Dr. James Girard Summer Undergraduate Research Program

Faculty Name: Dr. Osama “Sam” Abuomar

Department: Engineering, Computing, and Mathematical Sciences (ECaMS).

Research Project Title: Classification and Self-Organizing Maps of Oceanography Data Using Supervised and Unsupervised Learning Techniques.

Research Project Summary (Please provide an overview of your project -- this will be shared with students as a project description; maximum 500 words):

In this project, several supervised and unsupervised learning techniques will be employed to validate their efficacy in acquiring information about the physical properties of oceans, and bio-physical and mechanical properties of different materials, from data derived from designed experimental studies. The dataset(s) that will be studied consist of different formulation and processing factors as inputs and different responses as outputs. In SURE 2020, visual assessment of clustering tendency (VAT) and K-means clustering, and multivariate linear regression were applied on the oceanography datasets. In SURE 2021, different techniques will be applied on the same datasets. These include, self-organizing maps (SOMs), support vector machines (SVMs), and convolutional neural networks (CNNs) or what is referred to as deep learning. SOMs will be used to extract the input(s) of the most significant effect on the output responses. SVMs will be also applied to predict certain classes (categories) that describe the dataset’s physical/mechanical behavior. In addition, linear discriminate analysis will be implemented as a dimensionality reduction technique so that certain patterns and trends can be visually discovered. Particularly, these techniques will be used to separate the specimens into different classes and patterns based on different mechanical, physical, or chemical properties. CNNs will be used to predict certain responses given a set of inputs whose output(s) is unknown or cannot be measured experimentally. This project will highlight the significance and utility of supervised and unsupervised learning techniques in the context of data analytics and knowledge discovery.
1. Introduction and Background
   1.1 Broader Impacts on Field of Study and Undergraduate Research

An increase in large datasets with high dimensionality has created a need for more efficient ways to process high volumes of data. Machine learning, defined as the “extraction of patterns or models from observed data” [Goebel and Gruenwald, 1999], is a combination of classical statistics, computer science, and large-scale data analytics [Hand, 2007]. It aids in processing large datasets, discovering new patterns and trends in the data, and classifying data or predicting new responses based on the trends found. This method of classification/ prediction can then be applied to previously unclassified or new data points to sort this data into classes or to predict some outputs for further analysis. Supervised machine learning techniques, such as the use of support vector machines or artificial neural networks (ANNs), require the use of a training set of data to design a classifier or a predictive model and a testing set of data to validate that the designed classifier/model is functional for unseen samples. Unsupervised machine learning techniques seek to look for certain patterns, groups, and behavior that can be inferred from the dataset(s) that belong to different domains. Such datasets do not have a clear distinction between input (processing or formulation) and output (response) features.

Machine learning has been proven useful in the field of oceanography, as the datasets tend to be high dimensional. Previous applications of data mining techniques have been useful on oceanographic data regarding current patterns and phytoplankton density [Healey, 1998] as well as ocean temperatures and salinity data [Huang et al., 2007]. Recently, k-means clustering has been applied to a principal component analysis (PCA) reduced set of Acoustic Doppler Current Profiler (ADCP) oceanography data in order to classify the data into clusters based on different features and determine which are the driving features between clusters and across clusters [Werr et al., 2021]. Other data mining and machine learning techniques have also been applied to ocean climate indices to find patterns and trends in the time series [Steinbach et al., 2002].

Therefore, the broader impacts of the project are to (1) continue the research that has been accomplished as part of Dr. James Girard Summer Undergraduate Research Program (SURE) 2020 as well as to continue the published research that was conducted with other students and a colleague at Coastal Carolina University (Dr. Diane Frabant) where the oceanography datasets have been taken [Werr et al., 2021], (2) develop a complete informatics based model that is able to classify and predict certain physical and mechanical properties as well as to cluster several ADCP datasets into different groups and patterns in order to ultimately conduct a sensitivity analysis where the most important features that are driving the behavior of the entire dataset can be extracted, and (3) provide a continuous yearly basis to promote undergraduate research. The premise is that this work will be published in a good peer-reviewed conference as was accomplished in 2020 and the plan is to continue this trend not only in 2021 but in the years to come as well.

   1.2 Specific Aims and Goals of the project

The goals of this project are to discover patterns in the data collected from ADCP datasets, to classify this data into the transects it was collected from based on chemical, mechanical, and physical properties, and to predict certain physical and mechanical responses based on the collected data using several supervised machine learning techniques. In addition, unsupervised learning techniques will be implemented as not only a proof of concept of what was accomplished in SURE 2020, but also to conduct further clustering and sensitivity analyses on two new ADCP datasets that were not studied before in order to discover new patterns and trends and to extract the most important features that are driving the behavior of the entire dataset(s) [Hall and Smith, 1998].

2. Proposed Research Plan
   2.1 Experimental design and Methodology
The acoustic Doppler current profiler (ADCP) has been used in several applications to detect ocean currents and circulation [Ursella and Gacic, 2001], sediment transport [Kostaschuk, et al., 2005], and zooplankton distribution [Lorke, et al., 2004]. This instrument emits a fixed frequency and measures the echoes off of sound scatterers. These scatterers reflect sounds back to the ADCP, and can be any small particles such as particulate matter, copepods, and euphausiids. It then calculates the velocity in three directions using the Doppler shift, or the difference between the frequency heard from an object that is static and the frequency heard from an object in motion [Teledyne RD Instruments 2011]. The ADCP operates under the assumption that the scatterers are moving at the same speed as the water’s speed.

The ADCP, manufactured by Teledyne RD Instruments, uses four beams in order to obtain velocity measurements in three directions: north-south, east-west, and vertical. Each beam is pointed in a different direction to obtain a different velocity component. Vertical velocity is measured twice in order to obtain an error velocity. In addition, the ADCP is able to use bottom-tracking capabilities to determine ship location and velocity based on the sea floor in addition to current velocity based on scatterers in the water. However, there are some limitations to the ADCP. Because it operates on sound scatter and detection, any obstruction to the acoustic signals will potentially affect the data. Factors such as the fluid’s attenuation and penetration of the acoustic pulse, the concentration of scatterers in the water, and the amount of bubbles near the ADCP can reduce the data quality received by the ADCP [Flagg et al., 1998]. Bubbles can attenuate and scatter both incoming and outgoing signals, thus skewing the data and potentially causing a large shear [Flagg et al., 1998, New, 1992]. However, corrections may be applied to help adjust for the shear and reduce the bias of the data should it occur.

2.1.1 Data Collection:

Data was collected using a 300 kHz ADCP (Teledyne RD Instruments) on two cruises by Dr. Diane Fribance of the Department of Marine Science at Coastal Carolina University in July 2014 off the coast of McIntosh County, GA. The first cruise covered transects A, B, and C1-C9, with bin size 0.5 meter. The second cruise covered transect C with bin size 1 meter (Fig. 1). Bin size refers to the depth intervals at which data was collected. Data collected includes temperature, location coordinates, date and time, pitch, roll, heading, north velocity, east velocity, vertical velocity, error velocity, bottom-tracking displacement, bottom-tracking north, east, vertical, and error velocities.

Figure 1. Map of transects covered by Dr. Diane Fribance. Data from SAV_14_18001 dataset was taken from transects A, B, and C1:C9. Data from SAV_14_18002 dataset was taken from transect C.

2.1.2 Methods of data analysis:

In this project, several supervised and unsupervised machines learning techniques will be applied on the oceanography data. These include artificial neural networks (ANNs), convolutional neural networks
(CNNs), Self-Organizing maps (SOMs), and Support Vector Machines (SVMs) [Tong and Koller (2001), Abuomar et al. (2013), and Ravì et al. (2016)].

Learning in an ANN can occur in either a supervised or an unsupervised fashion. A supervised approach uses a learning algorithm that creates an input/output mapping based on a labeled training set; thus, creating a mapping between an \( n \)-dimensional input space (i.e., different experimental conditions in this case) and \( m \)-dimensional output space (certain physical and mechanical responses). In this case, the network will learn a functional approximation from the input/output pairings and will have the ability to recognize or classify a new input vector into a correct output vector (generalization). An unsupervised learning architecture, in contrast, presents the network with only a set of unlabeled input vectors from which it must learn. That is, the unsupervised ANN is expected to create characterizations about the input vectors and to produce outputs corresponding to a learned characterization (i.e., knowledge discovery).

ANNs that use unsupervised learning will determine natural clusters or feature similarity within the input dataset and to present results in a meaningful manner. Since no labeled training sets are used in this approach, the outputs from the unsupervised learning network must be examined by a domain expert to determine if the classification provides any new insight into the dataset. If the result is not reasonable, then an adjustment is made to one of the training parameters used to guide the network’s learning, and the network is presented the patterns again.

Deep learning has in recent years set an exciting new trend in machine learning. The theoretical foundations of deep learning are well rooted in the classical ANN literature. But different to more traditional use of ANNs, deep learning accounts for the use of many hidden neurons and layers (typically more than two) as an architectural advantage combined with new training paradigms. While resorting to many neurons allows an extensive coverage of the raw data at hand, the layer-by-layer pipeline of nonlinear combination of their outputs generates a lower dimensional projection of the input space. Every lower-dimensional projection corresponds to a higher perceptual level. Provided that the network is optimally weighted, it results in an effective high-level abstraction of the raw data or images. This high level of abstraction renders an automatic feature set, which otherwise would have required hand-crafted or bespoke features.

SOMs, sometimes referred to as Kohonen maps are utilized to map patterns of arbitrary dimensionality into 2-D or three-dimensional (3-D) arrays of neurons (maps). A SOM may be thought of as a self-organizing cluster. The basic components of a 2-D SOM for assessing ADCP oceanography data are shown in Fig. 2. The inputs are the dimensions of the dataset being analyzed. Note that each element of the input vector \( x \) is connected to each of the processing units on the map through the weight vector \( w \).

After training, the SOM will define a mapping between the ADCP input data space and the 2-D map of neurons. The ADCP output feature \( y \) of a processing unit is then a function of the similarity between the input vector and the weight vector. The nonlinear mapping of the SOM utilizes a technique developed by Sammon that preserves the higher dimensional closeness on the map.

In Fig. 2, a trained feature map and its response to a winning output neuron, when excited by an original training pattern or an unknown similar input vector pattern, is shown. This figure is a general illustration to show the logic of the SOM and the ANN techniques. Knowledge about the significance of the area around the winning neuron will then help the domain expert in knowledge discovery.

![Figure 2](image_url). Representation of the ADCP data analysis using ANN and a SOM. In the processing unit, the input vector \( x \) is multiplied by the weight vector \( w \) to create a mapping to the output vector \( y \).
The goal of an SVMs classifier is to define a separating hyperplane between the points belonging to two or more distinct classes (different ranges of transects as shown in Fig. 1) and maximize the distance between these points and the hyperplane. This maximum distance is referred to as the margin. This concept is illustrated in Fig. 3 for non-linearly separable data.

![Figure 3](image)

**Figure 3.** An example of the SVMs model for non-linearly separable data.

The margin \( m \) (indicated in Fig. 3) is given by the relation:

\[
m = \frac{|g(x)|}{||w||}
\]

where \( g(x) \) is the discriminant function used to separate and classify the data vectors into corresponding classes and \( w \) is the weight vector used by SVMs model. The weight vector is scaled so that the value of \( g(x) \) at the closest point to the separating hyperplane is equal to 1 for class one and -1 for class two.

Linear discriminant analysis (LDA) is a predictive modeling algorithm for multi-class classification. It can also be used as a dimensionality reduction technique, providing a projection of a training dataset that best separates the examples by their assigned class [Vanhatalo et al., 2017]. In practice, LDA for multi-class classification is typically implemented using the tools from linear algebra, and like PCA, uses matrix factorization at the core of the technique [Bro and Smilde, 2014]. If time allows, this study will use LDA to visualize and analyze multidimensional data obtained from the ADCP data.

### 2.2 Preliminary Results

As described above, this work is a continuation of SURE 2020 project. In SURE 2020, PCA was applied to reduce dimensionality and extract the most common and important features of ADCP dataset. K-means clustering was applied on the dataset to cluster the data points into distinct clusters (groups). The main common features across clusters were bottom tracking range, navigation north velocity, heading, and error velocity. Virtual assessment of cluster tendency algorithm (VAT) and improved virtual assessment of cluster tendency algorithm (iVAT) were applied after k-means to determine that the data in fact could be clustered as well as to give an estimate of the optimal number of potential clusters. In addition, multivariate linear regression was used to examine the influence of one or more independent variables on a dependent variable in ADCP dataset.

### 2.3 Expected Outcomes

The student who will participate in this project should build upon what was accomplished in SURE 2020 in order to develop a complete informatics-based model where a hybrid of supervised and unsupervised machine learning techniques can be implemented on several ADCP datasets. These techniques include LDA, SOMs, SVMs, and CNNs as outlines above (section 2.1.2) where the student is expected to understand how machine learning can be used in classification, prediction, clustering, and sensitivity...
analysis when applied on several datasets that belong to either one domain or several domains. In addition, by the end of the project, the SURE student should be able to not only deliver informative and professional presentations but also to contribute effectively to scientific writing/reporting as well.

### 3. Mentorship Plan

The student is expected to read some background and preliminary material distributed by the faculty mentor in order to prepare him/her for the project. Then, the student will be required to conduct a thorough literature review in order to discover other contributions of data science techniques and algorithms on oceanography ADCP data including what has been accomplished in SURE 2020 and published by Lewis students [Werr et al., 2021]. This helps both the student and faculty mentor to align a novel research agenda that has not been explored *a priori*.

The next step involves applying and developing all the new machine learning and data mining techniques explained by the faculty mentor on the oceanography ADCP data. In addition, the student has to meet with the faculty mentor in a regular basis (once a week) in order to check the accuracy of the results and/or if the student has any questions or concerns about the obtained results or about the project in general.

Furthermore, in order to enrich the writing and presentation skills, the student should conduct a 1-slide presentation at the beginning of the project in order to check his/her understanding of the basic concepts and experiments that will be studied and conducted. In addition, he/she has to write a separate draft of the introduction, methods, and literature review to be included in the final project report. After all the required analysis and tasks are accomplished and approved by the faculty mentor, the student is expected to write a detailed project report to be reviewed by the faculty mentor and/or other committee(s) determined by the SURE program faculty panel.

Finally, before conducting the final presentation in the Summer Research Symposium, the student is required to prepare a mock poster/PowerPoint presentation to prepare him/her for the final meeting at the Symposium. The timeline of the above tasks is illustrated in section 4 (Proposed Timeline with Project Goals).

### 4. Proposed Timeline with Project Aims and/or Goals

The following is a *tentative* timeline along with the project’s goals:

<table>
<thead>
<tr>
<th>Date (Approximate)</th>
<th>Project’s Goals</th>
</tr>
</thead>
</table>
| June 12            | One-slide project presentation (assign June 5)  
*Read one paper related to project topic along with other background material*  |
| June 23            | Finish data cleaning and pre-processing and get the required results  |
| June 30            | Draft of Introduction, Methods, and Literature Review (assign June 13)  |
| July 7             | Finish Support Vector Machines along with all results and start on SOMs  |
| July 14            | Finish SOMs and start on CNNs  |
| July 21            | Finish CNNs along with all results and analysis  |
| July 24            | Mock poster presentation (assign July 1)  |
| July 27            | Conduct linear discriminant analysis (LDA) for dimensionality reduction (if time allows)  |
| July 28            | Draft of Technical Report including all results and analysis (assign June 30; after the Introduction, Method, and Literature Review draft).  |
| August 1           | Final Technical Report  |
| August 5-9         | Summer Research Symposium Presentation  |
5. References


