**INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND COMPUTER SYSTEMS**
E-ISSN: 2180-0650
VOLUME 10, ISSUE 2, 2024, pp. 173 – 186
DOI: https://doi.org/10.15282/ijsecs.10.2.2024.14.0132

**RESEARCH ARTICLE**

# Hybrid Learning for Bottlenose Dolphin Motion Approximation Using Weighted Polling KNN and Bayes Mechanism

**Farid Morsidi**

Computing Department, Faculty of Computing & Meta-Technology, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia

**ABSTRACT -** Artificial intelligence has been extensively employed to examine social animal behaviours, such as feeding habits, reproduction, swarm structures, and hibernation. For marine biologists, grasping essential habitat characteristics like swimming depth and temperature preferences is vital for evaluating ecological boundaries, habitat deterioration, and conservation initiatives. Machine learning has been used to help identify bottlenose dolphins and their cohabitation patterns, as computer vision and deep learning are often used in studies of fish habitats involving predation, feeding habits, breeding, and other fish. Existing techniques do not provide sufficient variety in estimating classification characteristics for assigning unique animals to their corresponding subclasses, particularly for differentiating various marine species along with their respective groups. The objective for this proposed research framework is aimed at integrating supervised clustering and probabilistic classifiers such as K-Nearest Neighbors (KNN) or Bayes classifying with weighted voting to classify dolphin types, apart from consolidated behavioral patterns. The study further integrated the use of oxygen usage into the dataset for monitoring movement activity after pup pre-feeding intervals in addition to metabolic rates of other animals. Data was cross verified, encompassing trial laps, dolphin identifiers, timing, segment length, and carbon dioxide emission rates, for the purpose of classification. Findings showed an average classification accuracy of 97% for weighted KNN types and 91% for weighted Bayes classifiers, emphasizing their effectiveness in differentiating dolphin classes in swarming settings. These machine learning methods provide essential tools for enhancing marine research and aiding conservation and habitat preservation through improved identification of ecological traits.

**ARTICLE HISTORY**
Received : 11 December 2024
Revised : 6 March 2025
Accepted : 17 March 2025
Published : 24 March 2025

**KEYWORDS**
*Probability classification Supervised learning*
*Clustering*
*KNN*
*Machine learning*

## 1.0 INTRODUCTION

Probabilistic clustering has evolved into a robust method in data science for classifying data points according to probabilistic connections among various clusters. In contrast to conventional techniques such as K-Means, which allocate data points to a single cluster [1], probabilistic clustering models view clusters as probability distributions. This adaptability enables every data point to be associated with unique probability scores for each cluster, yielding greater understanding of intricate datasets [2]. For example, in genetics, this method reveals gene expression patterns across diverse conditions, aiding researchers in understanding how genes co-express within multiple pathways and spotting disease biomarkers. Furthermore, probabilistic clustering is essential for customer segmentation in e-commerce and enhances targeted marketing approaches by allowing users to be classified into overlapping groups such as regular buyers and seasonal shoppers. This naturally includes customers who exhibit different behaviours and improves personalization. Recommender systems provide a more adaptive method by helping to recommend products based on the probabilistic preferences of users across multiple interest groups. This strategy is useful for dealing with noisy data, as seen in financial fraud detection to detect suspicious transactions and environmental monitoring that uses sensor information to find pollution sources. Probabilistic clustering finds application in healthcare, enabling natural language processing to analyze patient information by disease subtype in the realm of precision medicine. This approach also helps in structuring text into intersecting themes relevant to social media and extensive dataset analysis. In general, the variety and influence of probabilistic clustering offer significant insights across various domains, enabling decision-makers to assess data more effectively and obtain valuable information from intricate datasets.

Supervised clustering is crucial in automated decision making and data analysis because it improves clustering accuracy by leveraging domain-specific knowledge [3-4]. By combining labelled data with traditional clustering methods, the accuracy of supervised approaches is combined with the exploratory advantages of unsupervised learning.

***CORRESPONDING AUTHOR | F. Morsidi** | ✉ farid_m2mfan@yahoo.com.my

This integrated approach includes partial tags as a data point, enhancing cluster relevance and reliability, and also includes clear and important groups of expertise. Furthermore, supervised clustering improves text classification and increases the effectiveness of information retrieval and recommendation systems [3-5]. In the field of fraud detection, the analysis of labelled transaction data is essential for identifying fraudulent activities, especially for monitoring cases of credit card fraud. This approach can also help supply chain management by organizing warehouses and production facilities to optimize logistics and inventory management [6]. In addition, supervised clustering is used in ecological research to classify wildlife using limited behavioral data, which helps conservation efforts by identifying critical habitats [7-8]. Overall, this approach improves on traditional clustering techniques by combining labelled data and expert opinion, providing valuable insights across a range of fields, including healthcare, social sciences and environmental research.

Although each approach has certain advantages, the integration of probabilistic classification and supervised clustering has yet to be thoroughly explored, especially when solving complex domain-specific problems that require both predictability and interpretability. Probabilistic models are particularly effective at recognizing unique data points in noisy, overlapping or uncertain datasets by assessing the probability that each point belongs to various clusters [9-10]. On the contrary, the controlled clustering improves data placement and clarity using the labelled data specific to the domain to manage the clustering procedure [11-12]. This cooperation can overcome the restrictions inherent in each approach when used separately. For example, in the field of online retail customer segmentation, probabilistic clustering can successfully identify overlapping customer groups such as "bargain hunters" and "occasional shoppers" based on their purchasing patterns. But these clusters can be further enhanced by combining supervised clustering with imperfect behavioral tags such as purchase frequency, brand loyalty and browsing patterns. This improved segmentation allows companies to develop more precise marketing strategies. For example, promotions can be customized for consumers who are more willing to buy but are more price sensitive, while premium buyers can be engaged with personalized offers. Similarly, in healthcare, probabilistic classification is often applied to distinguish patient subtypes from gene expression information and medical images. Researchers can use supervised clustering of labelled clinical findings and partial diagnostic indicators to improve classification of disease subtypes [13], a development that could lead to better predictions of treatment response and disease progression in conditions such as cancer and autoimmune diseases. Probabilistic clustering techniques have been used in environmental monitoring to identify pollution hotspots through analysis of sensor data, including variables such as temperature, pH levels and particle concentrations. However, integrating supervised clustering with labelled environmental data such as pollution sources and weather conditions can significantly improve understanding of pollution dispersion patterns. This thorough approach enables researchers to identify the precise locations and timings where environmental threats are most severe, aiding in the formulation of targeted response plans.

In ecology, probabilistic classification helps analyze species and habitats by modeling shared characteristics like migration paths and feeding zones [14-15]. Integrating supervised clustering with labelled behavioral data such as predator-prey interactions or cross-seasonal mating patterns can help researchers more effectively identify groups of species [15], improving conservation efforts for threatened and endangered species. Moreover, the integration of these methods has proven very promising in terms of fraud detection: while probabilistic models are successful in detecting anomalous transactions by identifying deviations from expected patterns, supervised clustering can improve detection accuracy by using labelled data such as confirmed fraudulent activity or suspicious account behavior. This comprehensive model is designed to minimize false positives while simultaneously detecting emerging fraud patterns with greater flexibility, thereby enhancing its significance in financial security frameworks. However, current studies have largely concentrated on probabilistic classification and supervised clustering as separate entities.

This study aims to provide contribution to fill an existing gap in baseline classification measures for marine biology by investigating the integration of probabilistic classification and supervised clustering adapted to specific field applications, especially in terms of performing intelligent classification for dominant features within a clustering domain. Emphasis is placed on the classification of bottlenose dolphins based on characteristic behavioral traits such as energy expenditure during migration, breathing patterns and swimming behavior. By utilizing weighted polling techniques, this research advances the capabilities of KNN and Bayes classification algorithms. This study promotes the objective of utilizing the advantages of the proposed methods as a measure to identify the most accurate classification traits from the training dataset in relation to recognized output values for dolphin classes. This method underscores that hybrid methodologies can yield more precise and dependable insights. Initial research indicates that the integrated methodology significantly improves prediction rates and enhances accuracy in distinguishing behavioral groups relative to independent models, which is an encouraging indication that this hybrid approach is versatile enough to handle complex and cross-platform datasets. The results are anticipated to enhance the progress of machine learning techniques by presenting effective solutions for challenging issues that demand predictive capability and flexibility in data analysis.

The paper's format is as follows: Section 2 provides interdisciplinary reviews on related topics such as probability clustering and supervised learning applications in automated classification, with emphasis on their implementation for industrial and simulation objectives; Section 3 describes the research methodology, which includes classification prediction techniques applied to the bottlenose dolphin locomotive dataset, including behavioral patterns, characteristics, application strategies, and the importance of each classification algorithm in the analysis process; Section 4 presents a discourse analysis that evaluates the effectiveness of the proposed method in enhancing probability classification and integrated learning about feature identification, where Section 5 concludes the discussion with suggestions on future

endeavors for the assessed fields involving subjugation of probability classification and supervised learning in feature classification.

## 2.0 RELATED WORKS

Supervised learning and probability classification are distinct techniques utilized in data clustering and intelligent categorization within a particular domain [7], [16], [17]. Their main goal is to identify and enhance the input mapping of data points to their corresponding outputs. Intelligent systems are being utilized in marine ecosystems to monitor ecological and habitat conditions using video feed and live satellite images [4], [16], [18]. Supervised learning and probability classification can both break down input data, like animal features and habitat characteristics, into layers of varying levels of abstraction to help understand the data and define complex features for labelling training examples with new inputs [9], [14]. This type of implementation is commonly found in computer vision and deep learning, with researchers trying to develop decision-making tools to address habitat-related problems and improve current marine life monitoring systems [16]. When working with nonparametric data, supervised learning and probability classification can help enhance mapping for advanced data processing, specifically in identifying recurring and synonymous data entries to correct discrepancies in the predicted output compared to the target value [14]. Marine scientists utilize modern facilities such as remote underwater cameras and videos to study fish in their natural habitats, helping them understand and predict their reactions to climate change, loss of habitat, and fishing [7], [16]. However, analysing extensive data such as acoustic and observer data to find useful information may take a significant amount of time [17]. Deep learning technology assists marine biologists in swiftly analysing extensive comparative datasets, revealing fresh insights that cannot be obtained through traditional manual monitoring methods [4], [16]. For instance, a review study explored the fundamental underworking of fish classification in underwater environments utilizing computer vision and deep learning methods from 2003 to 2021 [16]. Among the key features discussed in the research context explores key ideas for marine animal trait classifications, obstacles in applying deep learning to analyze underwater images, and offers perspectives on research tracking oceanic environments and how both supervised learning and probability classification had been improvised to enhance current existing methodologies.

Different techniques for supervised clustering and probability classification are used in the field of marine biology [2], [7], [19]. An example of literary research on a combination of both supervised clustering and probability classification includes creating a categorization system that organizes four types of edible fish by examining texture extraction and analysing colour properties [4]. The K-nearest neighbor (KNN) algorithm is used in this system for probabilistic classification which is a very widely used nonparametric technique in statistical pattern recognition, with fish meat and fish scale being key features for classification. The HVs colour model and GLCM technique are used for meat image analysis. The study effectively illustrates the application of the KNN algorithm; however, it does not provide an in-depth analysis of its scalability and computational efficiency when utilized with larger datasets or more intricate fish classification tasks. Additionally, focusing exclusively on specific features such as texture and colour may restrict its effectiveness across various marine species or in diverse environmental contexts. The KNN classifier served as the primary algorithm for classification in the evaluation dataset. The results showed an accuracy of 90% for tilapia meat, 97.5% for mackerel meat, 87.5% for tilapia scales and 95% for mackerel scales. Despite these good results, the study ignored the potential impact of noise and variations in capture conditions on the performance of the classifier, which could have a significant impact on its practical application. Another work on supervised clustering has performed an in-depth analysis of the problem of classification using the KNN rule, and one issue related to this rule, is the sensitivity to the size of the neighborhood parameter, $k$ [20]. Aiming to address this issue and to increase classification effectiveness, an improved dual-weighted voting scheme for KNN is presented and developed [20]. This method mitigates the impact of k by employing a dual-weighted voting function for the k nearest neighbors of the target object. However, the evaluation is constrained by a small amount of artificial data, with real datasets only serving as benchmarks. To enhance the validation of this approach, it would be beneficial to apply it to a wider range of datasets, especially those characterized by high-dimensional features, to assess its robustness and broader applicability. The results suggest that the proposed classifier is a good algorithm for many real-life estimation and pattern recognition problems because it yields a satisfactory performance even when $k$ is varied over a wide range. However, the study could benefit from a comparative analysis against other state-of-the-art classifiers to provide a clearer benchmark of its performance. One study combined the traits from both probability classification and supervised clustering in the evaluation of various types of fish classification that were applied to fish landing records between 2005-2019 which also include those from the eastern region of Peninsular Malaysia [8]. The classifiers that have been used in the study are sequential minimum optimization (SMO), naïve Bayes (NB), multi-layer perceptron (MLP), instance-based k-nearest neighbor (IBK), and random forest (RF). The classification accuracy and the confusion matrix measures are used in this study, while multi-classification methods are also integrated to increase the performance of each classifier. The testing instances illustrate that the most favourable results are obtained from the following combinations: RF + SMO + NB + MLP and SMO + RF + NB + MLP. The combination and multipliers of single classifiers render an accuracy of 80.95 percent. This indicates that multi-classifier approaches in fish classification systems should be looked into for further research and practical application [8]. The study does not address the issues of computational overhead and scalability associated with the integration of multiple classifiers, which could restrict its applicability in large-scale or real-time scenarios. Another attempt at classifying characteristics of marine animals based on probability classification had been made on the evaluation of fish's freshness by examining digital images and evaluating the success rate upon implementation of a supervised learning approach in sorting mechanics [21].

The procedure consists of cropping, segmentation and capturing RGB values followed by training and testing a dataset of 210 images categorized into three categories: fresh, deteriorating and fully deteriorating. Experimental results show that the naive Bayes algorithm effectively evaluates the freshness level of fish based on fisheye images, achieving a test accuracy of 79.37%. This study highlights the importance of understanding the fish market distribution and the future benefits of applying this approach. The study underscores the potential of this method to enhance fish market distribution strategies; however, it overlooks the limitations associated with solely depending on RGB values for classification. For instance, the incorporation of supplementary sensory data or advanced imaging techniques could improve accuracy and provide a more robust evaluation of fish freshness. This research illustrates the importance of understanding fish market distribution and the prospective benefits of implementing this approach, while also encouraging the investigation of improvements in feature selection and integration with alternative classification systems.

Categorization is an essential method of data analysis that is employed to create models from incoming data or forecast future patterns. The Naive Bayesian classifier is a method utilized to forecast the likelihood of a sample being part of a specific class, under the assumption of attributes being conditionally independent [18]. The presumption condition in which the condition is fulfilled would determine the proficiency of the systems' performance as compared in tandem with other similar classifier implementations. An instance of the Bayes classifier implementation in feature prediction is used to train a set of fish data and receive test data input from users [9]. This approach employs a model based on probability, prior knowledge, and observed data to determine the fish's safety for consumption. The accuracy of testing data is determined by the system through the holdout method. This study showcases the Naïve Bayesian classifier's success in efficiently and effectively addressing the classification challenge of fish, specifically when dealing with extensive datasets. However, the study exhibits lacking in terms of exploring how variations in data quality or feature relevance might impact classifier accuracy. Additionally, the reliance on the holdout method for accuracy testing is seen possible with cross-validation techniques to ensure robustness across different datasets. The proposed system produces precise results, with many results exceeding the anticipated optimal percentage. Another related study in probabilistic classification suggested a fish recognition system that utilizes both local and global features in fish images [18]. Local characteristics are obtained through the utilization of Local Binary Pattern (LBP), Speeded-Up Robust Feature (SURF), and Scale Invariant Feature Transform (SIFT), whereas global characteristics are obtained via Colour Coherence Vector (CCV). Five popular machine learning algorithms such as Decision Tree, k-Nearest Neighbor (KNN), Support Vector Machines (SVM), Naïve Bayes, and Artificial Neural Network (ANN) are utilized to predict fish species, and the ultimate classification is decided by a majority vote. Examination of a segment of the *QUT_fish_data* dataset revealed an accurate rate of 98.46%, establishing it as a formidable competitor in the field. Despite the impressive accuracy, the study would benefit from an analysis of the computational complexity of integrating local and global features, as well as a comparison with simpler, feature-specific models to evaluate trade-offs between accuracy and efficiency.

## 3.0 METHODS AND MATERIAL

This section examines the classification prediction techniques utilized on the selected dataset showing the behavioral patterns of the targeted sample traits for marine animals, selected for this purpose in the bottlenose dolphin locomotive data annotation. The discussion centers on describing the features, techniques for application, advantages, and importance of every classification algorithm in the context of classification analysis. This study utilizes weighted polling in both the KNN algorithm and Bayes classification to make slight improvements to the existing frameworks of both algorithms. The study scope examines the advantages of adding weighted polling methods as an aspiration criterion to choose the most accurate classification traits from the training dataset compared to the known output for each dolphin class in the testing domain. Figure 1 on the following page illustrates the stages involved in the classification approach employed in this study for bottlenose dolphins, utilizing the proposed probability classification method through weighted polling metrics for both K-Nearest Neighbour and naive Bayes classifiers. This approach encompasses three primary categories, starting with classification algorithms and polling methods, feature extractions, and the generation of independently forecasted dolphin classes. Regarding the category that involves the application of a weighted polling mechanism for probability classification, various modifications of a similar algorithm were established as comparison metrics across each iteration, including standard forms, weighted forms, and weighted-squared iterations.

### 3.1 Implementation of Probabilistic Classifiers with Supervised Learning

This portion of the discussion focuses on utilizing a classification strategy that combines a probabilistic classifier with supervised learning to identify and categorize the optimal dolphin class sample in the tested clusters. Weighted polling measures are used to enhance the basic aggregation strategy for feature classification.

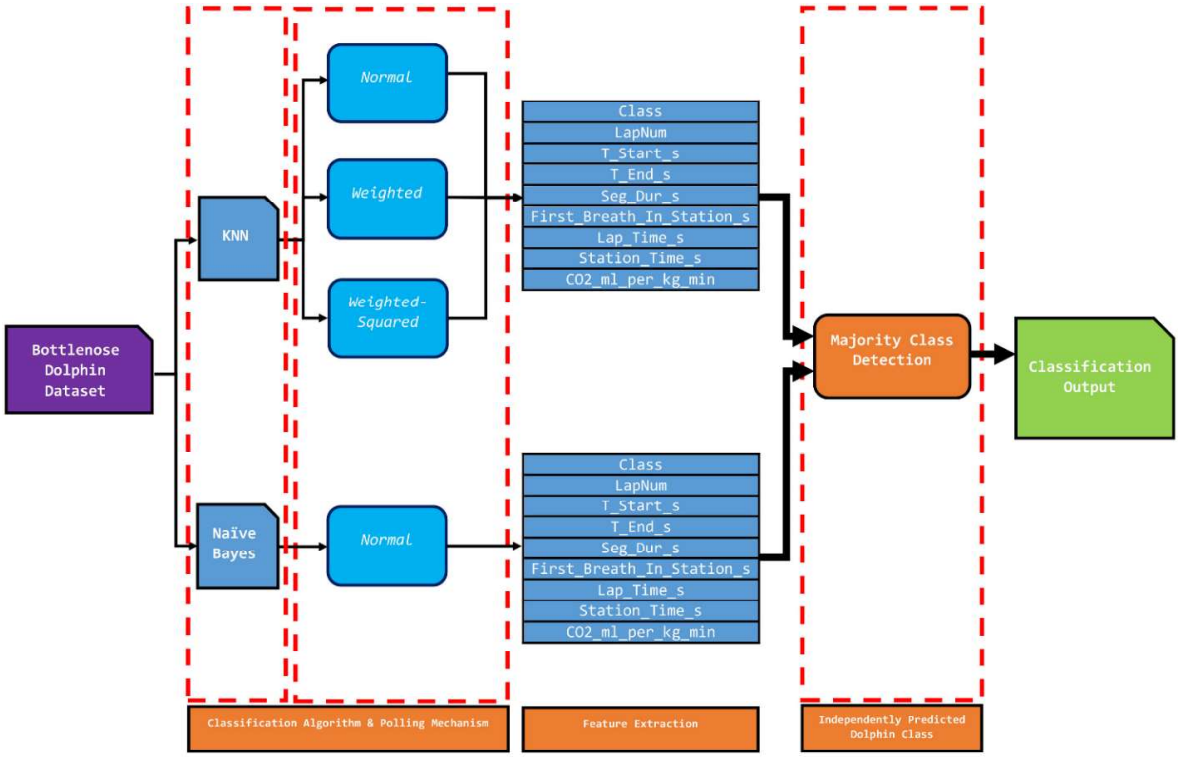


Figure 1. Framework for the implemented classification strategy for bottlenose dolphin dataset classification using KNN & naïve Bayes' weighted polling

### 3.1.1 Dataset Description

This study attempts to perform probabilistic classification on marine animal activities, particularly regarding their swimming patterns, locomotion, rate of metabolism among submerging/emerging habits, and oxygen consumption within each submersion. Several data classes belonging to tuna, marlin, and dolphin had been analyzed by the identified prediction traits, however, the scarcity of open-source datasets had limited the selection of compatible testing instances. This study improvised the dataset categories obtained from bottlenose dolphin locomotive activities [15] and identified accordingly major characteristics to be included during the predictive classification process as highlighted below. The dataset classes contain dolphin locomotive activities that are measured for oxygen consumption and resting metabolic rate which are post-prandial for over 12 hours within the interval of subsequent measurements. The training datasets are devised via cross-matching activity data for categorizing the proliferation of dolphins during trial laps (starting/ending), distinctive dolphin ID, the time index in which the lap is initiated, the time index where the next lap is scheduled, the segment duration interval between laps, first breadth at station time index, total swimming duration, station transit intervals, along with carbon dioxide production rate. The prefix of the dataset numeration is annotated as follows.

Table 1. The participating variables for bottlenose dolphin dataset instances for all subset dimensionality

| Variable List | Description |
|---|---|
| *Class* | The identifier used to classify each distinctive dolphin categories |
| *LapNum* | The collective accumulation of elapsed laps without factoring in the last lap that had been omitted pre-emptively |
| *T_Start_s* | The time index accrued from initiated and ongoing laps in which the dolphin had swam through |
| *T_End_s* | The time index represents the current departing point where the dolphin would continue its swim. |
| *Seg_Dur_s* | The accrued duration representing the entire segment involving the traversed lap and the arrival time at the destination station |

| | |
|---|---|
| *First_Breath_In_Station_s* | The collective time index represents the initial breath after consuming a lap, implying the lap section between the traversal among the partaking stations. |
| *Lap_Time_s* | The accumulated time representing the entire dolphin swim session |
| *Station_Time_s* | Total time accrued when being positioned at the station |
| *CO2_ml_per_kg_min* | The metric summing the production rate of carbon dioxide in the rate of milliliters CO2 per kg per min, evaluated by dividing the overall CO2 production during the full dolphin swimming trip duration |

Table 2. Dataset reference on the locomotive traits for each partaking bottlenose dolphin category

| ID | Body mass (kg) | Length (cm) | Age (yr) | RMR (ml $O_2$ $min^{-1}$) | *N* | Resting $\dot{V}_{CO2}$ (ml $min^{-1}$) | S | Lap (s) | Station (s) |
|---|---|---|---|---|---|---|---|---|---|
| *83H1* | 141–146 | 234 | 10 | 583±20 | 12 | 513±194 | 8 | 24, 22–26 | 18, 12–31 |
| *01L5* | 149–156 | 239 | 23 | 430±20 | 7 | 352±17 | 9 | 32, 28–36 | 20, 16–26 |
| *63H4* | 177 | 254 | 27 | 495±171 | 4 | 467±177 | NA | NA | NA |
| *9ON6* | 186–190 | 249 | 21 | 577±262 | 11a | 474±212 | 9 | 25, 24–26 | 13, 10–21 |
| *6JK5* | 209–210 | 259 | 24 | 388±222 | 13a | 343±182 | 6 | 28, 25–31 | 21, 16–26 |
| *9FL3* | 243–247 | 274 | 34b | 608±410 | 9 | 489±344 | 3 | 20, 20–21 | 22, 19–28 |

### 3.1.2 Dataset Training/Testing

Although training and testing data are designated at a 4:1 ratio (80%:20%), the output was assessed relative to the accumulated average accuracy rate. The rationale of selecting a 4:1 ratio (80:20) dataset split for training and testing is a widely accepted standard in machine learning that seemingly provides a good balance in terms of learning versus evaluation of the model. 80% of the data for training ensures that the model gets a fair bit of information to learn some meaningful patterns, thereby reducing underfitting. The subsequent remaining 20% is left for testing just to provide a fair estimate of the model's performance on the unseen data. This ratio maintains an ideal bias-variance equation and avoids overfitting or underfitting. This also acts like a standard metric that allows comparison of model performance more simply across tasks and applications. The 80:20 ratio is especially suitable for a dataset with many points so that sufficient training and test data points are available for model learning and performance evaluation respectively.

### 3.1.3 Execution Strategy

*1. K-Nearest Neighbour (KNN)*

KNN is a non-parametric machine learning technique classified as supervised learning pioneered by Evelyn Fix and Joseph Hodges in 1951 [22], [23], and a revered dynamically referenced approach in classification and regressive analysis since KNN refrains from *a priori* knowledge about the distribution of data. The progressive dynamic nature of KNN enables the management of alphanumerical data subtypes, ensuring it is a workable selection for multiple instances of dataset categories [4]. Among the feature traits of KNN is a least sensitive exposure to outlier imposition in opposition to myriads of similar data clustering algorithms [19]. KNN mechanics initiate with the scouring of the distinct number of the neighbor cluster within the domain relative to the supplied data point, where the distance metric is aggregated by methods specified in the classification category, for instance, the most widely applied method to calculate distance is Euclidean, Manhattan, and Minkowski which is gaining traction in recent times [10], [14], [24]. The equation representing the formulation of KNN mechanics is incorporated as follows in Equation 1 [14]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

In terms of deciding on the classification of each data point, feature extraction is done by pre-processing criteria such as weighting and dimension reduction to attempt to elevate the accuracy as well as possible, with the assumption to predicate the algorithm's adaptability to various clustering patterns and anticipate viabilities for the local data structure [3]. The governing rule for KNN is to designate an unclassified object to a binding class about its *k* nearest neighbors within the domains of the training set. The main distinction stemming from this trait is the almost guaranteed preferable asymptotic performance in terms of infinite optimal Bayesian error rate where this value would consecutively double should the independent value of *k* be designated as 1 [25]. The weighted KNN rule had been proposed to address the trade-offs of contemporary nearest neighbor classification where weighted approximations could bridge the clusters to the queried objects more adjacent and precisely [3], [24]. In terms of determining the sensitivity of *k* neighborhood size [7], [8], any selection of the values depicting the optimal *k* would directly influence the output of the probabilistic classification where a minute value would produce below-average local estimates, and vice versa. Feature extraction is imposed as a scaffold to complement the deficiencies of KNN in terms of improving the estimation of output performance even when there are affluent factors such as data points that are not gleaned (mislabelled, vague, containing too many outliers).

*2. Naïve Bayes*

The Naive Bayesian Classifier is a well-known and effective classification method commonly employed in supervised learning [5], [8], [18], [26]. Naive Bayes models are commonly employed in machine learning due to their efficiency and simplicity. They belong to the classification methods category that utilizes Bayes' Theorem [8], [18]. One of the unique ways Bayesian classifies data is by assuming that data attributes are independent of the data class, which is advantageous for handling datasets with unclear and noisy input data [9]. By examining each pair of features, these classifiers develop simple yet effective classification methods. Due to their high-dimensional data, Bayesian classifications are especially beneficial in tasks like text categorization, spam filtering, sentiment analysis, and rating classification. Naïve Bayes operates as a classifier by considering probabilities and assuming that the features in the model are unrelated [5], [8], [18]. In the representation of Bayes' theorem, the property of *j* classes is denoted in an array of$y_i$, $j \in \{1, \cdots, j\}$ whereas the X is denoted by $X = (X_1, X_2, \cdots X_m)$ consisted of *m* properties. The most common formulation assumption for representing the Bayesian theorem is denoted as follows in Equation 2 [14].

$$p(y_j I X) = \frac{p(X|y_j)p(y_j)}{p(X)} \tag{2}$$

This indicates that each characteristic impacts the forecasts individually, a situation uncommon in data forecasting. Even though this belief is frequently proven wrong, recent studies have shown its practical success and challenges in systematic enhancement. Bayesian learning algorithms use both prior knowledge and training data to estimate the posterior probability of a hypothesis [9]. The algorithm utilizes Bayes theorem during both training and prediction, making it a beneficial tool for tackling classification issues and improving machine learning models [14], [21]. Bayesian learning algorithms use both prior knowledge and training data to estimate the posterior probability of a hypothesis.

### 3.1.4 Polling Cycle

The main objective of the K-Nearest Neighbors (KNN) polling stage is to determine the predominant class within the nearest neighbors for each data point needing classification. Distances to all points in the training set are calculated, the nearest neighbors are determined, and the class label with the highest vote is selected [2], [7], [19], [27]. This repetitive pattern can consume a lot of computational power [7], particularly when dealing with extensive datasets or data with many dimensions. Distance weighting, KD-trees, and dimensionality reduction methods are utilized to enhance this process [8], [17], [28]. The optimal *k* varies based on the data, with higher values decreasing noise but blurring class boundaries. Optimizing hyper-parameters can help identify an optimal value for *k* [8], [28]. Weighted KNN and Weighted Naive Bayes are popular techniques in the field of probability classification because they improve the predictive accuracy of the assignment of importance to relevant data points. In weighted k-NN, the closer neighbors have more influence on the decision that leads to improved classification, especially in cases with complex datasets and lessening the noise impact. Conversely, Nave Bayes with weighted models are refined and modify probabilities to match the significance given to specific features, resulting in better performance when dealing with unbalanced and correlated data. Both methods are effective classifiers in the presence of noise and better manage class imbalance due to their weight assignment; as a result, they are already used in various fields, including fraud detection, medical diagnosis, and marine animal ecology studies. In marine ecology, they aid in species classification, tracking migration patterns, and predicting the suitability of habitats using environmental factors for conservation and biodiversity management. Because they adapt well and show varied behavior, they can be classified as one of the approaches much needed for real-life classification.

Table 3. Polling methods with weighting formulation for the tested classification approaches

| Classification | Approach | Elaboration |
|---|---|---|
| K-Nearest Neighbour | *Normal* | $\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n)\ + 1$ |
| | *Weighted* | $\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n) + (samples - n)$ |
| | *Weight-Squared* | $\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n) + (samples - n)^2$ |
| Bayes | | $\frac{Prevalence\ of\ classifications\ in\ NN}{Prevalence\ of\ classification\ for\ the\ respective\ test\ set}$ |

Table 3 shows the fundamental weighted polling approaches improvised in this study. The polling cycle's precision can be significantly reduced by noisy or irrelevant characteristics or inconsistent feature measurements [3], [7]. Naive Bayes utilizes probability calculations in place of a fundamental polling cycle such as K-Nearest Neighbors. By utilizing Bayes' theorem and assuming features are independent, it calculates the probability of each test instance belonging to each class. The chosen class is the one with the highest probability. This process can be done for each test case in applications that need real-time monitoring, just like a polling cycle. Precomputing prior probabilities and feature likelihoods can speed up calculations, especially for large datasets or frequent predictions [16], [18].

### 3.1.5 Evaluation Analysis

For the proposed methodology, the performance of the probability classifier on bottlenose dolphin data instances is evaluated in concurrent execution to further analyze the proficiency of weighted polling on both appropriate clustering approaches applied in this study (KNN & Bayes). Selection of the best accuracy is derived when the algorithm is executed simultaneously on the training dataset, where the motivation of comparing the polling rate from all 4 algorithm variants is to exemplify the aptitude of the model performance. Within the conjecture of the testing instance, it is presumed that the execution strategy is heuristic and nonparametric, thus no singular classifier would be biased over other participating variants in the same testing class. The systemic evaluation predicated the prediction traits to hover around 80-90% with every recurrent ascent in the value of *k* dimensionality. The accuracy rate determines the proficiency of the weighted polling methods in comparison with the conventional implementation of similar algorithm variants on a similar purpose in terms of properly classifying data points subjected to the actual foundational ground truth dataset. The general rule of thumb for classification has deemed the lowest error rate to possess an opposite aptitude with accuracy, where the higher accuracy rate would in turn produce a lower error rate.

#### 3.1.5.1 Accuracy

In probabilistic classification and supervised learning, the model's output conveys the specific characteristic of a particular data instance associated with a certain class, where threshold values are applied to transform the classification probabilities into a single or multiple prediction features [18]. In the context of research, accuracy in assessing the skill of probability classification is defined as the ratio of correctly labelled instances compared to the total number of instances in the dataset's population [8], [29]. Assessing accuracy is significant in terms of enhancing the overall effectiveness of the model, in addition to serving as a signal of the even distribution of class clusters within that specific dataset division. The accuracy rate is also less complicated to assess and offers a basic benchmark for comparing the old and updated datasets. There are numerous identifiable uses of accuracy in evaluation metrics for assessing the performance of the unique system, in that (i) *the speed at which accuracy builds influences performance across all prediction instances regardless of the probability threshold* [14], [18], [30], (ii) *establishes a definite baseline representing the model's total prediction count* [6], and (iii) *generalizes the optimal probability of population sampling in representation of the majority class, unaffected by the size of minority class predictions* [24]. In this research scope, the accuracy rate is considered the most effective metric to depict the degree of balance among the different class distributions of bottlenose dolphin swarms, aiding in selecting the optimal sample to represent the whole data cluster, while also delivering the most insightful statistical analysis for individual dolphin locomotion patterns in identifying their environmental behaviours and distinct actions. The fundamental formulation representation for evaluating accuracy rate is deemed as shown in Equation 3:

$$\text{Accuracy } = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

In data science, the accuracy metric is employed to evaluate classification models. Accuracy is defined as the ratio of correctly predicted instances to total cases [4], [30], [31]. This criterion is often used in classification systems because of its simplicity, ease-of-use, and clearness. The accuracy rate within classification processes is determined by using true positives and false negatives divided by all predicted and actual positive and negative instances. By measuring the overall performance of a dataset with balanced models, this measure is effectively used. The presence of true negative but not incorrectly stated negative values makes it suitable for balanced dataset analysis. The level of model competence is used to assess the quality of various classification models, particularly in cases where binary classification tasks have an equal class distribution. Accuracy is primarily used as an alternative to computational and clarity methods, which are desirable properties of classically straightforward systems [4]. Its performance is evaluated by considering both true positive and true false negatives, making it a good fit for datasets with balanced characteristics. As with the case for classifying best dolphin categories from its relative classes, the measurement of accuracy rate would determine the extent to which the quality predictions could be done on desired feature traits for marine animals' ecosystems.

## 4.0 RESULTS AND DISCUSSION

The testing is done on the Leave-One-Out (LOO) method [32], [33], where clusters with specific numbers containing a myriad assortment of dimensions are examined. From the collective number of 299 samples, 298 samples have been utilized where 1 of the data points is omitted to be stored as a sample. The classification is done with data points within the range of the remaining 298 samples, where this test set is standardized. The classification algorithm is run using premeditated samples, where the result is examined according to the output relevancy with the expert's prediction. The final output will portray the identified percentage of the correct classification. The limitations of the selected dataset size is in its relatively smaller sample size since the number of populated dolphin specimen are fairly distinguishable where the purpose of probability classification is the identification of the better swarm of bottlenose dolphin that would indicate the health level and population compatibility, thus indirectly assists in aiding and establishing a better evaluation metric on their environmental habitat preference. The predictive capabilities on selected dolphin locomotive traits is also more catered towards marine mammals such as dolphin categories themselves rather than generalized approach towards approximating the best specimen with other animal groups within similar ecosystems. The probability classification supplemented with supervised learning as implicated within the study conjecture also tends towards marine animal habits that persist of breathing and swimming patterns among their designated locations, thus will produce computational complexity that varies according to the dataset density and subset dimensionality.

### 4.1 Effects of Dimensionality Clusters on Probability Classification

The general rule of thumb for dimensionality complexity in affecting the performance of feature prediction is that the selection of optimal $k$ is influenced by the congruence of finite space domain for the problem instance, dictated by the size of $k$ in producing sensitive results about the sparseness of the data point and the vague state of non-gleaned data. Any bloated size of $k$ would also crowd the neighborhood to be too inclusive of outliers relative to the subclass residing within the domain. In terms of evaluating classification performance for probability classification, weighted polling could assist in mitigating the sensitivity of whether to bloat or minimize the size of $k$ when selecting data points within the cluster grouping. Weighted polling approaches have been proven to be less perceptive with the $k$ dimensionality [20], [24]. This comes to terms with the choice of $k$ in the adjacency of the nearest neighbor principle, however, this still recedes in the dependency of classification on the relative $k$ size.

The number of neighbors, $k$ is a crucial factor in the performance of the k-nearest neighbors (KNN) algorithm, which depends on the selection of hyperparameters [3], [4]. The magnitude of a change in the size of $k$ can either cause overfitting or excess fitting, depending on whether it is small or large instance of data points. A simple heuristic suggests that selecting unbalanced number representing $k$ leads to bias and high variance, while selecting larger objects also results in bias but low variance. The preference for an odd $k$ in datasets with imbalances is to avoid ties. Therefore, a sufficiently high KNN model is required to create essentially the same well-balanced and effective KNN model, which is known as delta-bias equilibrium. The best $k$ is often cross-validated [14], depending on the dataset. Weighted distance-based voting can refine predictions and ensure that closer neighbors have a greater impact. The weight function affects how all neighbors contribute to predictions, with uniform weighting treating all neighboring equally and distance–based weight affecting closer neighbours [24]. Bayesian optimization and cross-validation are among the techniques that could be utilized to optimize these hyperparameters. In the study context, an even number of $k$ dimensionality was decided as the refining factor for optimization maximization in relative with the subset dimensionality and the limited sparsity of the datapoint size.

Table 4. Percentages of Tested Probability Classifier Ranges for All Participating Dimensions

| Number of $k$ | KNN | | | Naive Classifier |
|---|---|---|---|---|
| | Normal | Weighted | Weighted-Squared | |
| 1 | 92.31 | 92.31 | 92.31 | 92.31 |
| 2 | 92.31 | 92.31 | 92.31 | 90.97 |

| | | | | |
|---|---|---|---|---|
| 3 | 92.31 | 92.31 | 92.31 | 91.97 |
| 4 | 91.97 | 92.31 | 92.31 | 92.64 |
| 5 | 91.30 | 92.31 | 92.31 | 92.98 |
| 6 | 91.30 | 91.64 | 92.31 | 88.63 |
| 7 | 89.97 | 90.97 | 92.31 | 90.97 |
| 8 | 89.97 | 90.64 | 91.97 | 89.97 |
| 9 | 88.96 | 90.30 | 91.97 | 90.64 |
| 10 | 88.96 | 89.97 | 91.30 | 91.30 |
| **Average** | **90.936** | **91.507** | **92.141** | **91.238** |

Table 4 displays the percentages of tested probabilities across implemented classifier ranges inclusive of all participating dimensions. Within this conjecture, KNN weighted-squared had produced the best result at average of 92.141% whereas the baseline KNN implementation performs the best at average of 90.936%. The trend of accuracy rate across all tested *k* dimensions displayed inconsistent performance with occasional fluctuation and decrement in between all domains.

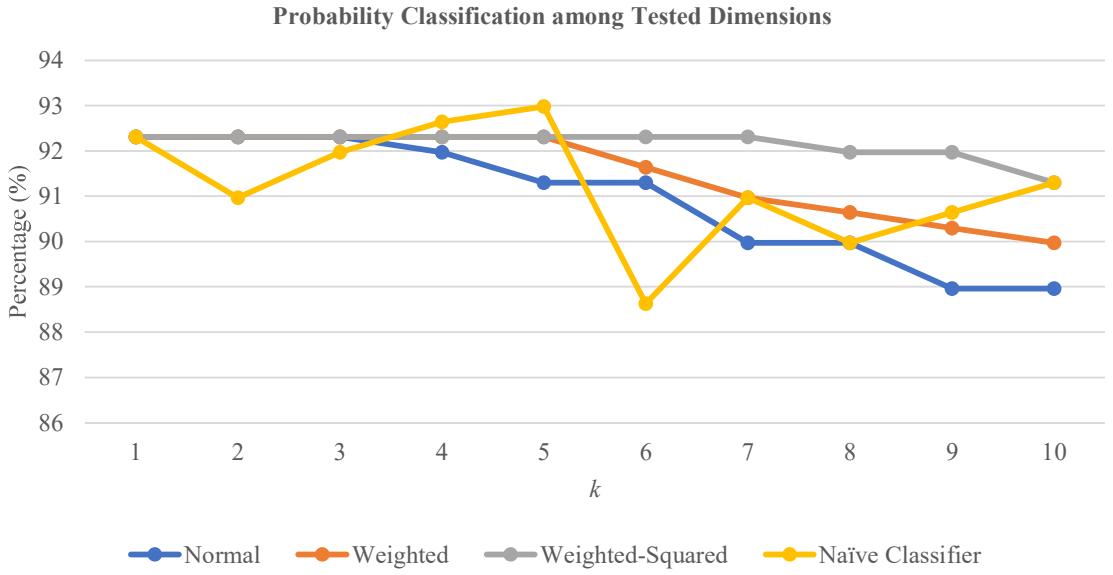


Figure 2. Tested instances for all applied dimensions ($k$ = 1-10) in terms of percentage (%)

Figure 2 shows the results obtained for the probability classification among the tested dimensions. To elaborate, the observation from a comparison of probability classifications across all dimensions within range ($k$ = 0-10) portrays an evenly consistent percentage pattern among the tested instances, albeit the Bayes classifier displays a slightly inconsistent range when tested in between $k = 4$ and $k = 7$. One point that can be denoted from the testing across all dimensions is the weighted-squared polling implementation provides the most consistent gradient with a better percentage across all sectors with each increment of $k$, however, the improvement of accuracy percentage still depends on the relative dimensionality size. This fact is due to the nature of the nearest neighbor assigned with greater values with each $k$ increment thus elevating the accuracy value. All other instances of weighted polling display relative reliability in terms of percentage consistency as there are no wide deviations from the previous values with each distinctive increase or decrease in accuracy percentage.

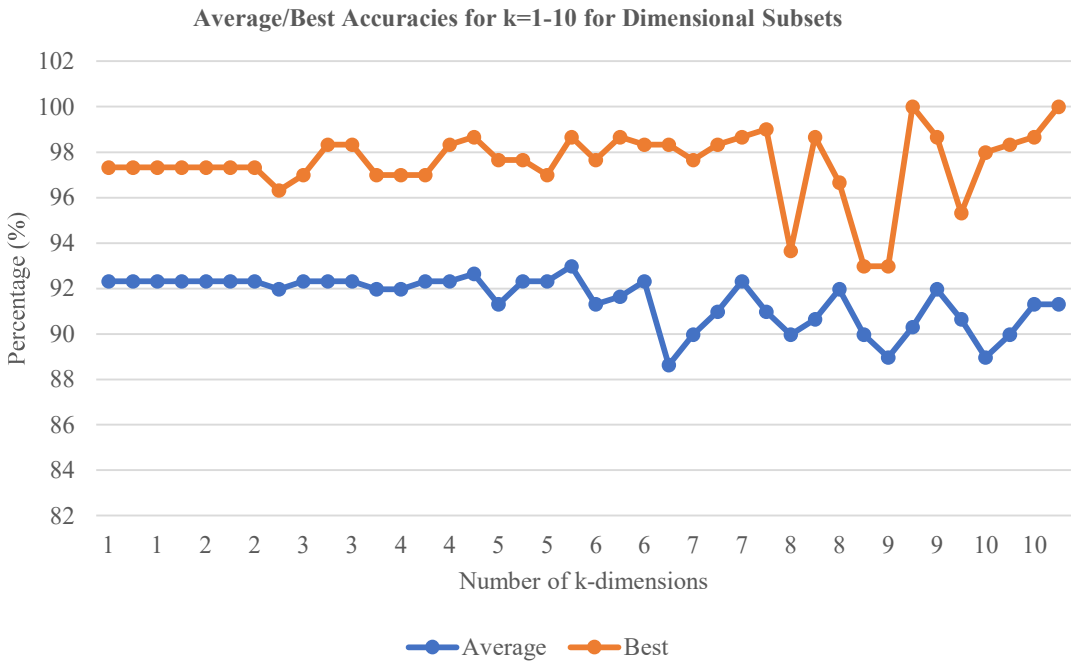


Figure 3. Polling results across all numbers of *k*-dimensions (*k*=1-10) for tested weighted polling methods

As shown in Figure 3, the comparison between the accracy percentages exhibited by both the weighted polling mechanisms demonstrated distinctive performances across the tested *k* dimensions where there are fluctuations and decrement according to the inclusion of subset dimensionality items for all tested k dimensionality domains ranging from as high as 100% to as low as 89%.

### 4.2 Benchmarking

The following tables illustrate the overall prediction performance for the tested data instances for the custom testing dataset derived from the reference bottlenose dolphin dataset.

Table 5. The tested instance and the best dimension assortment to achieve the highest accuracy rate

| Algorithm | Method | Dimension subset used to achieve the best | Average | Best | Class |
|---|---|---|---|---|---|
| KNN | 1 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| | 2 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| Bayes | 3 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| | 4 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| KNN | 1 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| | 2 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| Bayes | 3 | LapNum/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 97.32 | 83H1 |
| | 4 | Seg_Dur_s/Lap_Time_s/Station_Time_s/ CO2_ml_per_kg_min | 91.97 | 96.32 | 83H1 |
| KNN | 1 | LapNum/Lap_Time_s/Seg_Dur_s | 92.31 | 96.99 | 83H1 |
| | 2 | LapNum/Station_Time_s/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 98.33 | 83H1 |
| Bayes | 3 | LapNum/Station_Time_s/Seg_Dur_S | 92.31 | 98.33 | 83H1 |
| | 4 | LapNum/Lap_Time_s/Seg_Dir_s | 91.97 | 96.99 | 83H1 |
| KNN | 1 | LapNum/Lap_Time_s/Seg_Dur_S | 91.97 | 96.99 | 83H1 |
| | 2 | LapNum/Lap_Time_s/Seg_Dur_S | 92.31 | 96.99 | 83H1 |
| Bayes | 3 | LapNum/Station_Time_s/Seg_Dur_s/CO2_ml_per_kg_min | 92.31 | 98.33 | 83H1 |
| | 4 | LapNum/Lap_Time_s | 92.64 | 98.66 | 83H1 |
| KNN | 1 | LapNum/T_Start_s | 91.30 | 97.66 | 83H1 |
| | 2 | LapNum/Station_Time_S | 92.31 | 97.66 | 83H1 |
| | 3 | LapNum/Lap_Time_s/Seg_Dur_s | 92.31 | 96.99 | 83H1 |

| | | | | | |
|---|---|---|---|---|---|
| Bayes | 4 | LapNum/Lap_Time_s | 92.98 | 98.66 | 83H1 |
| | 1 | LapNum/T_Start_S | 91.30 | 97.66 | 83H1 |
| KNN | 2 | LapNum/CO2_ml_per_kg_min | 91.64 | 98.66 | 83H1 |
| | 3 | LapNum/CO2_ml_per_kg_min | 92.31 | 98.33 | 83H1 |
| Bayes | 4 | LapNum//CO2_ml_per_kg_min | 88.63 | 98.33 | 83H1 |
| | 1 | LapNum/T_Start_s/T_End_s | 89.97 | 97.66 | 83H1 |
| KNN | 2 | LapNum/T_Start_s/T_End_s | 90.97 | 98.33 | 83H1 |
| | 3 | LapNum/CO2_ml_per_kg_min | 92.31 | 98.66 | 83H1 |
| Bayes | 4 | LapNum/C02_ml_per_kg_min | 90.97 | 99.00 | 83H1 |
| KNN | 1 | LapNum/T_Start_s /Seg_Dur_s/CO2_ml_per_kg_min | 89.97 | 93.65 | 83H1 |
| | 2 | LapNum/ /T_Start_s | 90.64 | 98.66 | 83H1 |
| | 3 | LapNum/T_Start_s/CO2_ml_per_kg_min | 91.97 | 96.66 | 83H1 |
| Bayes | 4 | LapNum/T_Start_s/Lap_Time_s/T_End_s/Seg_Dur_S/First_Breadth_in_Station_s/CO2_ml_per_kg_min | 89.97 | 92.98 | 83H1 |
| | 1 | LapNum/T_Start_s/Station_Time_s | 88.96 | 92.98 | 83H1 |
| KNN | 2 | LapNum | 90.30 | 100.0 | 83H1 |
| | 3 | LapNum/CO2_ml_per_kg_min | 91.97 | 98.66 | 83H1 |
| Bayes | 4 | LapNum/First_Breadth_in_Station_s/CO2_ml_per_kg_min | 90.64 | 95.32 | 83H1 |
| | 1 | LapNum/T_End_s | 88.96 | 97.99 | 83H1 |
| KNN | 2 | LapNum/Station_Time_s | 89.97 | 98.33 | 83H1 |
| | 3 | LapNum/CO2_ml_per_kg_min | 91.30 | 98.66 | 83H1 |
| Bayes | 4 | LapNum/CO2_min_per_kg_min | 91.30 | 100.0 | 83H1 |

As seen in Table 5, all classes in predicting the best dolphin specimen had annotated 83H1 as the best dolphin class irrespective of weighted polling methods between KNN and Naïve Bayes. In retrospect, KNN algorithm with all its variants produced a marginally better accuracy rate as compared with Naïve Bayes' implementations. Although Naïve Bayes have produced a result of 100% prediction rate in several runs, the entirety of accuracy rate represented via this weighted polling method exhibited a subtle inconsistency across all domains in comparison with the consistency produced via KNN in terms of KNN (normal, weighted, weighted-squared). In Section 3.1.3 it was hypothesized that testing could be done on probability classification to predict the bottlenose dolphin class with above-average performance based on locomotive activities like swimming patterns, metabolism rate during submerging/emerging, and oxygen consumption while submerged. All 255 dimension subsets have been included in the test for each unique *k* value and classification method. One goal is to determine the subset of dimensions that produces the best accuracy in every test as a whole. The results of the test demonstrate that it is crucial to include multiple subsets of dimensions for each classification method to increase the accuracy rate. Table 5 lists the main dimensional factor needed to achieve the best accuracy rate, including variables like *LapNum*, *Seg_Dur_s*, and *CO2_ml_per_kg_min* which are crucial for enhancing accuracy. Each probability assessment has predicted that dolphin subgroup 83H1 outperforms all other subgroups consistently.

Table 6. the frequency of dimensional variables on producing the better accuracy rate among all the tested dimensional subset

| Dimension variable | Frequency for each respective best dimension |
|---|---|
| *LapNum* | 39 |
| *T_Start_s* | 9 |
| *T_End_s* | 4 |
| *Seg_Dur_s* | 17 |
| *First_Breath_In_Station_s* | 0 |
| *Lap_Time_s* | 9 |
| *Station_Time_s* | 7 |
| *CO2_ml_per_kg_min* | 20 |

Table 6 displays the dimensional variable that was included in the referenced dataset and added alongside it. The individual variables for each variable were already listed in Table 2, but these variables provided valuable insights into dolphin group breathing and swimming metabolism. Furthermore, the results of the study revealed several observations on the impact of subset dimensionality on accuracy rate for all dolphin classes implicated in the experimentation. *LapNum* shows the greatest impact in all dimensional subsets as inclusion for improving accuracy rates, while *First_Breadth_in_Station_s* shows the most negative impact among the dimensional arrays tested. An average accuracy of 96.9% was attained using only a few of the most important dimensions from the classified data points list. These findings have demonstrated several important points to note involving the measurement and inclusion of crucial subset dimensionality elements in order to influence the outcome of the prediction result. Firstly, any attempts to approximate and conject the prime dolphin class as best performing candidate need to include the number of laps (*LapNum*), the segments of swimming traversal (*Seg_Dur_s*), and concentration of exhaled/inhaled air (carbon dioxide/oxygen) for a

better representation of the specific dolphin class that fulfill the aspiration criterion (*CO2_ml_per_kg_min*). The exclusion of any of these subset variables would drastically decline the probability classification augmenting relevant feature selections. Furthermore, the inclusion of several dimension variable such as initial breathing when departing from stationary state (*First_Breath_In_Station_s*), the end point of each swimming segment (*T_End_s*), and the period in which the dolphin is stationary at the current starting point (*Station_Time_s*) does not play a significant role in boosting the efficiency of the polling result where in some instances their omittance displays a relevance in further elevating the accuracy rate. As with the case of the better counterpart of weighted polling measures in properly approximate the prime dolphin class, the Bayes classifier was the only one with inconsistent accuracy distribution as the subset dimensions increased, while other weighted polling methods imposed within the KNN algorithm demonstrated a consistent result across all dimensions. This outcome showed that additional dimensional subsets were added to improve the overall accuracy rate of the entire domain spectrum by compensating for inaccuracies in another related cluster. The inclusion of several subset dimensionality variable would produce desirable predictive rate in favour of the niche groups and their respective traits.

## 5.0 CONCLUSIONS

This study evaluates the effectiveness of weighted probing for probabilistic classification and compares its performance with traditional methods in similar classification situations. The study focused on testing probabilistic classification using a locomotive dataset of banded dolphins. The evaluation is performed by measuring the accuracy of predicting the exact dolphin classification within a group of banded dolphins. To see how the number of dimensions k affects the ranking values and to obtain more reliable results, all variants of the classifier model were tested over 10 clusters depending on the number of dimensions k and a summary of the final results was obtained. The study showed that the use of weighted probing is a more reliable option compared to traditional classification techniques. Weighted techniques give more weight to near neighbors, improving prediction accuracy and overall success. One potential area of future research is to use classifier strategies to identify features associated with each category that meet specific requirements while addressing limitations or potential low-dimensional integration, in particular helping to demonstrate probabilistic classification of targeted features better than the average. To address the importance of a subset of dimensional sectors that are considered feasible to increase accuracy rates. Future research could also investigate how often each classification is performed consecutively to assess retention rates. This would help to identify the adjustments needed to improve the accuracy of classifications that are predicted correctly less frequently than other classifications.

## ACKNOWLEDGEMENTS

The author expresses gratitude to Universiti Malaysia Pahang and Universiti Pendidikan Sultan Idris for their support of this research and publication. Additionally, the author values the reviewers for their helpful comments on the article. This research did not receive any financial support from funding organizations in the public, private, or nonprofit sectors.

## AUTHORS CONTRIBUTION

Farid Morsidi (Conceptualisation; Methodology; Validation; Formal analysis; Data curation; Formal analysis; Investigation; Resources; Software; Visualisation; Writing - original draft; Writing - review & editing)

## CONFLICT OF INTEREST

The author states that there is no conflict of interest regarding the publication of the materials in this paper.

## REFERENCES

[1] N. Suryani and A. S. Fitrani, "Prediction Of Election Participant With Malang City Demographic Data Using The K-Nn Algorithm," *J. Mantik*, vol. 6, no. 36, pp. 2369–2376, 2022, [Online]. Available: https://ejournal.iocscience.org/index.php/mantik/article/view/2802%0Ahttps://ejournal.iocscience.org/index.php/mantik/article/download/2802/2223

[2] K. Kaharuddin and E. W. Sholeha, "Classification of Fish Species with Image Data Using K-Nearest Neighbor," *Int. J. Comput. Inf. Syst.*, vol. 2, no. 2, pp. 54–58, 2021, doi: 10.29040/ijcis.v2i2.33.

[3] R. Taylor, "Machine Learning Techniques for Fish Breeding Decision Making," *2023 Wellingt. Fac. Eng. Symp.*, pp. 1–12, 2023, [Online]. Available: https://ojs.victoria.ac.nz/wfes/article/view/8422

[4] S. Winiarti, F. I. Indikawati, A. Oktaviana, and H. Yuliansyah, "Consumable Fish Classification Using k-Nearest Neighbor," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 821, no. 1, 2020, doi: 10.1088/1757-899X/821/1/012039.

[5] I. F. Dehkordi, K. Manochehri, and V. Aghazarian, "Internet of Things ( IoT ) Intrusion Detection by Machine Learning ( ML ): A Review," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 12, no. 1, pp. 13–38, 2023, doi: dx.doi.org/10.17576/apjitm-2023-1201-02.

[6] J. K. C. Revanna and N. Y. B. Al-Nakash, "Metaheuristic link prediction (MLP) using AI based ACO-GA optimization model for solving vehicle routing problem," *Int. J. Inf. Technol.*, vol. 15, no. 7, pp. 3425–3439, 2023, doi: 10.1007/s41870-023-01378-5.

[7] M. K. Alsmadi and I. Almarashdeh, "A survey on fish classification techniques," *J. King Saud Univ. - Comput.*

*Inf. Sci.*, vol. 34, no. 5, pp. 1625–1638, 2022, doi: 10.1016/j.jksuci.2020.07.005.
[8] R. Rosly, M. Man, A. Ngah, and N. S. A. Manan, "Multi-classifier models to improve the accuracy of fish landing application," *Int. J. Adv. Technol. Eng. Explor.*, vol. 11, no. 111, pp. 145–159, 2024, doi: 10.19101/IJATEE.2023.10102060.
[9] S. S. Latt and M. Myint, "Classification of Fish based on Naive Bayesian Classifier," in *LISS 2013*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 585–589. doi: 10.1007/978-3-642-40660-7_86.
[10] A. Fix and V. Efix, "Comparison between the KNN, W-KNN, Wc-KNN and Wk-KNN models on a CDC heart disease dataset," no. April, 2024, doi: 10.21203/rs.3.rs-4505140/v1.
[11] M. H. Alobaidi, M. A. Meguid, and T. Zayed, "Semi-supervised learning framework for oil and gas pipeline failure detection," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-16830-y.
[12] S. Sulaiman, R. A. Wahid, and F. Morsidi, "Feature extraction using regular expression in detecting proper noun for Malay news articles based on KNN algorithm," *J. Fundam. Appl. Sci.*, vol. 9, no. 5S, p. 210, Jan. 2018, doi: 10.4314/jfas.v9i5s.16.
[13] R. R. Rani and D. Ramyachitra, "Krill Herd Optimization algorithm for cancer feature selection and random forest technique for classification," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2018, pp. 109–113. doi: 10.1109/ICSESS.2017.8342875.
[14] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazanci, and O. Oral, "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture," *J. King Saud Univ. - Sci.*, vol. 35, no. 6, 2023, doi: 10.1016/j.jksus.2023.102754.
[15] A. S. Allen *et al.*, "Dynamic body acceleration as a proxy to predict the cost of locomotion in bottlenose dolphins," *J. Exp. Biol.*, vol. 225, no. 4, pp. 1–13, 2022, doi: 10.1242/jeb.243121.
[16] A. Saleh, M. Sheaves, and M. Rahimi Azghadi, "Computer vision and deep learning for fish classification in underwater habitats: A survey," *Fish Fish.*, vol. 23, no. 4, pp. 977–999, 2022, doi: 10.1111/faf.12666.
[17] Y. Baidai, L. Dagorn, M. J. Amande, D. Gaertner, and M. Capello, "Machine learning for characterizing tropical tuna aggregations under Drifting Fish Aggregating Devices (DFADs) from commercial echosounder buoys data," *Fish. Res.*, vol. 229, no. September, 2020, doi: 10.1016/j.fishres.2020.105613.
[18] S. M. M. Islam, S. I. Bani, and R. Ghosh, "Content-based Fish Classification Using Combination of Machine Learning Methods," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 1, pp. 62–68, 2021, doi: 10.5815/ijitcs.2021.01.05.
[19] Y. Gültepe, "Analysis of Alburnus tarichi population by machine learning classification methods for sustainable fisheries," *SLAS Technol.*, vol. 27, no. 4, pp. 261–266, 2022, doi: 10.1016/j.slast.2022.03.005.
[20] J. Gou, T. Xiong, and Y. Kuang, "A novel weighted voting for K-nearest neighbor rule," *J. Comput.*, vol. 6, no. 5, pp. 833–840, 2011, doi: 10.4304/jcp.6.5.833-840.
[21] S. Saputra, A. Yudhana, and R. Umar, "Implementation of Naïve Bayes for Fish Freshness Identification Based on Image Processing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3, pp. 412–420, 2022, doi: 10.29207/resti.v6i3.4062.
[22] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.
[23] I. Fondón *et al.*, "Automatic classification of tissue malignancy for breast carcinoma diagnosis," *Comput. Biol. Med.*, vol. 96, no. December 2017, pp. 41–51, 2018, doi: 10.1016/j.compbiomed.2018.03.003.
[24] M. Açıkkar and S. Tokgöz, "An improved KNN classifier based on a novel weighted voting function and adaptive k-value selection," *Neural Comput. Appl.*, vol. 36, no. 8, pp. 4027–4045, 2024, doi: 10.1007/s00521-023-09272-8.
[25] W. L. Buntine, "A Theory Of Learning Classification Rules," 1992. doi: 10.1.1.49.5614.
[26] K. Demir and O. Yaman, "A HOG Feature Extractor and KNN-Based Method for Underwater Image Classification," *Firat Univ. J. Exp. Comput. Eng.*, vol. 3, no. 1, pp. 1–10, 2024, doi: 10.62520/fujece.1443818.
[27] Tavish Srivastava, "Guide to K-Nearest Neighbors Algorithm in Machine Learning," *Https://Www.Analyticsvidhya.Com/Blog/2018/03/Introduction-K-Neighbours-Algorithm-Clustering/*, pp. 1–20, 2024.
[28] D. S. Pedroche, D. Amigo, J. García, and J. M. Molina, "Architecture for trajectory-based fishing ship classification with AIS data," *Sensors (Switzerland)*, vol. 20, no. 13, pp. 1–21, 2020, doi: 10.3390/s20133782.
[29] F. Idachaba and O. Tomomewo, "Surface pipeline leak detection using realtime sensor data analysis," *J. Pipeline Sci. Eng.*, vol. 3, no. 2, p. 100108, 2023, doi: 10.1016/j.jpse.2022.100108.
[30] T. De Kerf, J. Gladines, S. Sels, and S. Vanlanduit, "Oil spill detection using machine learning and infrared images," *Remote Sens.*, vol. 12, no. 24, pp. 1–13, 2020, doi: 10.3390/rs12244090.
[31] I. Md Jelas, M. A. Zulkifley, M. Abdullah, and M. Spraggon, "Deforestation detection using deep learning-based semantic segmentation techniques: a systematic review," *Front. For. Glob. Chang.*, vol. 7, no. February, 2024, doi: 10.3389/ffgc.2024.1300060.
[32] Y. Xu, W. Yang, and J. Wang, "Air quality early-warning system for cities in China," *Atmos. Environ.*, vol. 148, pp. 239–257, 2017, doi: 10.1016/j.atmosenv.2016.10.046.
[33] D. R. Tobergte and S. Curtis, "The 54th Annual Meeting of the Association for Computational Linguistics," in *Climate Change 2013 – The Physical Science Basis*, vol. 53, no. 9, Cambridge University Press, 2016, pp. 1–30. doi: 10.18653/v1/P16-2