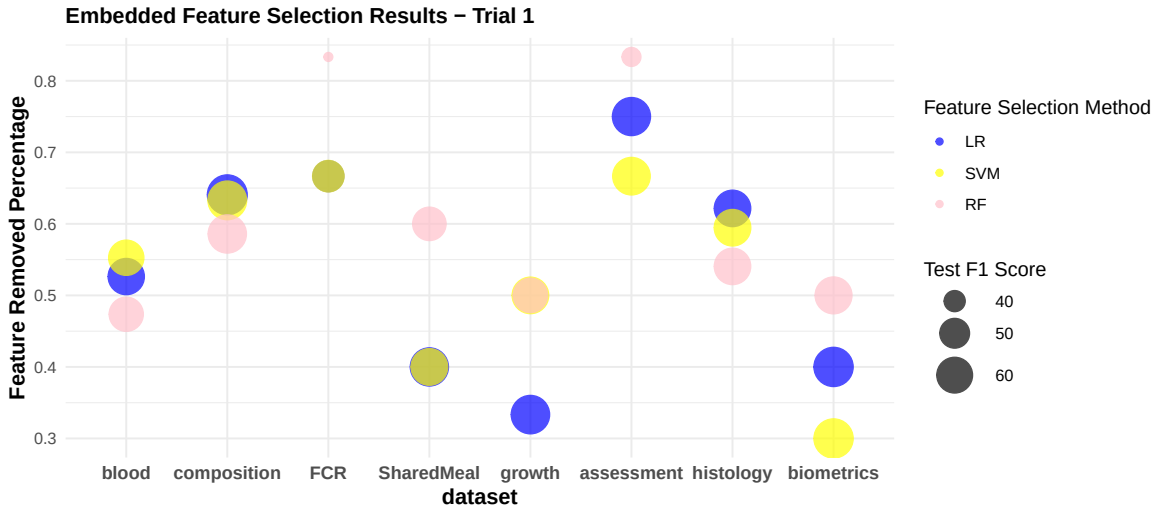

Figure 5.7: Comparison of embedded feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 1

gorithm selection in conjunction with the embedded method. From this table, the embedded method using RF seems the best one.

In Table 5.24, the average count of features chosen using the embedded method for Trial 2 is detailed. Across the board, the embedded method demonstrates a propensity to opt for a more concise feature set for all datasets. Notably, for the 'FCR' dataset, the application of the embedded method with the RF algorithm remarkably manages to achieve comparable performance using just a single feature.

Upon examining the Figure 5.8, it becomes clear that the 'FCR' dataset occupies a position higher than both Figure 5.1 and Figure 5.4. However, for the remaining datasets, either the intersection of the filter method or the wrapper method appears to be more advantageous.

Embedded Methods Results on Trial 3 The training F1 scores, as presented in Table 5.25, offer a mixed bag of results. For instance, the "blood" dataset shows a modest improvement when processed through the RF algorithm, reinforcing the embedded method's ability to discern feature importance intrinsically. However, the "composition" dataset experiences a decline in performance with the RF algorithm, which could be attributed to the algorithm's sensitivity to feature interdependencies. The "SharedMeal" dataset, interestingly, shows an improvement with the SVM algorithm, suggesting that the embedded method's performance can vary depending on the algorithm and dataset in question.

Moving to the testing phase, as depicted in Table 5.26, the outcomes are multifaceted.

Datasets	Allfeatures	LR	SVM	RF
blood	79.68(2.15)	80.35(1.29)(≈)	80.90(2.00)**(+)**	81.74(2.12)**(+)**
composition	80.27(3.34)	81.54(2.98)(≈)	83.23(3.03)**(+)**	83.23(2.30)**(+)**
FCR	57.19(5.54)	60.19(8.19)(≈)	60.39(8.41)(≈)	60.39(8.41)(≈)
SharedMeal	73.32(5.66)	69.50(7.13)**(–)**	65.54(8.58)**(–)**	69.78(6.85)**(–)**
growth	61.12(5.74)	58.98(7.31)(≈)	59.83(7.22)(≈)	67.53(3.34)**(+)**
assessment	61.07(3.24)	55.35(5.27)**(–)**	54.91(4.91)**(–)**	71.37(2.56)**(+)**
histology	78.43(1.87)	74.70(2.27)**(–)**	73.28(3.00)**(–)**	80.32(1.73)**(+)**
biometrics	74.03(5.08)	75.32(5.56)(≈)	75.11(4.76)(≈)	76.84(3.48)**(+)**

Table 5.22: Embedded method - Training F1 score - Trial 2.

Datasets	Allfeatures	LR	SVM	RF
blood	72.91(3.77)	75.33(3.24)**(+)**	75.78(3.36)**(+)**	76.62(3.45)**(+)**
composition	70.48(7.77)	71.19(7.89)(≈)	72.40(6.67)(≈)	74.17(5.95)**(+)**
FCR	52.30(9.37)	50.55(12.58)(≈)	50.06(13.36)(≈)	50.06(13.36)(≈)
SharedMeal	69.63(9.38)	64.60(10.46)**(–)**	61.46(11.10)**(–)**	65.00(9.94)**(–)**
growth	59.37(7.06)	57.56(8.81)(≈)	58.69(8.64)(≈)	65.13(5.32)**(+)**
assessment	57.97(5.58)	53.08(7.88)**(–)**	53.03(6.81)**(–)**	69.13(4.84)**(+)**
histology	71.76(4.26)	68.64(5.42)**(–)**	68.22(5.24)**(–)**	73.44(3.57)(≈)
biometrics	68.20(5.49)	70.28(6.02)(≈)	69.63(6.34)(≈)	69.33(5.50)(≈)

Table 5.23: Embedded method - Test F1 score - Trial 2.

For the ”SharedMeal” dataset, the application of the SVM algorithm under the embedded method yields a slight performance enhancement, akin to the results achieved by the wrapper method using SVM. However, the embedded method distinguishes itself by selecting only 2 features shown in Table 5.27, as opposed to the 4 chosen by the wrapper method, showcasing its efficiency. Furthermore, the embedded method, when employing the RF algorithm, manages to significantly boost model performance with just one additional feature compared to the filter method on the ’growth’ dataset. Analyzing the remaining datasets, a pattern emerges: the embedded methods tend to opt for fewer features in datasets with a smaller feature set, while gravitating towards a more extensive feature selection in larger datasets, especially when compared to wrapper methods using the same algorithm.

Upon analyzing Figure 5.9, it’s evident that the embedded method, when employing the RF algorithm for the ’assessment’ collection, occupies a higher position compared to both Figure 5.1 and Figure 5.4 like in Trial 1. Besides, given its performance is similar to the model with all features, this should be considered as the most suitable choice for this dataset. For the remaining datasets, either the intersection of the filter method or the wrapper method appears to be more advantageous.

Summary Embedded methods incorporate feature selection within the model training process, offering an approach to discerning feature importance. Across three trials, the results highlight the potential and challenges of the embedded method. In Trial 1, while datasets like ”composition” benefited, others such as ”FCR” saw diminished performance. The embedded method typically favored a concise feature set, yet not as aggressively as the wrapper method. By Trial 2, it became evident that the method’s performance was closely tied to the nature of the dataset and the algorithm used. Some datasets improved, while others, like ”SharedMeal”, declined when using certain algorithms. The RF algorithm emerged as a prominent choice in embedded methods. In Trial 3, results were mixed; the ”SharedMeal” dataset saw improvements with the SVM algorithm, whereas ”composition” underperformed with RF. An observable trend was the method’s inclination towards fewer features for smaller datasets and a larger feature subsets for larger datasets. This pat-

Datasets	Allfeatures	LR	SVM	RF
blood	40	15	15	16
composition	117	48	47	57
FCR	6	2	1	1
SharedMeal	5	2	2	2
growth	6	3	4	3
assessment	15	5	5	4
histology	37	15	14	17
biometrics	16	7	6	10

Table 5.24: Embedded method - number of feature selected(AVG) - Trial 2

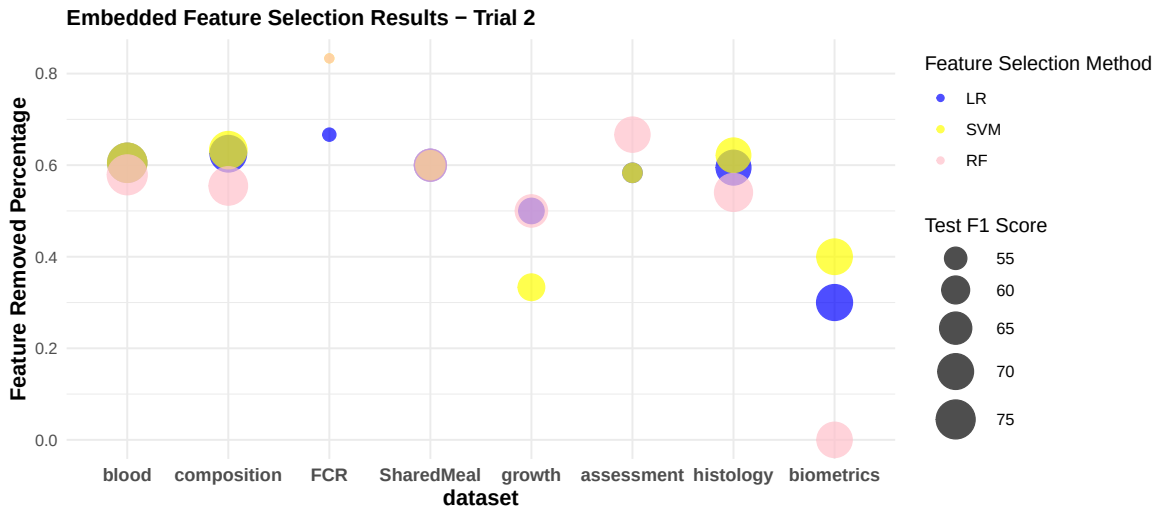

Figure 5.8: Comparison of embedded feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 2

tern contrasted with the wrapper method, which consistently pursued aggressive feature reduction. Comparatively, while embedded methods show promise, the wrapper method outperformed several datasets in terms of feature reduction and comparable performance, underlining the importance of dataset-specific and algorithm-specific considerations.

5.6 Discussions and Conclusions

The results presented in Table 5.28 offer a comprehensive overview of the best feature selection methods we investigated above across three trials for our 24 datasets. The selection of the most suitable feature selection method for each dataset is based on the performance(the average F1 score) and the number of features removed. All methods mentioned here achieved our goal of selecting fewer features while maintaining similar performance compared with using all features. The selection of the most suitable feature selection method for each dataset is rooted in a multifaceted criterion.

For the blood dataset, the filter method with intersection is deemed the best method in Trial 1, selecting 16 out of the 39 features. However, in subsequent trials, the Embedded and Wrapper methods, specifically with SVM, were favored, reflecting their ability to achieve comparable or superior performance with a reduced feature set.

The composition dataset exhibited a preference for the Wrapper method using SVM in Trials 1 and 3, while the Filter method with intersection is the best method in Trial 2. This suggests that while the Wrapper method might be more computationally intensive, its abil-

Datasets	Allfeatures	LR	SVM	RF
blood	66.70(2.39)	67.46(3.12)(≈)	65.50(3.03)(≈)	69.31(2.58)**(+)**
composition	66.50(2.62)	67.95(2.18)(≈)	67.53(3.00)(≈)	64.24(3.21)**(–)**
FCR	66.28(2.58)	63.66(4.49)**(–)**	63.95(4.49)(≈)	53.62(6.60)**(–)**
SharedMeal	54.31(4.17)	55.20(2.72)(≈)	57.00(1.79)**(+)**	56.60(2.45)**(+)**
growth	59.85(2.76)	56.14(3.17)**(–)**	54.62(5.62)**(–)**	60.73(1.29)(≈)
assessment	60.38(2.17)	59.15(3.19)(≈)	58.32(3.11)**(–)**	54.69(3.48)**(–)**
histology	75.95(1.95)	69.87(2.09)**(–)**	66.19(2.89)**(–)**	75.48(1.93)(≈)
biometrics	62.89(2.35)	60.85(1.94)**(–)**	55.61(5.62)**(–)**	60.43(2.56)**(–)**

Table 5.25: Embedded method - Training F1 score - Trial 3.

Datasets	Allfeatures	LR	SVM	RF
blood	51.92(8.19)	53.24(6.50)(≈)	54.18(5.04)(≈)	54.00(7.72)(≈)
composition	52.21(7.40)	52.67(7.83)(≈)	52.36(6.99)(≈)	48.08(7.04)**(–)**
FCR	54.98(11.14)	53.72(11.73)(≈)	54.07(11.57)(≈)	41.25(12.70)**(–)**
SharedMeal	47.81(9.37)	49.58(8.53)(≈)	54.04(6.70)**(+)**	50.93(5.85)(≈)
growth	54.87(6.13)	50.60(5.72)**(–)**	50.97(6.50)**(–)**	58.20(5.37)**(+)**
assessment	54.81(7.53)	52.52(6.63)(≈)	51.41(7.43)(≈)	51.19(6.43)(≈)
histology	55.89(4.52)	55.04(5.85)(≈)	53.81(4.94)(≈)	55.54(3.91)(≈)
biometrics	57.48(6.60)	55.67(7.37)(≈)	51.31(6.68)**(–)**	54.80(6.02)(≈)

Table 5.26: Embedded method - Test F1 score - Trial 3.

ity to select a subset of features (20 out of 117 in Trial 1 and 18 out of 128 in Trial 3) without compromising on performance makes it a viable choice.

For datasets like FCR and ShareMeals, the filter method with intersection consistently emerged as a preferred choice in multiple trials, indicating its effectiveness in these specific contexts.

The ”growth” dataset displayed a shift in preference from the Filter method in Trials 1 and 2 to the Embedded method with RF in Trial 3. This transition underscores the dynamic nature of feature selection, where the best method method can vary based on the specific trial or dataset nuances.

In the ”assessment” dataset, the Embedded method, particularly with LR and RF, is favored in Trials 1 and 3. In contrast, the Filter method with intersection is the method of choice in Trial 2. This dataset serves as a testament to the adaptability of the Embedded method in selecting a minimal feature set while maintaining robust performance.

The ”histology” and ”biometrics” datasets further accentuate the versatility of the Embedded method, especially with LR, in selecting a concise set of features without sacrificing model accuracy.

Compare with the statistical methods Upon examination of Tables 4.3 and 5.29, representing features with significant differences between health groups and those identified by their average feature importance score respectively, some notable observations emerge. Primarily, there exists a convergence on a subset of six shared features across the two tables. For each category, they are listed as the first or second feature in Table 5.29. The overlapping features, including ”alkaline_phosphatase” from the Enzymes category and ”chloride” and ”urea” from Blood Biochemistry, among others, underscore their pivotal role in discerning health conditions, as both statistical analysis and machine learning feature selection methods independently identified them. However, the divergence in the remaining features between the two tables suggests that while certain parameters are significant in distinguishing health conditions, they may not necessarily possess the highest predictive importance when implementing machine learning models. Conversely, some features that aren’t statistically

Datasets	Allfeatures	LR	SVM	RF
blood	38	15	16	21
composition	128	52	54	70
FCR	6	2	2	2
SharedMeal	5	3	2	2
growth	6	3	4	3
assessment	12	5	7	2
histology	37	16	17	18
biometrics	10	5	5	6

Table 5.27: Embedded method - number of feature selected(AVG) - Trial 3.

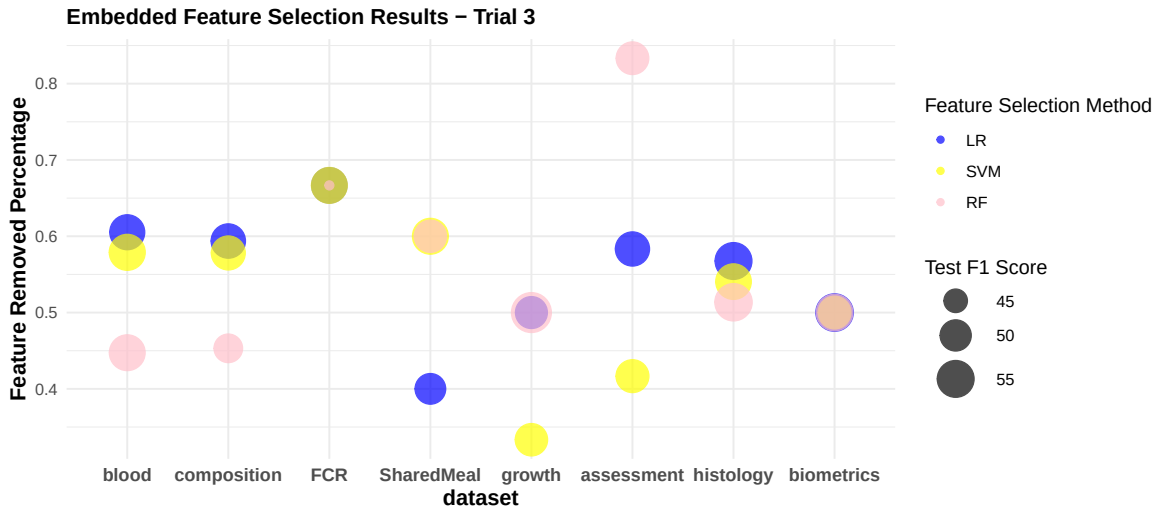

Figure 5.9: Comparison of embedded feature selection methods and their impact on model performance (test F1 Score) across different datasets on Trial 3.

divergent might still carry substantial weight in a machine learning context.

Comparing Table 5.28 with Table 4.2, it becomes evident that the presence of statistically significant differences in features for each health group doesn't necessarily translate to their utility in classification tasks. For the blood dataset, Table 4.2 shows a decrease in significant features across trials, while Table 5.28 indicates a more pronounced reduction in features selected by the best feature selection method. This suggests that as the number of statistically significant features diminishes, the feature selection methods become more selective. The FCR datasets present a scenario where significant features decrease to none in the latter trials in Table 4.2. However, the best feature selection methods continue selecting features, suggesting that certain non-significant features still hold predictive information.

Upon a comparison of Table 5.30 and Table 5.31, it becomes evident that the Machine Learning Feature Selection (MLFS) method consistently outperforms the traditional statistical (STAT) methods in feature selection across various collections and trials.

In most collections, the MLFS method either demonstrates a marked improvement or maintains comparable performance to the entire feature set. This is particularly evident in the Blood, Composition, and Biometrics collections, where the MLFS method consistently achieves higher or nearly equivalent F1 scores compared to the STAT methods on the test set. The MLFS method's ability to discern and prioritize features that contribute most to predictive power is evident.

Conversely, the STAT methods, while occasionally achieving comparable results, often

	Trial1		Trial2		Trial3	
Dataset	**Method**	**Feature**	**Method**	**Feature**	**Method**	**Feature**
blood	Filter(intersection)	39 - 16	Embedded(SVM)	38 - 15	Wrapper(SVM)	38 - 6
composition	Wrapper(SVM)	117 - 20	Filter(intersection)	117 - 18	Wrapper(SVM)	128 - 18
FCR	Filter(intersection)	6 - 3	Embedded(LR)	6 - 1	Filter(intersection)	6 - 1
ShareMeals	Filter(intersection)	5 - 2	Filter(intersection)	5 - 2	Embedded(SVM)	5 - 2
growth	Filter(intersection)	6 - 3	Filter(intersection)	6 - 2	Embedded(RF)	6 - 3
assessment	Embedded(LR)	12 - 3	Filter(intersection)	15 - 2	Embedded(RF)	12 - 2
histology	Wrapper(RF)	37 - 4	Filter(intersection)	37 - 12	Embedded(LR)	37 - 16
biometrics	Embedded(LR)	17 - 6	Filter(intersection)	16 - 3	Embedded(LR)	10 - 5

Table 5.28: The best feature selection methods across three trials for various datasets. Each entry in the "Feature" column represents the total number of features in the dataset followed by the number of features selected by the method.

Enzymes	alkaline_phosphatase
Blood Biochemistry	chloride, urea, cortisol, sodium, cholesterol, colour, magnesium, phosphate, cortisol, calcium
Blood Cell Related	monocytes_abs, haptoglobin, Albumin_globulin_ratio, buffy_coat_thickness
Other	event, temperature_celsius, satiation_ration

Table 5.29: Important features identified by feature selection method on blood collection in trial 01 for health classification.

fall short in certain collections, such as SharedMeals, Growth, and Histology. In these instances, the F1 scores using STAT methods are notably lower, indicating a potential limitation in relying solely on traditional statistical tests for feature selection in these datasets.

In summary, while both methods have their merits, the MLFS method demonstrates a more consistent and often superior performance across the board. This suggests that leveraging machine learning algorithms for feature selection can offer a more robust and adaptive approach, especially when dealing with complex real-world datasets.

Datasets	Trial 1		Trial 2		Trial 3	
	All	MLFS	All	MLFS	All	MLFS
blood	61.01(6.39)	66.03(10.80)(≈)	72.91(3.77)	75.78(3.36)**(+)**	51.92(8.19)	55.99(6.05)(≈)
composition	66.36(9.79)	68.34(9.21)(≈)	70.48(7.77)	74.27(5.97)(≈)	52.21(7.40)	51.22(7.79)(≈)
FCR	64.47(6.92)	65.00(6.24)(≈)	52.30(9.37)	50.06(13.36)(≈)	54.98(11.14)	52.45(12.87)(≈)
SharedMeals	63.40(6.11)	61.21(8.85)(≈)	69.63(9.38)	73.28(6.66)(≈)	47.81(9.37)	54.04(6.70)**(+)**
growth	64.57(6.04)	62.55(7.43)(≈)	59.37(7.06)	69.57(7.22)**(+)**	54.87(6.13)	58.20(5.37)**(+)**
assessment	62.99(7.94)	64.62(5.74)(≈)	57.97(5.58)	74.27(10.66)**(+)**	54.81(7.53)	51.19(6.43)(≈)
histology	59.88(7.19)	59.20(7.84)(≈)	71.76(4.26)	73.71(4.48)(≈)	55.89(4.52)	55.04(5.85)(≈)
biometrics	62.98(5.90)	66.93(7.19)**(+)**	68.20(5.49)	71.28(7.22)**(+)**	57.48(6.60)	55.67(7.37)(≈)

Table 5.30: The test F1 score of all three trials using the features that have statistical differences between two health groups.

Datasets	Trial 1		Trial 2		Trial 3	
	All	STAT	All	STAT	All	STAT
blood	61.01(6.39)	62.90(6.15)(≈)	72.91(3.77)	75.69(3.52)**(+)**	51.92(8.19)	52.02(6.29)(≈)
composition	66.36(9.79)	68.34(9.21)(≈)	70.48(7.77)	67.92(6.65)(≈)	52.21(7.40)	52.98(6.21)(≈)
FCR	64.47(6.92)	65.80(7.32)(≈)	52.30(9.37)	52.30(9.37)(≈)	54.98(11.14)	54.98(11.14)(≈)
SharedMeals	63.40(6.11)	52.08(8.72)**(–)**	69.63(9.38)	65.26(10.02)**(–)**	47.81(9.37)	50.01(6.81)(≈)
growth	64.57(6.04)	63.38(6.41)(≈)	59.37(7.06)	39.18(5.88)**(–)**	54.87(6.13)	54.59(5.60)(≈)
assessment	62.99(7.94)	62.08(7.64)(≈)	57.97(5.58)	51.80(6.78)**(–)**	54.81(7.53)	54.52(5.80)(≈)
histology	59.88(7.19)	61.73(8.08)(≈)	71.76(4.26)	63.16(5.84)**(–)**	55.89(4.52)	56.50(5.00)(≈)
biometrics	62.98(5.90)	55.78(7.37)**(–)**	68.20(5.49)	57.33(4.61)**(–)**	57.48(6.60)	54.59(5.60)(≈)

Table 5.31: The test F1 score of all three trials using the features that have statistical differences between two health groups.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The project embarked on an in-depth exploration across multiple datasets, starting from the data preprocessing step until machine learning with feature selection step. The overarching goals shown below, including enhancing feature selection for improved predictive outcomes were successfully achieved, cementing the project's significance in the realm of data analytics and machine learning.

Initially, the extensive datasets collected provided a rich foundation for analysis. Critical to the success of subsequent exploration was the meticulous preprocessing undertaken. The quality of the data was enhanced, ensuring noise reduction and normalization of feature scales. Through a comprehensive cleansing process, missing values were effectively managed, and potential outliers were addressed, thus optimizing the data for subsequent analytical stages. This rigorous data processing stage set the stage for meaningful insights and highlighted the importance of clean and prepared datasets in the machine learning pipeline.

The EDA phase provided invaluable insights into the underlying structure and patterns within the data. By visualizing the distributions, correlations, and potential relationships among features, a deeper understanding of the datasets was established. Through techniques like scatter plots, histograms, and heatmaps, intrinsic characteristics, and anomalies were identified. This chapter was not just about revealing patterns but also about identifying potential pitfalls and challenges that might arise during modeling. By investigating datasets, the EDA set the stage for a more informed approach to feature selection, emphasizing the necessity to approach each dataset with a tailored strategy.

Building upon the insights from EDA, the feature selection phase delved into optimizing models by identifying the most influential variables. Through multiple trials, the efficacy of various methods - filter, wrapper, and embedded - was applied. While certain datasets flourished under one method, others presented challenges, revealing the method's sensitivity to dataset characteristics and algorithm choices. The intersection of filter methods appears to possess a broad generalization capability, as it can effectively select important features in approximately half of the datasets. The embedded method's integrated approach to feature selection within model training showcased its potential but also revealed its sensitivity to dataset nature and algorithm choice. Notably, the wrapper method often emerged as the most adept at aggressive feature reduction while maintaining performance. In summary, we successfully identified the important features of all datasets and achieved a superior or statistically equal F1 score with these features. Through our efforts, it became clear that while traditional statistical methods hold value in specific contexts, the machine learning feature selection methods present a more adaptive and often superior alternative in most scenarios.

A general observation reveals that the accuracy, as indicated by the F1 scores, remains commendably high across the datasets and trials. Notably, values such as 75.78 in the "blood" dataset during Trial 2 and 74.27 in both the "composition" and "assessment" datasets are particularly striking. The overarching trend indicates that the application of machine learning feature selection often leads to competitive, if not superior, F1 scores compared to using all features across the trials. This pattern indicates the efficacy and potential of machine learning with feature selection methods in discerning and utilizing influential features to achieve optimal classification outcomes.

6.2 Future Work

We have identified the following list of tasks for future work.

Pre-processing Process For the preprocessing phase, there exist several methodologies that have not been explored in this study but may potentially enhance model performance. Future work may consider implementing these additional preprocessing techniques to ascertain their impact on improving the predictive accuracy and robustness of the model.

Data Integration In preceding chapters, various methods have been applied to individual datasets without exploring the possibility of integrating these datasets into fewer or a singular consolidated dataset. Two viable approaches for future investigation include: (1) merging features within the same collection, and (2) merging instances within the same trial. Additionally, with another trial currently underway and yielding more data, the endeavor of integrating datasets post-collection of new data is a worthwhile pursuit that promises to offer valuable insights and enhance the robustness of the model's predictive capabilities.

Feature Engineering There is a potential for exploring feature engineering to either create new features or modify existing ones, thereby enhancing their information content[57]. However, caution should be exercised to avoid being overly stringent in feature selection. Over-aggressive trimming of features may lead to the omission of those that are crucial for interpreting complex datasets, undermining the model's predictive prowess in the process. A viable approach entails presenting the selected significant features to domain experts and soliciting their insights regarding the potential relationships and interdependencies among these features.

Resampling Techniques Resampling is a crucial technique that is worth further exploration, specifically focusing on the oversampling of minority classes and undersampling of majority classes [14]. This approach aims to address class imbalance, which is a common issue in dataset preparation. By creating a balanced dataset, standard classification approaches can shatter all classes adequately, thereby improving the model's predictive accuracy. This method has already been applied in the biology area[24].

However, such techniques are not utilized in this study. Techniques such as the oversampling, while useful, have the potential to introduce bias into the model due to their alteration of the class distribution within the dataset[89]. To the best of our knowledge, part of the data employed in this study represents the inaugural collection of data pertaining to King Salmon. This data is not only pivotal for the current analysis but also invaluable as a reference for future studies seeking to understand the baseline levels of various features in King Salmon. Consequently, each data instance is of paramount importance. Techniques that under-sample the majority class by removing instances can inadvertently lead to the

loss of crucial information. Similarly, methods that generate synthetic samples, exemplified by SMOTE[55], may precipitate overfitting, as the model becomes excessively tailored to the training data. Given that our preliminary results indicate a tendency towards overfitting in more than half of the datasets analyzed, the selection of appropriate methods warrants careful consideration and deliberation.

Anomaly Detection Anomaly detection[12] is another area that necessitates deeper investigation. Outliers, while ostensibly deviating from the norm, may encapsulate crucial, albeit infrequent, information instrumental in refining the predictive accuracy of the model. In this study, we do not remove any outliers. The main reason is that the absence of established reference levels impedes the effective identification of outliers, rendering traditional statistical methods, such as the Interquartile Range (IQR), somewhat constrained and limited in their applicability [54]. Consequently, a collaborative discourse with researchers affiliated with the Cawthron Institution is imperative to facilitate a more nuanced understanding and treatment of outliers within the dataset.

High dimensional Visualization Currently, the 2 dimension visualization seems not enough for our datasets like Figure 4.4, which could not separate the two classes well. The visualization techniques not only offer intuitive insights into intricate datasets but also facilitate the identification of patterns, clusters, and anomalies that might be obscured in higher dimensions. Thus, we should seek for high dimensional visualization methods.

Ensemble Feature Selection Methods Ensemble feature selection methods adeptly address the constraints inherent to individual feature selection algorithms that rely on disparate selection criteria, thereby generating feature subsets through the amalgamation of various feature selection outcomes. These ensemble techniques, which incorporate filter, wrapper, and embedded feature selection methodologies, have demonstrated superior performance compared to singular methods when applied to medical datasets[15, 4]. By harnessing the strengths of multiple selection criteria, ensemble feature selection methods offer a robust and comprehensive approach to identifying pivotal features, thereby enhancing the reliability and efficacy of the resultant predictive models.

Ensemble Evaluation The application of ensemble methods, specifically bagging and boosting, can be explored further to improve the imbalance in the dataset. These techniques, known for their ability to enhance model performance, can be particularly useful in addressing the challenges posed by complex datasets. For instance, Random Forests (RF), a prominent ensemble method, has garnered widespread application in the fields of biology and bioinformatics for tasks encompassing feature selection and classification[9]. Besides, the evolutionary computation approaches, such as the GA-based feature selection method for ensemble classifiers also worth trying[87].

These techniques offer promising potential for enhancing the robustness and accuracy of health prediction models for King Salmon, warranting further investigation and application in future research endeavors.

Bibliography

[1] ABEEL, T., HELLEPUTTE, T., VAN DE PEER, Y., DUPONT, P., AND SAEYS, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics 26*, 3 (2010), 392–398.

[2] ACUNA, E., AND RODRIGUEZ, C. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004* (2004), Springer, pp. 639–647.

[3] ALMUTIRI, T., AND SAEED, F. Review on feature selection methods for gene expression data classification. In *Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing 4* (2020), Springer, pp. 24–34.

[4] AMINI, F., HU, G., AND WANG, L. Application of the two-layer wrapper-embedded feature selection method to improve genomic selection. In *2022 17th Annual System of Systems Engineering Conference (SOSE)* (2022), IEEE, pp. 232–237.

[5] ARAÚJO, B. C., LOVETT, B., PREECE, M. A., BURDASS, M., SYMONDS, J. E., MILLER, M., WALKER, S. P., AND HEASMAN, K. G. Effects of different rations on production performance, spinal anomalies, and composition of chinook salmon (oncorhynchus tshawytscha) at different life stages. *Aquaculture 562* (2023), 738759.

[6] ASSEFA, A., ABUNNA, F., ET AL. Maintenance of fish health in aquaculture: review of epidemiological approaches for prevention and control of infectious disease of fish. *Veterinary medicine international 2018* (2018).

[7] BAAK, M., KOOPMAN, R., SNOEK, H., AND KLOUS, S. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis 152* (2020), 107043.

[8] BOLTANA, S., SANHUEZA, N., DONOSO, A., AGUILAR, A., CRESPO, D., VERGARA, D., ARRIAGADA, G., MORALES-LANGE, B., MERCADO, L., REY, S., ET AL. The expression of trpv channels, prostaglandin e2 and pro-inflammatory cytokines during behavioural fever in fish. *Brain, behavior, and immunity 71* (2018), 169–181.

[9] BOULESTEIX, A.-L., JANITZA, S., KRUPPA, J., AND KÖNIG, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*, 6 (2012), 493–507.

[10] CARABALLO, C., DESAI, N. R., MULDER, H., ALHANTI, B., WILSON, F. P., FIUZAT, M., FELKER, G. M., PIÑA, I. L., O'CONNOR, C. M., LINDENFELD, J., ET AL. Clinical

implications of the new york heart association classification. *Journal of the American Heart Association 8*, 23 (2019), e014240.

[11] CASANOVAS, P., WALKER, S. P., JOHNSTON, H., JOHNSTON, C., AND SYMONDS, J. E. Comparative assessment of blood biochemistry and haematology normal ranges between chinook salmon (oncorhynchus tshawytscha) from seawater and freshwater farms. *Aquaculture 537* (2021), 736464.

[12] CHANDOLA, V., BANERJEE, A., AND KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR) 41*, 3 (2009), 1–58.

[13] CHANDRASHEKAR, G., AND SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering 40*, 1 (2014), 16–28.

[14] CHAWLA, N. V., JAPKOWICZ, N., AND KOTCZ, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter 6*, 1 (2004), 1–6.

[15] CHEN, C.-W., TSAI, Y.-H., CHANG, F.-R., AND LIN, W.-C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems 37*, 5 (2020), e12553.

[16] CIESLAK, M. C., CASTELFRANCO, A. M., RONCALLI, V., LENZ, P. H., AND HARTLINE, D. K. t-distributed stochastic neighbor embedding (t-sne): A tool for eco-physiological transcriptomic analysis. *Marine genomics 51* (2020), 100723.

[17] COUCH, C. E., NEAL, W. T., HERRON, C. L., KENT, M. L., SCHRECK, C. B., AND PETERSON, J. T. Gut microbiome composition associates with corticosteroid treatment, morbidity, and senescence in chinook salmon (oncorhynchus tshawytscha). *Scientific Reports 13*, 1 (2023), 2567.

[18] DEEKSHATULU, B., CHANDRA, P., ET AL. Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia technology 10* (2013), 85–94.

[19] DISSANAYAKE, K., AND MD JOHAR, M. G. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Applied Computational Intelligence and Soft Computing 2021* (2021), 1–17.

[20] ELVY, J. E., SYMONDS, J. E., HILTON, Z., WALKER, S. P., TREMBLAY, L. A., CASANOVAS, P., AND HERBERT, N. A. The relationship of feed intake, growth, nutrient retention, and oxygen consumption to feed conversion ratio of farmed saltwater chinook salmon (oncorhynchus tshawytscha). *Aquaculture 554* (2022), 738184.

[21] ESMAEILI, N., CARTER, C. G., WILSON, R., WALKER, S. P., MILLER, M. R., BRIDLE, A. R., AND SYMONDS, J. E. Protein metabolism in the liver and white muscle is associated with feed efficiency in chinook salmon (oncorhynchus tshawytscha) reared in seawater: Evidence from proteomic analysis. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 42* (2022), 100994.

[22] FROESE, R. Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations. *Journal of applied ichthyology 22*, 4 (2006), 241–253.

[23] GARCÍA, S., LUENGO, J., AND HERRERA, F. *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[24] GARCÍA-PEDRAJAS, N., PÉREZ-RODRÍGUEZ, J., GARCÍA-PEDRAJAS, M., ORTIZ-BOYER, D., AND FYFE, C. Class imbalance methods for translation initiation site recognition in dna sequences. *Knowledge-Based Systems 25*, 1 (2012), 22–34.

[25] GEMS, D., KERN, C. C., NOUR, J., AND EZCURRA, M. Reproductive suicide: similar mechanisms of aging in c. elegans and pacific salmon. *Frontiers in Cell and Developmental Biology 9* (2021), 688788.

[26] GLENCROSS, B. D., BLYTH, D., BOURNE, N., CHEERS, S., IRVIN, S., AND WADE, N. M. An analysis of partial efficiencies of energy utilisation of different macronutrients by barramundi (lates calcarifer) shows that starch restricts protein utilisation in carnivorous fish. *British Journal of Nutrition 117*, 4 (2017), 500–510.

[27] GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine learning 46* (2002), 389–422.

[28] HAN, S., QUBO, C., AND MENG, H. Parameter selection in svm with rbf kernel function. In *World Automation Congress 2012* (2012), IEEE, pp. 1–4.

[29] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., AND FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[30] HINDAR, K., FLEMING, I. A., MCGINNITY, P., AND DISERUD, O. Genetic and ecological effects of salmon farming on wild salmon: modelling from experimental results. *ICES Journal of Marine Science 63*, 7 (2006), 1234–1247.

[31] HOOLE, D., LEWIS, J., SCHUWERACK, P., CHAKRAVARTHY, C., SHRIVE, A., GREENHOUGH, T., AND CARTWRIGHT, J. Inflammatory interactions in fish exposed to pollutants and parasites: a role for apoptosis and c reactive protein. *Parasitology 126*, 7 (2003), S71–S85.

[32] HUANG, X., ZHANG, L., WANG, B., LI, F., AND ZHANG, Z. Feature clustering based support vector machine recursive feature elimination for gene selection. *Applied Intelligence 48* (2018), 594–607.

[33] HUDA, S., YEARWOOD, J., JELINEK, H. F., HASSAN, M. M., FORTINO, G., AND BUCKLAND, M. A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE access 4* (2016), 9145–9154.

[34] INIESTA, R., STAHL, D., AND MCGUFFIN, P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological medicine 46*, 12 (2016), 2455–2465.

[35] ISWARI, N. M. S., ET AL. Fish freshness classification method based on fish image using k-nearest neighbor. In *2017 4th international conference on new media studies (CONMEDIA)* (2017), IEEE, pp. 87–91.

[36] JIN, X., XU, A., BIE, R., AND GUO, P. Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In *Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, April 9, 2006. Proceedings* (2006), Springer, pp. 106–115.

[37] JOHNE, A. S., CARTER, C. G., WOTHERSPOON, S., HADLEY, S., SYMONDS, J. E., WALKER, S. P., AND BLANCHARD, J. L. Modeling the effects of ration on individual growth of oncorhynchus tshawytscha under controlled conditions. *Journal of Fish Biology* (2023).

[38] JOHNSON, J. M., AND KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. *Journal of Big Data 6*, 1 (2019), 1–54.

[39] JORDAN, M. I., AND MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science 349*, 6245 (2015), 255–260.

[40] JOVIĆ, A., BRKIĆ, K., AND BOGUNOVIĆ, N. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (2015), Ieee, pp. 1200–1205.

[41] KANG, C., HUO, Y., XIN, L., TIAN, B., AND YU, B. Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine. *Journal of theoretical biology 463* (2019), 77–91.

[42] KAUR, H., PANNU, H. S., AND MALHI, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR) 52*, 4 (2019), 1–36.

[43] KHANJI, C., LALONDE, L., BAREIL, C., LUSSIER, M.-T., PERREAULT, S., AND SCHNITZER, M. E. Lasso regression for the prediction of intermediate outcomes related to cardiovascular disease prevention using the transit quality indicators. *Medical Care 57*, 1 (2019), 63–72.

[44] KIM, H.-Y. Statistical notes for clinical researchers: the independent samples t-test. *Restorative Dentistry & Endodontics 44*, 3 (2019).

[45] KING, A. P., AND ECKERSLEY, R. *Statistics for biomedical engineers and scientists: How to visualize and analyze data*. Academic Press, 2019.

[46] KOTSIANTIS, S. B., KANELLOPOULOS, D., AND PINTELAS, P. E. Data preprocessing for supervised leaning. *International journal of computer science 1*, 2 (2006), 111–117.

[47] KRASKOV, A., STÖGBAUER, H., AND GRASSBERGER, P. Estimating mutual information. *Physical review E 69*, 6 (2004), 066138.

[48] LIU, H., AND YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering 17*, 4 (2005), 491–502.

[49] MA, S., AND HUANG, J. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics 9*, 5 (2008), 392–403.

[50] MARTÍNEZ-PORCHAS, M., MARTÍNEZ-CÓRDOVA, L. R., AND RAMOS-ENRIQUEZ, R. Cortisol and glucose: reliable indicators of fish stress? *Pan-American Journal of Aquatic Sciences* (2009), 158–178.

[51] MATHEW, T. E. A logistic regression with recursive feature elimination model for breast cancer diagnosis. *International Journal on Emerging Technologies 10*, 3 (2019), 55–63.

[52] MCDONALD, M., SMITH, C., AND WALSH, P. The physiology and evolution of urea transport in fishes. *The Journal of membrane biology 212* (2006), 93–107.

[53] MENA, L. J., AND GONZALEZ, J. A. Machine learning for imbalanced datasets: Application in medical diagnostic. In *Flairs Conference* (2006), pp. 574–579.

[54] MILLER, J. N. Tutorial review—outliers in experimental data and their treatment. *Analyst 118*, 5 (1993), 455–461.

[55] MOHAMMED, A. J., HASSAN, M. M., AND KADIR, D. H. Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering 9*, 3 (2020), 3161–3172.

[56] NANDA, A., MOHAPATRA, B. B., MAHAPATRA, A. P. K., MAHAPATRA, A. P. K., AND MAHAPATRA, A. P. K. Multiple comparison test by tukey's honestly significant difference (hsd): Do the confident level control type i error. *International Journal of Statistics and Applied Mathematics 6*, 1 (2021), 59–65.

[57] NARGESIAN, F., SAMULOWITZ, H., KHURANA, U., KHALIL, E. B., AND TURAGA, D. S. Learning feature engineering for classification. In *Ijcai* (2017), vol. 17, pp. 2529–2535.

[58] NATIONAL INSTITUTE OF WATER AND ATMOSPHERIC RESEARCH. Chinook salmon, 2023. Accessed: 2023-07-21.

[59] NEW ZEALAND KING SALMON. Chinook salmon, 2023. Accessed: 2023-07-21.

[60] NGUYEN, B. H., XUE, B., AND ZHANG, M. A survey on swarm intelligence approaches to feature selection in data mining. *Swarm and Evolutionary Computation 54* (2020), 100663.

[61] NGUYEN, H. V., AND BYEON, H. Prediction of parkinson's disease depression using lime-based stacking ensemble model. *Mathematics 11*, 3 (2023), 708.

[62] ORIMO, H. The mechanism of mineralization and the role of alkaline phosphatase in health and disease. *Journal of Nippon Medical School 77*, 1 (2010), 4–12.

[63] OSTERTAGOVA, E., OSTERTAG, O., AND KOVÁČ, J. Methodology and application of the kruskal-wallis test. *Applied mechanics and materials 611* (2014), 115–120.

[64] PERRY, S. F. The chloride cell: structure and function in the gills of freshwater fishes. *Annual review of physiology 59*, 1 (1997), 325–347.

[65] POORE, J., AND NEMECEK, T. Reducing food's environmental impacts through producers and consumers. *Science 360*, 6392 (2018), 987–992.

[66] RICHHARIYA, B., TANVEER, M., RASHID, A. H., INITIATIVE, A. D. N., ET AL. Diagnosis of alzheimer's disease using universum support vector machine based recursive feature elimination (usvm-rfe). *Biomedical Signal Processing and Control 59* (2020), 101903.

[67] ROUDER, J. N., ENGELHARDT, C. R., MCCABE, S., AND MOREY, R. D. Model comparison in anova. *Psychonomic bulletin & review 23* (2016), 1779–1786.

[68] SABERIOON, M., CÍSAŘ, P., LABBÉ, L., SOUČEK, P., PELISSIER, P., AND KERNEIS, T. Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features. *Sensors 18*, 4 (2018), 1027.

[69] SADEGHI, J., CHAGANTI, S. R., AND HEATH, D. D. Regulation of host gene expression by gastrointestinal tract microbiota in chinook salmon (oncorhynchus tshawytscha). *Molecular Ecology* (2023).

[70] SAEYS, Y., INZA, I., AND LARRANAGA, P. A review of feature selection techniques in bioinformatics. *bioinformatics 23*, 19 (2007), 2507–2517.

[71] SALCEDO-SANZ, S., GHAMISI, P., PILES, M., WERNER, M., CUADRA, L., MORENO-MARTÍNEZ, A., IZQUIERDO-VERDIGUIER, E., MUÑOZ-MARÍ, J., MOSAVI, A., AND CAMPS-VALLS, G. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion 63* (2020), 256–272.

[72] SALMON, N. Z. K. New zealand king salmon annual report 2022, 2022. Accessed: 21 Oct 2023.

[73] SANTOS, R. D. O., GORGULHO, B. M., CASTRO, M. A. D., FISBERG, R. M., MARCHIONI, D. M., AND BALTAR, V. T. Principal component analysis and factor analysis: Differences and similarities in nutritional epidemiology application. *Revista Brasileira de Epidemiologia 22* (2019).

[74] SCHOLTENS, M., DODDS, K., WALKER, S., CLARKE, S., TATE, M., SLATTERY, T., PREECE, M., ARRATIA, L., AND SYMONDS, J. Opportunities for improving feed efficiency and spinal health in new zealand farmed chinook salmon (oncorhynchus tshawytscha) using genomic information. *Aquaculture 563* (2023), 738936.

[75] SENAN, E. M., AL-ADHAILEH, M. H., ALSAADE, F. W., ALDHYANI, T. H., ALQARNI, A. A., ALSHARIF, N., UDDIN, M. I., ALAHMADI, A. H., JADHAV, M. E., AND ALZAHRANI, M. Y. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering 2021* (2021).

[76] SINGH, V., POONIA, R. C., KUMAR, S., DASS, P., AGARWAL, P., BHATNAGAR, V., AND RAJA, L. Prediction of covid-19 corona virus pandemic based on time series data using support vector machine. *Journal of Discrete Mathematical Sciences and Cryptography 23*, 8 (2020), 1583–1597.

[77] STEINER, K., LAROCHE, O., WALKER, S. P., AND SYMONDS, J. E. Effects of water temperature on the gut microbiome and physiology of chinook salmon (oncorhynchus tshawytscha) reared in a freshwater recirculating system. *Aquaculture 560* (2022), 738529.

[78] STENTON-DOZEY, J. M., HEATH, P., REN, J. S., AND ZAMORA, L. N. New zealand aquaculture industry: research, opportunities and constraints for integrative multi-trophic farming. *New Zealand Journal of Marine and Freshwater Research 55*, 2 (2021), 265–285.

[79] THORSTAD, E. B., BLISS, D., BREAU, C., DAMON-RANDALL, K., SUNDT-HANSEN, L. E., HATFIELD, E. M., HORSBURGH, G., HANSEN, H., MAOILÉIDIGH, N. Ó., SHEEHAN, T., ET AL. Atlantic salmon in a rapidly changing environment—facing the challenges of reduced marine survival and climate change. *Aquatic Conservation: Marine and Freshwater Ecosystems 31*, 9 (2021), 2654–2665.

[80] URBANOWICZ, R. J., MEEKER, M., LA CAVA, W., OLSON, R. S., AND MOORE, J. H. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics 85* (2018), 189–203.

[81] WATKINSON, S. Life after death: the importance of salmon carcasses to british columbia's watersheds. *Arctic 53*, 1 (2000), 92–96.

[82] WEAVER, K. F., MORALES, V. C., DUNN, S. L., GODDE, K., AND WEAVER, P. F. *An introduction to statistical analysis in research: with applications in the biological and life sciences*. John Wiley & Sons, 2017.

[83] WERGEDAHL, H., LIASET, B., GUDBRANDSEN, O. A., LIED, E., ESPE, M., MUNA, Z., MØRK, S., AND BERGE, R. K. Fish protein hydrolysate reduces plasma total cholesterol, increases the proportion of hdl cholesterol, and lowers acyl-coa: cholesterol acyltransferase activity in liver of zucker rats. *The Journal of nutrition 134*, 6 (2004), 1320–1327.

[84] WOLD, S., ESBENSEN, K., AND GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52.

[85] WU, H. A deep learning-based hybrid feature selection approach for cancer diagnosis. In *Journal of Physics: Conference Series* (2021), vol. 1848, IOP Publishing, p. 012019.

[86] WUERTZ, S., AND REISER, S. Creatine: A valuable supplement in aquafeeds? *Reviews in Aquaculture 15*, 1 (2023), 292–304.

[87] XUE, B., ZHANG, M., BROWNE, W. N., AND YAO, X. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation 20*, 4 (2015), 606–626.

[88] YOUNG, T., LAROCHE, O., WALKER, S. P., MILLER, M. R., CASANOVAS, P., STEINER, K., ESMAEILI, N., ZHAO, R., BOWMAN, J. P., WILSON, R., ET AL. Prediction of feed efficiency and performance-based traits in fish via integration of multiple omics and clinical covariates. *Biology 12*, 8 (2023), 1135.

[89] YU, C. H. Resampling methods: concepts, applications, and justification. *Practical Assessment, Research, and Evaluation 8*, 1 (2002), 19.

[90] ZEALAND, F. N. Lca of king salmon from new zealand.

[91] ZHAO, R., SYMONDS, J. E., WALKER, S. P., STEINER, K., CARTER, C. G., BOWMAN, J. P., AND NOWAK, B. F. Relationship between gut microbiota and chinook salmon (oncorhynchus tshawytscha) health and growth performance in freshwater recirculating aquaculture systems. *Frontiers in Microbiology 14* (2023), 1065823.