Analysing Unlabeled Data with Randomness and Noise: The Case of Fishery Catch Reports

Aida Ashrafi[1] and Bjørnar Tessem[1] and Katja Enberg[2]

***Abstract*— Detecting violations within fishing activity reports is crucial for ensuring the sustainable utilization of fish resources, and employing machine learning methods holds promise for uncovering hidden patterns within this complex dataset. Given that these violations are infrequent occurrences, as fishermen generally adhere to regulations, identifying them becomes akin to an anomaly outlier detection task. Since labeled data distinguishing between normal and anomalous instances is not available for catch reports from Norwegian waters, we have opted for more conventional approaches, such as clustering methods, to identify potential clusters and outliers. Moreover, the catch reports inherently exhibit randomness and noise due to environmental factors and potential errors made by fishermen during report registration which complicates the processes of scaling, clustering, and anomaly detection. Through experimentation with various scaling and clustering techniques, we have observed that many of these methods tend to group the data based on the species caught, exhibiting a high level of agreement in cluster formation, indicating the stability of the clusters. Anomaly detection methods, however, yield varying potential outliers as it is a more challenging task.**

I. INTRODUCTION

Leveraging machine learning and data science for the United Nations' Sustainable Development Goals (SDGs) offers a promising contribution towards their effective implementation. Among the SDGs, SDG14 highlights the importance of life below water and the imperative to enhance sustainability within the fisheries industry [1]. An essential aspect of achieving this goal involves combating Illegal, Unreported, and Unregulated (IUU) fishing [2], for which AI-driven monitoring systems offer significant utility. Malde et al. [1] and Handegard et al. [2] underscore the significance of employing machine learning techniques in marine science and promoting sustainable fisheries practices. Initially, scientists utilized traditional machine learning models ([3]), but have since transitioned to employing deep learning models ([4], [5], and [6]), for tasks such as fishing activity detection and preventing overfishing.

Our research focuses on harnessing AI to analyze fishing catch reports from Norwegian waters, aiming to support regulatory authorities - in this case, the Norwegian Directorate of Fisheries (NDF) [3]—in gaining comprehensive insights into fishing activities over time. We aim to find any hidden patterns in the required catch reports by fishermen, a huge amount of data over the last decades. These data are, however, not annotated with kind of labels one normally expects to have for machine learning. Hence, unsupervised approaches for analysis is necessary to get insights into the data.

A. Problem Relevance

While the majority of vessels adhere to regulations most of the time, occasional violations occur. These deviations from the norm, being rare events that deviate from expected patterns, may be classified as anomalies or outliers [7].

A prevalent method for unsupervised anomaly detection involves utilizing an autoencoder to reconstruct the training data, which exclusively comprises normal data. A threshold for reconstruction error is established using this training data. During the testing phase, both normal and anomalous data can be employed, and anomalies are identified as those with errors significantly deviating from the threshold [8].

Monitoring fishing vessels comprehensively, especially while they are at sea, presents a daunting challenge. Although numerous catch reports are available, they consist of raw data provided by fishermen and lack the annotations indicating which are in some sense irregular or normal behavior. Consulting experts for such annotations is impractical due to both the efforts needed and the dynamic nature of fishing regulations across different regions and times.

The dataset exhibits features with a variety of distribution shapes, which needs to be handled according to recommended practice. However, if we look at the effect of fishing, namely the catch features (species and amount), they also exhibit extra high degrees of randomness, making the use of many unsupervised machine learning techniques challenging.

Our research still endeavors to uncover hidden patterns within this complex dataset using machine learning models, aiming to provide insights into fishing activities and facilitate anomaly detection.

B. The Contribution

The catch reports are tabular data with both categorical and numerical features including gear type, start and stop position (latitude and longitude) of the fishing interval, duration of the fishing interval, time of the catch activity, length of the vessel, ID of the vessel (called callsign), round weight, and species.

The objective is to analyze the dataset, identifying patterns and potential anomalies, which may include erroneous or suspicious reports. To our knowledge, this work marks a

[1]Dept. of Information Science and Media Studies, University of Bergen, Norway

[2]Dept. of Biological Sciences, University of Bergen, Norway

[1]https://www.un.org/sustainabledevelopment/oceans/

[2]https://www.fao.org/sustainable-development-goals-data-portal/data/indicators/1461-illegal-unreported-unregulated-fishing/

[3]https://www.fiskeridir.no

pioneering application of machine learning models to analyse fishery activity data in terms of deviating reports.

At a general level the research task is to enable analysis of data that exhibit some well-known problematic features, like randomness, sloppy incorrect reporting, missing values, and intended incorrect reporting. These issues still needs to be overcome to be able to support the main purpose, i.e., the application of data to support resource management.

Given the complex regulatory landscape established by the NDF, detecting irregularities within the data poses a significant challenge. Identifying deviations from legitimate fishing activities is not straightforward.

Traditional machine learning approaches, such as clustering techniques, have been used to address such issues. Distinct clusters represent groups of data points sharing similar patterns, while data points located far from any cluster may be regarded as anomalies [9].

Additionally, we employ different dimensionality reduction techniques to facilitate the visualization of the data in two dimensions, enhancing our ability to discern normal behaviour patterns and anomalies effectively.

We have started out by focusing on bottom trawlers; nevertheless, the methodology employed should hold relevance for other geographic regions and various types of fisheries.

The next section delves into the problem's background and the related work on the selected methodologies. In Section III, we provide an overview of the original dataset, detail the pre-processing steps undertaken, and elucidate the final dataset selection process. Moving forward, we illustrate the data visualization and outcomes derived from the clustering methods, along with identifying potential anomalies using various techniques in Section IV. Section V concludes with discussions and summarizing key findings.

II. BACKGROUND AND RELATED WORK

To comprehend and analyze this intricate dataset, we adhered to the following steps, which are common in machine learning and data science practices.

A. Dimensionality Reduction

Processing high-dimensional data, which often comprises numerous features, demands significant time, computational resources, and storage space. Dimensionality reduction techniques aim to alleviate these challenges by eliminating redundant information while preserving essential data with minimal loss, thus providing a more efficient low-dimensional representation. Additionally, dimensionality reduction facilitates data visualization, which is crucial for gaining insights into complex datasets. Dimensionality reduction can be achieved through either feature selection or feature extraction. Feature selection algorithms preserve the original features, whereas feature extraction algorithms transform the data into a new feature space.

One of the most widely used linear dimensionality reduction methods is Principal Component Analysis (PCA), which seeks orthogonal directions that explain the maximum variance in the data. Alternatively, autoencoders offer a non-linear approach to dimensionality reduction. An autoencoder is a neural network architecture designed to compress input data into its essential features through an encoder and then reconstruct the original input from this compressed representation efficiently though a decoder [10].

We employ both PCA and autoencoders to gain a better understanding of the data through 2D visualization and utilize the resulting 2D representations for clustering and detecting potential anomalies.

B. Clustering

Clustering methods have been in existence for approximately more than five decades. According to Saxena et al. [11] clustering characterized as unsupervised learning, where the labels for objects are not available. This makes the task more difficult compare to the supervised approach where the labels have the role of clues. Clustering entails grouping objects based on inherent similarities among them. The objects inside a cluster are more similar to each other than to the objects belonging to other clusters. Numerous clustering algorithms including hierarchical and partitional have been crafted over time to cater to specific domains, despite the absence of a universally acknowledged definition for a cluster. Partitional clustering techniques are also divided into distance-based, model-based and density-based methods.

In hierarchical clustering techniques, clusters are created through an iterative top-down or bottom-up approach. There are two main forms of hierarchical methods: agglomerative and divisive hierarchical clustering. Agglomerative clustering adopts a bottom-up strategy, initially forming clusters from individual objects and progressively merging these atomic clusters into larger ones until either all objects belong to a single cluster or certain termination criteria are met. Conversely, divisive hierarchical clustering employs a top-down approach, starting with a single cluster containing all objects and iteratively splitting it into smaller clusters until each object forms its own cluster or specific termination conditions are fulfilled.

In partitional clustering, unlike hierarchical clustering, data points are allocated into K clusters without any hierarchical arrangement by optimizing a certain criterion function. The Euclidean distance is the most frequently employed criterion, which determines the minimum distance between data points and existed clusters, thereby assigning the data point to a cluster [11].

Agglomerative clustering with single linkage can be a suitable method for our task because it uses minimum distance and the clusters merged in later stages may reveal potential anomalies, which aligns with our objective. However, we also tried some of the partitional methods such as K-means [12] as a distance-based method, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13] as a density-based method, and Self-organizing map (SOM) [14] as a model-based method since our data is complex and there is not a single ideal method for it.

C. Anomaly Detection

Anomaly detection pertains to identifying patterns within data that deviate from expected behavior. These deviant data points are termed anomalies or outliers. A direct approach to anomaly detection involves defining a region that encapsulates normal behavior and flagging any observation outside of this region as an anomaly. However, implementing this straightforward approach in real-world scenarios presents numerous challenges, including a scarcity of labeled data for training. To effectively learn the patterns within normal data, it's imperative to have annotations that help distinguish normal data from anomalies [7].

Given the absence of available labels for our problem, employing the anomaly detection methods described in [8] is not viable. Nonetheless, we have opted to examine the results generated by clustering methods as an alternative approach to uncover potential anomalies.

III. DATASET AND PRE-PROCESSING

The dataset utilized in our study is known as DCA, or daily catch reports, which is published by NDF and is publicly available [4]. This dataset encompasses the fishing activities of various fishing vessels in Norwegian waters. Given the variability in regulations and environmental conditions from year to year, we selected 2018 as a representative sample. The two-dimensional visualization of the data obtained through Principal Component Analysis (PCA) for both 2018 and 2019 is presented in Figures 1 and 2 . We observe that the 2019 version has a similar overall pattern to the 2018 visualisation, but they are presenting somewhat skewed distributions (relative to each other) along their respective principal components.

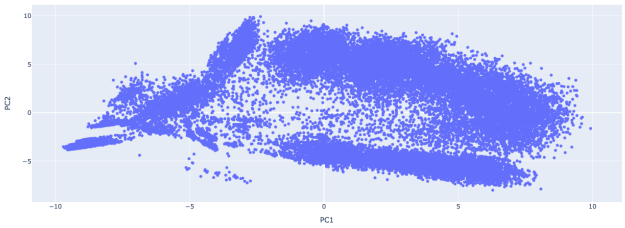

Fig. 1. Two-dimensional representation of data, showing the result of using PCA on DCA data 2018. The logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features.

This dataset comprises numerous features, including the start and stop positions (latitude and longitude) of each catch interval, the time and duration of each catch, the type of gear used for the catch, the species caught, the main species (wherein each catch consists of different species and the one with the highest weight is considered the main one), the length of the vessel, and the vessel's ID (callsign).

After applying pre-processing steps, the initial dataset of 120,000 datapoints representing partial reports for bottom trawlers was refined to approximately 35,000 reports. This reduction was achieved by selectively considering reports featuring species with over 2,000 occurrences and total round weight exceeding 100,000 kg. This focused approach aims to analyze common high catches, facilitating the identification of prominent patterns within the data. Additionally, we consolidated different species (each from a partial report) within each catch into a single row, enhancing the dataset's coherence and simplicity. Consequently, the dataset now exclusively comprises numerical features for streamlined analysis.

[4]A part of the electronic reporting by NDF: https://www.fiskeridir.no/Tall-og-analyse/AApne-data/elektronisk-rapportering-ers

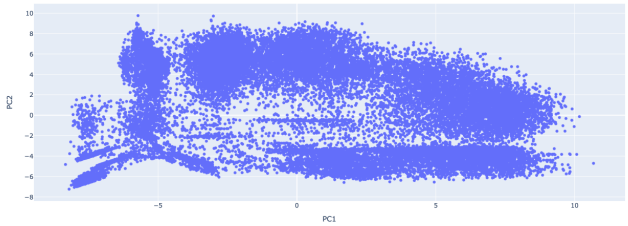

Fig. 2. Two-dimensional representation of data, showing the result of using PCA on DCA data 2019. The logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features.

A. Randomness and Distribution of Catch Data

The data, particularly the 'round weight' feature, encompasses a level of inherent randomness. One should, however, expect that modern industrial fisheries would enable us to get reasonable predictions of catches from data like location, gear and vessel size. The catch quantity is contingent upon environmental conditions, the presence of various species in a specific area on a given date, the ability of the fishermen, and even irregular registration of data contributing to the stochastic nature of this variable. All the data distribution deviates significantly from a normal distribution, posing challenges in identifying the optimal scaling method. Further, the unpredictable nature of the data introduces complexity to the task of discerning patterns within them.

To better understand the randomness in the 'round weight', we have developed a supervised model that predicts total catches for a bottom trawler data set, but slightly reduced in terms of data points and features. The regression value is the log with base 10 (log10) of total catch. A Xgboost (eXtreme Gradient Boosting) model was able to predict the log10 of the total catch with a coefficient of determination (R2) of 0.70 (5-fold cross-validation), meaning that 30% of the variation in the log10 catch could not be explained by the model. This indicates a fairly good model, and an analysis of the residuals or prediction errors showed that they had a mean of 0.0 and a standard deviation of 0.22.

When we look into the real values computed from the exponential of log10 values, we get results which are less convincing on behalf of the predictability of the catches. The errors in catch prediction ranged from 72,265 kilos too low to 17,868 kilos too high. The skewed interval indicates that the model is not able to predict the really big catches,

which are those with high economic value, but also with high environmental impact.

A histogram showing the distribution of the real sizes of catches compared to the predicted sizes is shown in Figure 3. The long tail effect is visible. The presence of numerous extremely small and large catches will undoubtedly pose challenges for anomaly detection and clustering within the dataset. In addition, there is the problem of which species will be caught in by-catches and the amount. These features may in themselves be even more challenging sources of randomness.

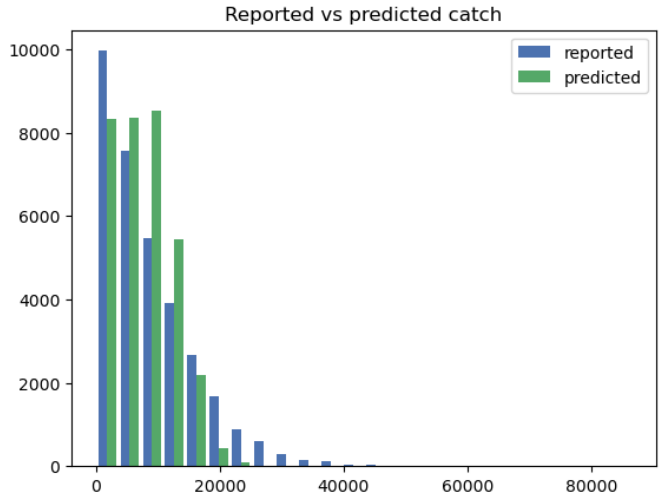

Fig. 3. Comparison of real catches and catches predicted by an xgboost model

B. Scaling The Data

The subsequent stage involves scaling the data, which is essential as we utilize PCA and autoencoder for dimensionality reduction and visualization purposes. We experimented with various methods, and three distinct approaches are outlined here for their potential insights into the data. The first approach entails applying the Standard Scaler to all features. This will transform our dataset such that each feature will have a mean of 0 and a standard deviation of 1. In the second approach, we altered the scaling method solely for the 'round weight' feature, while retaining the previous scaling for the remaining features. Given the considerable skewness in the distribution of 'round weight,' we opted to employ the logarithm function to scale its values. In the third method, we initially take the log10 of the 'round weight' and subsequently scale all features to fall within the range of -1 and 1. The distinctions among these methodologies become apparent in the visualizations presented in the following section.

IV. DATA VISUALIZATION AND ANALYSIS

A. Dimensionality Reduction and Two-dimensional Visualization of The Data

As discussed in the previous section, the choice of scaling method for round weight impacts the distribution of the data observed in two-dimensional visualization. Initially, we explore the application of PCA with all scaled versions of the dataset. Subsequently, we'll transition to using auto-encoder as the dimensionality reduction tool. By employing PCA, we aim to capture the underlying structure of the data and visualize it in a lower-dimensional space. Next, we'll explore the use of auto-encoder, which can potentially reveal additional insights into the data by reconstructing it from a compressed representation.

Additionally t-SNE [15] is employed for visualization purposes. However, as we did not achieve a clearer visualization compared to PCA, we report the results using PCA.

We ended up using Relational Autoencoder (RAE) [16] when utilizing logarithm of round weight, since it shows better performance. This is done utilizing the vanilla version of an autoencoder, where we scale all the features using standard scaling. The architecture of both is the same and quite simple, both the encoder and decoder part have a dense layer with 10 neurons as the only hidden layer. The input dimension is 22, while the latent dimension is 2. RAE captures both the relationships between input features and the relationships between individual data points which can help to improve the reconstruction task [16].

Visualizations of data using PCA are shown in Figures 1, 4, and 5 and the ones with autoencoder are depicted in Figures 6 and 7. In these visualizations we can see differences in the resulting distributions as a consequence of various scaling and dimensionality reduction methods. However, across all visualizations, discernible patterns, clusters, and anomalies are apparent. In the subsequent section, we will delve into these topics comprehensively.

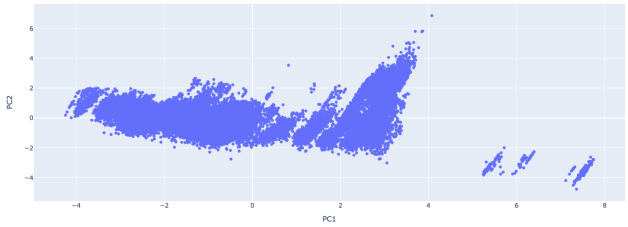

Fig. 4. Two-dimensional representation of data, showing the result of using PCA on DCA data 2018. Standard Scaler is used to scale all the features.

B. Clustering Results and Possible Anomalies

Regulatory conditions can vary greatly from one day to another, further complicating the identification of normal and anomalous instances. Even domain experts may not possess all the requisite details, exacerbating the difficulty of distinguishing between regular and exceptional occurrences. Given the absence of prior annotations, we have opted to employ more conventional machine learning approaches, such as clustering, to mitigate reliance on normal data during training. Our aim is to cluster the data and classify data points that are distant from any clusters as potential anomalies. This strategy allows us to approach anomaly detection in a

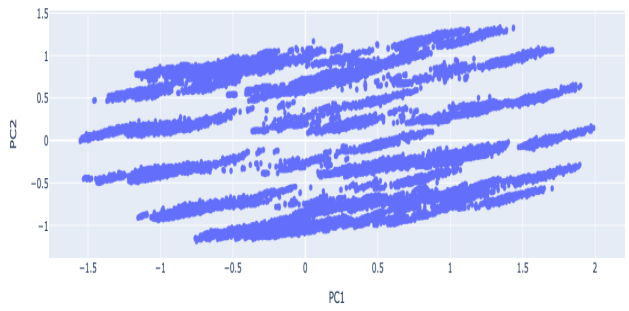

Fig. 5. Two-dimensional representation of data, showing the result of using PCA on DCA data 2018. First we take the logarithm of round weight then scale all features so that they are placed inside the range -1 and 1.

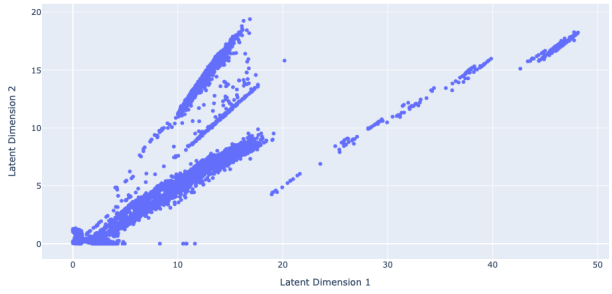

Fig. 6. Two-dimensional representation of data, the result of using RAE on DCA data 2018. The logarithm function is used to scale round weight and Standard Scaler for the rest of the features.

manner less dependent on pre-existing norms. Despite the persisting challenge posed by data randomness discussed in previous section, clustering methods are able to identify certain underlying patterns within the dataset.

We experimented with two scenarios for all clustering methods: firstly, utilizing all 22 features, and secondly, employing a 2D representation. We then examine the resulting clusters to determine which scenario produces more reasonable results. The scenario that yields more reasonable clusters is considered to have better performance.

In line with the details outlined in Section II-B, we em-

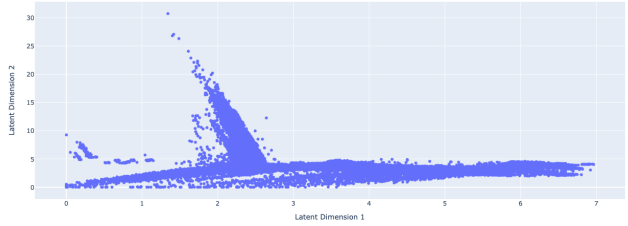

Fig. 7. Two-dimensional representation of data, the result of using autoencoder on DCA data 2018. Standard Scaler is used to scale all the features.

TABLE I

CLUSTERS WITH ONE DATAPOINT FROM FIGURE 8, SOME OF THE POTENTIAL OUTLIERS OBTAINED AFTER USING AGGLOMERATIVE CLUSTERING ON 2D DATA FROM PCA. THE THIRD ROW TO THE SIXTEENTH ROW INCLUDE THE ROUND WEIGHT OF COMMON SPECIES DURING THAT CATCH IN KG. THE LAST FOUR ROWS ARE THE START AND STOP POSITION OF THE CATCH INTERVAL.

	cluster 10	cluster 11	cluster 15	cluster 19
vessel length	33.95	29.92	33.95	19.75
month	4	5	**11**	7
duration	425.0	**283.0**	117.0	360.0
Cod	**3.0**	6.0	84.0	**5.0**
Saithe	**10.0**	0.0	4826.0	0.0
Haddock	0.0	0.0	32.0	0.0
Rosefish	0.0	0.0	0.0	0.0
Caridean shrimp	0.0	0.0	0.0	0.0
Ling	**30.0**	15.0	0.0	0.0
Beaked redfish	0.0	0.0	0.0	0.0
Greenland halibut	0.0	0.0	0.0	0.0
Spotted wolffish	0.0	0.0	0.0	0.0
Hake	0.0	4.0	0.0	0.0
Atlantic wolffish	0.0	0.0	0.0	0.0
Angler	0.0	30.0	0.0	0.0
Halibut	0.0	0.0	0.0	0.0
Pollack	0.0	0.0	0.0	0.0
start latitude	65.7	64.258	68.907	71.175
start longitude	9.433	8.723	13.508	28.434
stop latitude	65.683	64.371	68.824	71.149
stop langitude	65.683	9.139	13.275	28.646

ployed agglomerative clustering to simultaneously identify clusters and potential outliers within the dataset. Following parameter adjustments, we generated Figure 8, the clusters are achieved using the 2 principal components and depicted using 2D visualization in Figure 5. There are 8 main clusters which are grouped mainly based on the combination of species present in the catch, the rest seems to be deviations from the main ones. For example it is evident that cluster 7 is a notably small cluster, appearing to diverge from cluster 1. Upon closer examination of the features, cluster 7 comprises five data points, with one species shared with cluster 1. However, the vessel size and the duration of catch within this cluster is considerably smaller compared to those within cluster 1, despite capturing the same species.

Some of these potential outliers exhibit deviations from the nearest cluster in terms of the catch amount, either being excessively small or large, and sometimes they encompass different combinations of species. Interestingly, cluster 7 and 19 belong to the same vessel. Hence, it's apparent that certain vessels have experienced more deviations compared to others. In the case of cluster 19, there is only one data point with very small catch of only one species. Furthermore, clusters 14 and 15 are associated with the same vessel, yet the combination of species differs slightly, despite being caught in the same area.

The total count of data points distant from larger clusters but associated with very small clusters is 30. Table I displays the features of some of these data points, with the features contributing to the deviation highlighted in bold. All these data points are candidates for being classified as anomalies.

Furthermore, according to [9], in order to detect outliers using hierarchical clustering, we can generate a dendrogram of the clustering method applied to the data. This visualiza-

tion allows us to identify clusters that are distinctively distant from others. Data points belonging to such clusters can then be considered potential outliers. The dendrogram is depicted in Figure 9.

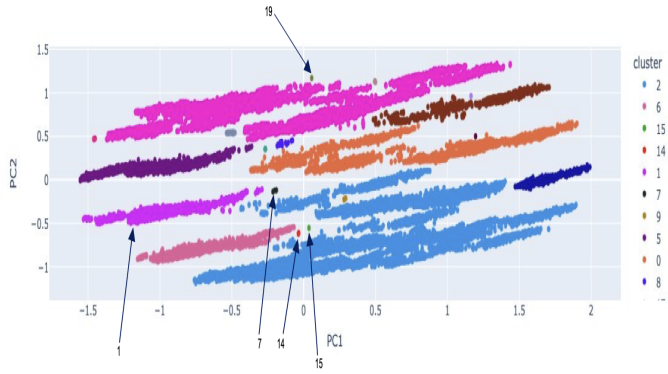

Fig. 8. Clusters and potential outliers using agglomerative clustering on two-dimensional representation that is the result of PCA, features are scaled to range -1 and 1. Numbers close to the arrows show the cluster number.

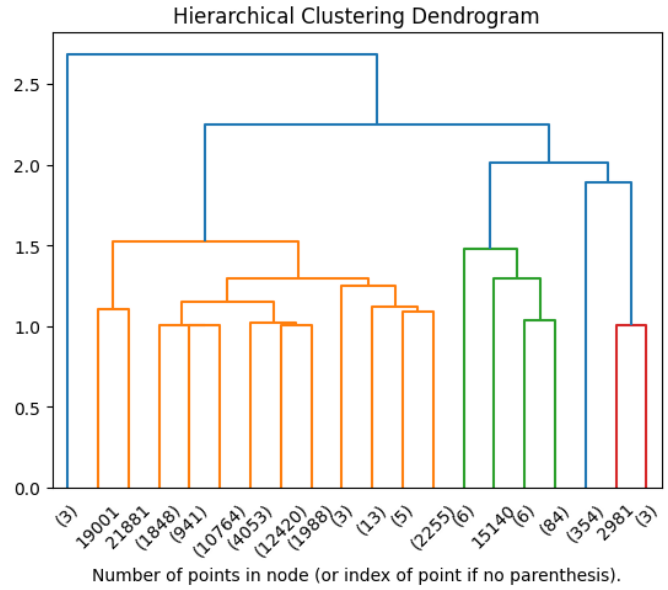

Fig. 9. Dendrogram of agglomerative clustering on 2D data achieved by PCA, number of data points within a cluster is written inside the parenthesis. Clusters without parenthesis have only one datapoint and the number written is the index of that datapoint. These types of clusters and clusters with very small number of datapoints are potential outliers that merge later to the closest cluster.

Additionally, there's the opportunity to examine the distribution of the data to determine the most suitable clustering method. Based on the 2D visualization of the data, it appears that there are distinct clusters with varying shapes, indicating that a density-based clustering method would be another suitable choice [13]. To delve deeper into potential clusters within the data, we employed the enhanced version of DBSCAN algorithm known as HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [17]. We observed improved performance when utilizing all 22 features compared to using fewer features. We also utilized the outlier detection functionality provided by the HDBSCAN library in Python. However, the identified outliers did not appear to be reasonable, which we attribute to the complexity of the dataset.

HDBSCAN identified 14 distinct primary clusters, each meticulously delineated in Figure 10. These main clusters predominantly center around a narrow selection of species, exhibiting a notable degree of purity in their composition unlike the striking lighter blue background (cluster -1), a sizable conglomerate encompassing all data points not affiliated with these main clusters. No discernible patterns emerge regarding the combination of species and their spatial distribution. For a clearer depiction of the main 14 clusters, we present them separately in Figure 11.

Despite the disparate nature of the data within this background cluster, our density-based method unified them into a single cohesive cluster. To explore this amalgam further, we applied alternative clustering techniques, namely K-means and agglomerative clustering. Remarkably, both methods yielded strikingly similar outcomes shown in Figures 12 and 13. The majority of the large dense areas are classified as the same cluster using both methods, as depicted with identical colors in both figures. Comparing the outcomes of various clustering methods to identify shared information is part of clustering ensemble problem, which is inherently more complex than comparing the outcomes of different classification methods. This complexity arises because cluster labels are symbolic, introducing the need to address a correspondence problem [18]. To tackle this challenge, we utilized the adjusted Rand Index, which quantifies the agreement between these methods in assigning clusters to data points, revealing a similarity score of approximately 0.8. We also incorporated the SOM clustering method into our analysis, Figure 14. While the similarity score between this method and the other two is slightly lower, it still demonstrates a significant degree of concordance. These methods primarily clustered the data based on the combination of species.

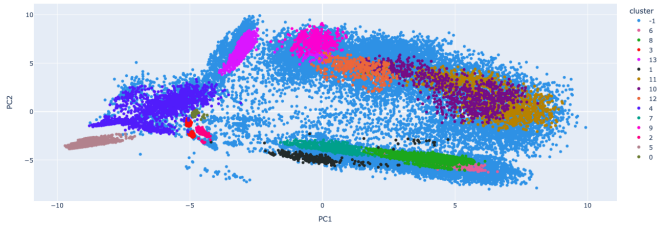

Fig. 10. Clusters obtained using HDBSCAN clustering. Logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features.

We also noted that when applying HDBSCAN to the data with all features scaled using standard scaling, one of the clusters (Cluster -1 in Figure 15) appeared exceptionally small. The data points inside this cluster seems to be far from any other cluster, indicating potential outliers. These data

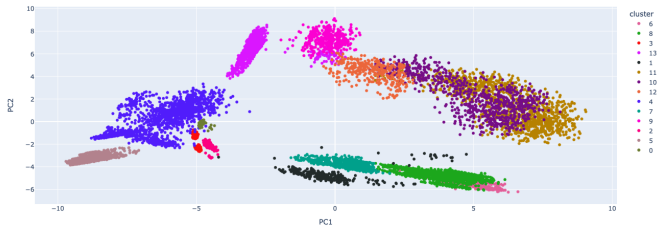

Fig. 11. 14 main clusters achieved using HDBSCAN clustering. Logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features. These are cluster 0 to 13 from Figure 10.

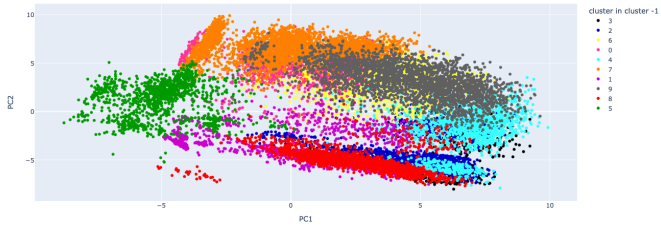

Fig. 12. Clusters achieved by K-means clustering. Logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features. K-means clustering is applied to cluster -1 from Figure 10.

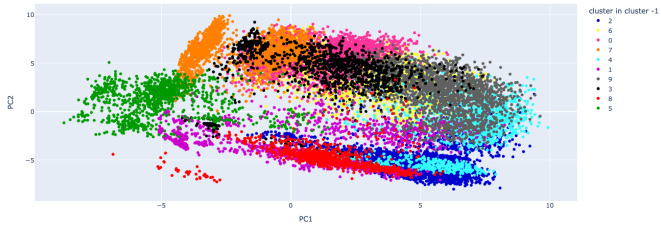

Fig. 13. Clusters achieved by agglomerative clustering. Logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features. Agglomerative clustering is applied to cluster -1 from Figure 10.

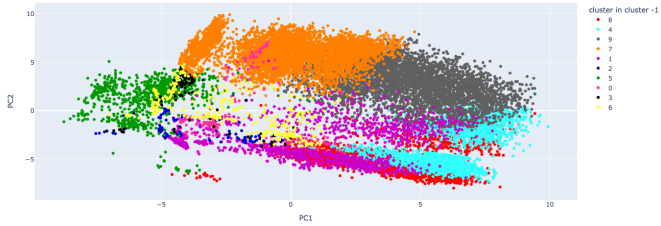

Fig. 14. Clusters achieved by SOM clustering. Logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features. SOM clustering is applied to cluster -1 from Figure 10.

points differ primarily in certain features, notably the amount of catch from the cluster they are closer to. Clusters 1, 2, and 3 exhibit higher purity in terms of species combination, encompassing only a limited number of types compared to Cluster 0, which includes all types of species. Additionally, Cluster 2 and Cluster 3 share the same geographic area and are distinct from Cluster 1.

Fig. 15. 5 clusters achieved using HDBSCAN clustering. Standard Scaler is used to scale all the features.

We also employed the two-dimensional representation generated from the RAE and applied the agglomerative clustering method to identify clusters and potential outliers. The result is depicted in Figure 16. For instance, Cluster 11 is situated between Cluster 4 and Cluster 6. The geographical area where this catch occurred aligns with Cluster 4, yet the species composition of this catch differs—it corresponds to one of the species caught in Cluster 6. Another example is Cluster 13, wherein a data point contains the same species and geographical area as Cluster 14, albeit with a lower catch amount than the minimum observed in Cluster 14. Furthermore, clusters that align along a diagonal line, such as 1, 2, 4, 16, 17, and even 11, share the same geographical area. They are a bit distant from Cluster 10 and even more so from Clusters 13 and 14.

As we discussed earlier, visualizing a dendrogram can help us recognizing potential outliers. The dendrogram for agglomerative clustering on the 2D representation obtained by RAE is shown in Figure 17.

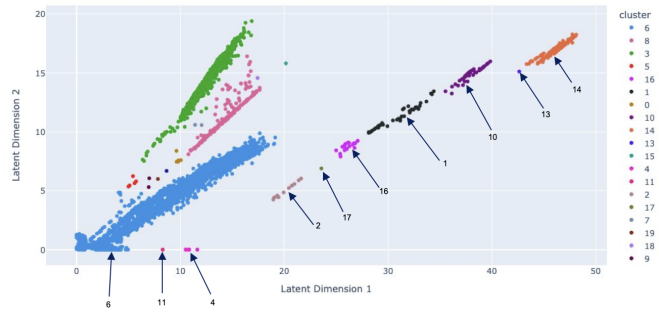

Fig. 16. Clusters and potential outliers achieved by agglomerative clustering on two-dimensional representation that is the result of using RAE on DCA data 2018. The logarithm function is used to scale round weight and Standard Scaler is used for the rest of the features. Numbers close to the arrows show the cluster number.

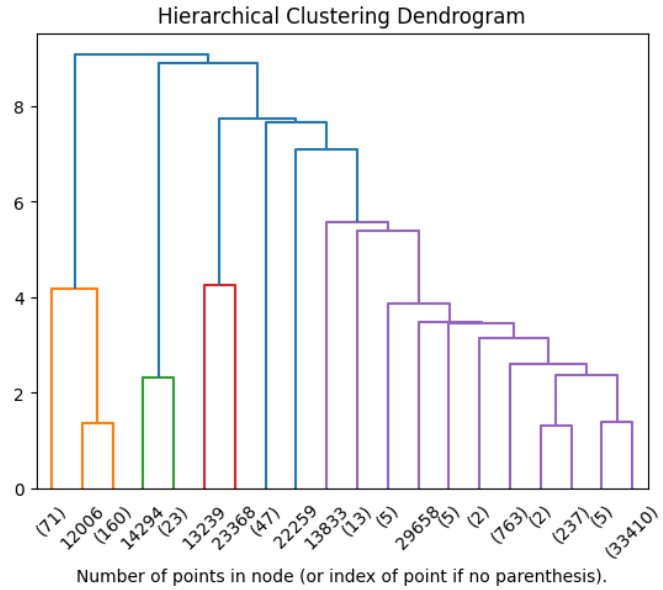

Fig. 17. Dendrogram of agglomerative clustering on 2D data achieved by RAE. The number of data points within a cluster is written inside the parenthesis. Clusters without parenthesis have only one datapoint and the number written is the index of that datapoint. These types of clusters and clusters with very small number of datapoints are potential outliers that merge later to the closest cluster.

V. CONCLUSION AND DISCUSSION

Machine learning offers a valuable tool for analyzing fishing activity reports submitted by fishermen, enabling us to identify and prevent violations of regulations and instances of overfishing. With vast datasets available from Norwegian waters, harnessing machine learning technologies holds significant potential in promoting sustainable fishing practices.

During the analysis of this complex dataset, we encountered several challenges while striving to uncover its underlying patterns. One of the most daunting features in the dataset is the variability in catch weight, influenced by a multitude of factors such as environmental conditions and potential errors made by fishermen during registration. The skewed distribution of the data adds another layer of complexity, making decisions regarding scaling, clustering, and anomaly detection more intricate.

Furthermore, the absence of labeled data restricted our choice of pattern detection algorithms. Without prior knowledge of normal reports and violations or anomalous data, we opted for an entirely unsupervised approach using clustering methods to identify clusters and potential outliers. Given the absence of an ideal definition for clusters or outliers, we experimented with various clustering techniques. While these methods exhibited a high level of agreement in identifying clusters, the identification of potential outliers differed among them. Another anomaly detection method to consider for further work can be Isolation Forest [19].

As expected, due to the intricate nature of the data and the inherent randomness involved, anomaly detection emerged as the most challenging aspect of the analysis. Although we sought assistance from experts, their input was limited due to the dynamic nature of regulations and their cautious approach in providing feedback on potential outliers at this stage.

Achieving a higher level of verification from experts would necessitate additional efforts, including detailed discussions about the desired user interface for inputting their insights. However, this process requires substantial time and resources and is thus earmarked for future endeavors.

While our focus was on reports concerning one type of gear in 2018, it's worth noting that this type of analysis can be extended to other gear types and across multiple years in the future. This approach can help explore similarities and differences over time and among different gear types.

After analyzing the dataset in our current work, we've identified a promising avenue for future research: employing transformer models for regression tasks on this tabular dataset. Additionally, upon gathering feedback from domain experts regarding anomalies, transformers can be leveraged for anomaly detection tasks having some annotated data. Given recent advancements in research focusing on attention mechanisms between data points besides attention between features, transformer models exhibit considerable potential for effectively handling tabular datasets [20], combining this technique with nearest neighbors can further enhance the efficiency [21].

REFERENCES

[1] K. Malde, N. O. Handegard, L. Eikvil, and A.-B. Salberg, "Machine intelligence and the data-driven future of marine science," *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1274–1285, 2020.

[2] N. O. Handegard, L. Eikvil, R. Jenssen, M. Kampffmeyer, A. B. Salberg, and K. Malde, "Machine learning+ marine science: critical role of partnerships in norway," 2021.

[3] E. N. de Souza, K. Boerder, S. Matwin, and B. Worm, "Improving fishing pattern detection from satellite ais using data mining and machine learning," *PLOS ONE*, vol. 11, no. 9, pp. 1–2, 2016.

[4] S. Arasteh, M. A. Tayebi, Z. Zohrevand, U. Glässer, A. Y. Shahir, P. Saeedi, and H. Wehn, "Fishing vessels activity detection from longitudinal ais data," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, C.-T. Lu, F. Wang, G. Trajcevski, Y. Huang, S. Newsam, and L. Xiong, Eds., 2020, p. 347–356.

[5] K. Shen, Y. Chu, S.-J. Chang, and S. Chang, "A study of correlation between fishing activity and ais data by deep learning," *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 14, pp. 527–531, 2020.

[6] A. Ashrafi, B. Tessem, and K. Enberg, "Detection of fishing activities from vessel trajectories," in *International Conference on Research Challenges in Information Science*. Springer, 2023, pp. 105–120.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[8] M. Alvarez, J.-C. Verdier, D. K. Nkashama, M. Frappier, P.-M. Tardif, and F. Kabanza, "A revealing large-scale evaluation of unsupervised anomaly detection algorithms," *arXiv preprint arXiv:2204.09825*, 2022.

[9] A. Barai and L. Dey, "Outlier detection and removal algorithm in k-means and hierarchical clustering," 2017.

[10] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.

[11] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[14] T. Kohonen, "Self-organizing maps," in *Springer Series in Information Sciences*, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:54122395

[15] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[16] Q. Meng, D. Catchpoole, D. Skillicom, and P. J. Kennedy, "Relational autoencoder for feature extraction," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 364–371.

[17] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[18] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

[19] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[20] J. Kossen, N. Band, C. Lyle, A. N. Gomez, T. Rainforth, and Y. Gal, "Self-attention between datapoints: Going beyond individual input-output pairs in deep learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 742–28 756, 2021.

[21] Y. Gorishniy, I. Rubachev, N. Kartashev, D. Shlenskii, A. Kotelnikov, and A. Babenko, "Tabr: Tabular deep learning meets nearest neighbors in 2023," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=rhgIgTSSxW