Explainable Contextual Data Driven Fusion

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

MATTHEW S. DEARDORFF

Dr. Derek Anderson, Thesis Supervisor

May 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

Explainable Contextual Data Driven Fusion

presented by Matthew S. Deardorff,

a candidate for the degree of Master of Science and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Derek T. Anderson

Dr. James M. Keller

Dr. Grant J. Scott

Dr. Mihail Popescu

ACKNOWLEDGMENTS

I've received a wealth of support throughout my time in graduate school. I would first like to thank my professor, Dr. Anderson, for providing the guidance, expertise, and resources required for me to be able to complete my requirements. I would also like to thank my mother and father, who encouraged me from the start to pursue whatever subject I've shown interest in while ensuring personal finances were never a hindrance. Finally, I would like to thank my friends, who have kept me sane and humble throughout these past years.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

ABSTRACT

Numerous applications require the intelligent combining of disparate sensor data streams to create a more complete and enhanced observation in support of underlying tasks like classification, regression, or decision making. This presentation is focused on two underappreciated and often overlooked parts of information fusion, explainability and context. Due to the rapidly increasing deployment and complexity of machine learning solutions, it is critical that the humans who deploy these algorithms can understand why and how a given algorithm works, as well as be able to determine when an algorithm is suitable for use in a particular instance of the problem. The first half of this paper outlines a new similarity measure for capacities and integrals. This measure is used to compare machine learned fusion solutions and explain what a single fusion solution learned. The second half of the paper is focused on contextual fusion with respect to incomplete (limited knowledge) models and metadata for unmanned aerial vehicles (UAVs). Example UAV metadata includes platform (e.g., GPS, IMU, etc.) and environmental (e.g., weather, solar position, etc.) data. Incomplete models herein are a result of limitations of machine learning related to under-sampling of training data. To address these challenges, a new contextually adaptive online Choquet integral is outlined.

Chapter 1

INTRODUCTION

In this age of data-driven machine learning we are experiencing an explosion of complexity in both the domains in which we apply our solutions, as well as the solutions themselves. The fact that we are able to tackle more sophisticated problems such as creating machines which can understand images or language is encouraging. As we solve these high-level problems, the applications to which we can apply machine learning grows to new heights. However, the fact that the complexity of the solutions themselves are growing at a similar or even higher rate can be alarming. For example, Google's BERT language model[1] is one of the industry-leaders in natural language processing, and has over 340 million parameters. These machines, where the user is only responsible for putting data in one end and receiving an answer out the other, are often called *black-box* algorithms. The term black-box refers to the fact that the user is ignorant of the machines inner-workings, and treats the specifics of how an answer was generated as being too complex to truly understand. This is a side-effect of the machine-learning field's focus on big data. As we tackle larger and larger problems, it is seen as a simple necessity that our algorithms grow in size to accommodate.

In contrast to these black-box solutions, we have the concept of a *glass-box* (sometimes called white-box) solution. That is, an algorithm that is transparent to its inner-workings, in such a way that they can be analyzed and understood by a user. There is already some nuance in the difference between these two definitions. For

example, the BERT model mentioned above can, at a literal level, be analyzed on a per-input basis to see how each of the 340 million parameters behaves. We are not truly ignorant of how the machine works (it was designed by humans after all), but that does not mean we have a good idea of *why* the machine works as well as it does. This question of *why does an algorithm perform well* is one of the primary tenets of a field called explainable AI (XAI). XAI is a subset of the machine learning field which concerns itself with generating glass-box solutions so that a user may better understand why and how an algorithm is able to solve a problem.

It is worth further dwelling on the kinds of questions we might wish for XAI to address. For example, a question as simple as *why* did you produce this answer is something that nearly any type of neural network is incapable of answering, yet would be undeniably useful in nearly any domain. Conversely, why did the algorithm *not* produce an alternative, seemingly-reasonable answer? One could imagine an image classifier that not only can distinguish between cats and dogs, but also provides human-readable explanations as to why a that determination was made (because the cat had whiskers! or, that's not a dog because it has scales!) We might ask of an algorithm *when* is it a good idea to listen to its solution, and when is it completely out of its depth? Can our cat-dog classifier recognize when it is ill-equipped to make a classification? This is something that is somewhat addressed in neural networks as many are able to provide a "confidence" level with an associated label, though it is still common to see algorithms which predict wrong labels at a high confidence[2]. Finally, (and perhaps the most desirable from an algorithm design perspective) we might ask the question of what needs to be done to *improve* this algorithm? If a machine were able to direct a human as to why it produced an incorrect solution, or what could be done to remedy this in the future, who knows what would be possible.

To explore how some of these questions can be answered, the rest of this thesis concerns itself with a specific tool called the fuzzy integral, specifically the Choquet

Integral (ChI). The ChI is an aggregation operator, which means it takes as input a number of sources (which can be thought of as evidence supporting a particular hypothesis) and combines them based on an estimation of the worth of each subset of these sources. For example, one could imagine asking three individuals: a biologist, a physicist, and a radiologist, to analyze an MRI scan and determine whether a patient has cancer. In this scenario we might be most willing to trust the radiologist's opinion as they are well suited to the problem, while trusting the physicist the least. Here we would take each of the experts' answers (yes the patient has cancer, no they do not, or maybe but it's hard to tell) and combine them into a single answer that is hopefully more trustworthy than any of them individually. In the interest of XAI we'd like to know why the answer was aggregated the way it was, how was that result calculated, and is it even an appropriate question to be asking these experts.

In working with the ChI, we do have a few advantages which can more easily enable XAI compared to million-parameter neural networks. One is that the ChI is an efficient encoding of a large number of fusion strategies. For N inputs, the ChI encodes $N!$ operators with $N(N!)$ parameters using only 2^N parameters. It is not often that we perform fusion on a large number of sources, so typically the number of parameters in a ChI are few enough that a human user can understand what is happening. Secondly, the ChI selects a fusion strategy based on a sorting of the input data. To this extent, one can think of the sort as the *context* of when a particular fusion is appropriate. For example, if an IR camera has a higher return than an RGB camera, that would result in one fusion strategy, while if the RGB camera were higher we would fuse differently.

In the rest of this paper we repeatedly come back to the idea of context informing the fusion process. We make the distinction between *internal* context and *external* context. We define internal context to be conditions necessary to utilize a specific operator within a ChI, e.g., the sort order. The external context is the conditions

under which the observations to be fused were made. For example, the fusion strategy for combining an IR and RGB camera is going to be different at noon compared to midnight. In this explore the role context plays in fusion, and how we can utilize that context to create more explainable machine learning solutions.

Chapter 2 explores the use of the Earth Mover's Distance (EMD) as a similarity measure for Linear Order Statistics (LOSs) and ChIs. We argue that the EMD maintains semantic expectations for similarity we would have in the domain of LOSs, something that a simple ℓ_p norm does not provide. The improved similarity measure allows us to use a decomposition process to determine the kind and quantity of unique operators within a single ChI. We further our goals of explainable AI by exploring what was learned from a data-derived integral, as well as attempting to visualize the parts of an integral in a way that is comprehensible and useful to the user.

Chapter 3 describes a system that uses metadata information which would otherwise go unused to determine the external contexts present in the data. Given that we can identify the unique scenarios in which this data was observed, we can construct specialized algorithms which perform well in their specified contexts, but are not required to complete the more difficult task of generalizing to all contexts in which the problem may present itself. This system also furthers our goal of explainable AI for a few reasons. If a context is encountered at run-time that was not present in the training data, our system is able to alert the user that it may be ill-prepared to handle that specific piece of data. Additionally, our system can be used to determine if certain algorithms are only relevant within a small subset of the data, or if they tend to perform well across the board.

Chapter 2

EARTH MOVER'S DISTANCE AS A SIMILARITY MEASURE FOR LINEAR ORDER STATISTICS AND FUZZY INTEGRALS

2.1 INTRODUCTION

Aggregation operators, like the *linear order statistic* (LOS)[1], *the ordered weighted average* (OWA)[2], *and the fuzzy integral* (FI), are widely used for tasks like regression and classification in contexts such as multi-criteria decision making and image processing. These operators combine data (aka inputs) relative to knowledge about the utility of the individuals and their interactions. An aspect of using these is, in our modern era of data-driven artificial intelligence (AI) and machine learning (ML), algorithm users are reluctant to deploy a technique if it is opaque and cannot be explained or trusted. Whereas a great deal of effort has gone into understanding the above operators at a fundamental level (e.g., [3]), there is more to be understood relative to their derivation from data. In the current paper, we confront data-driven XAI for fusion by exploring how to measure and use similarity within and between linear convex sums (LCSs), LOSs, and FIs.

The reader can refer to [4] for a recent survey on explainable AI (XAI) and the kinds of questions it allows us to answer, such as why did the algorithm learn what

[1]The LOS is also referred to commonly as linear functions of order statistics and linear combinations of order statistics.

[2]When the input and weights are real-valued numbers and not fuzzy sets, the OWA is an LOS.

it did, or *how do we explain* to the user what an algorithm is doing. In the past few years, our group has explored XAI for information fusion. Specifically, we have proposed statistical, visual, and linguistic XAI fusion explanations relative to tasks like classification and regression [5, 6, 7, 8, 9, 10, 11, 12].

The XAI challenge confronted in the current section is, how do we unearth and communicate a minimal set of underling logic behind a learned instance of the *Choquet integral* (ChI)? This XAI task fits into the category of generating local explanations [4]. For example, consider N sources of information, $X = \{x_1, x_2, ..., x_N\}$, which provide the input $\mathbf{h} = (h_1, h_2, ..., h_N)$. Herein, we let $h_i = h(x_i)$ for sake of equation simplification. It can be shown that the ChI can be decomposed into $N!$ underlying LCSs (one for each input sort) with 2^N variables (number of fuzzy measure (FM) variables). Without loss of generality, consider $N = 3$ inputs, where these inputs have the relation $h_2 > h_1 > h_3$. The output of the ChI for this input sort is the linear function $w_1h_2 + w_2h_1 + w_3h_3$. An important detail here is that the decomposition and resulting equation can be interpreted and explained. Interpretability of the entire process hinges on if the inputs are interpretable. The aspect that is unique to this paper is that in practice most applications do not make use of the full representational capability of the ChI, i.e., use all $N!$ unique LCSs. Usually, the ChI breaks down into a single, or handful of, simple, context-dependent operators like the max (t-conorm), min (t-norm), mean, soft and trimmed variants, or a more unique LOS. The question we aim to answer is, for a given set of data, which operator(s) did the ChI learn? Our goal is to demystify the ChI, helping inform users in their specific application domains about what their ChI is doing. Our current paper addresses this by proposing a similarity measure designed for the ChI, and then by using this to provide discovery and communication of the underlying hidden structure/logic in ChIs that are learned from data.

While we are interested in measuring similarity between any LOS and ChI, it

is helpful to restrict our discussion, and at moments only consider, the canonical aggregation operators **max**, **median**, **mean**, and **min**. These operators are different points on the LOS and ChI "spectrum," from the smallest of the inputs to the largest. They represent a tractable set of important commonly encountered operators that we believe outline a set of relationships that help us establish the semantics of similarity in this context. If we cannot satisfy these relationships then it is unlikely that a proposed similarity measure will work on more unique LOSs/ChIs. To this end, we consider the following three touchstones:

(T1) The **max** and **min** operators are least similar to each other and should have a similarity of 0.

(T2) Similarity between equivalent operators should be 1.

(T3) Similarity between (**max**, **median**) and (**min**, **median**) should be greater than 0 but less than 1, as the **median** is somewhat similar to the other two operators.

These touchstones are based on our intuitive understanding of what a partial ordering of LOSs and ChIs should look like, as we could not find any literature proposing an ordering structure across all LOSs or ChIs.

The remainder of this chapter is organized as follows. First, we succinctly review the LOS, Euclidean distance, and the EMD. Next, we review the ChI, the EMD is extended to measure similarity between ChIs, and clustering is proposed to discover underlying similarity structure within an integral that can be easily communicated to people. Lastly, synthetic and real-world experiments are provided to show the effectiveness of the proposed ideas.

2.2 LINEAR ORDER STATISTIC

Let X be a set of N sources of information, e.g., from humans, algorithms, or sensors. Furthermore, let h_i be the input from source i, e.g., a subjective belief, objective sensor

measurement, classifier output, etc. An LCS is defined as

$$f_{LCS}(\mathbf{w}, \mathbf{h}) = \sum_{i=1}^{N} w_i h_i, \tag{2.1}$$

where $\mathbf{w} = (w_1, ..., w_N)^t \geq \mathbf{0}^N$ and $\left(\sum_{i=1}^{N} w_i\right) = 1$.

The LOS is

$$f_{LOS}(\mathbf{w}, \mathbf{h}) = \sum_{i=1}^{N} w_i h_{\pi(i)}, \tag{2.2}$$

where π is a sorting such that $h_{\pi(1)} \geq h_{\pi(2)} ... \geq h_{\pi(N)}$. Thus, a LOS is a LCS with a "pre-sort" where $\mathbf{w} = (w_1, ..., w_N)^t \geq \mathbf{0}^N$ and $\left(\sum_{i=1}^{N} w_i\right) = 1$.

2.2.1 ℓ_p-Norm as an LOS Proximity Measure

The frequently used ℓ_p-norm is a rational place to start when it comes to capturing similarity between pairs of LCSs or LOSs. In this section, we show that while this norm can be useful as a baseline, it has semantic issues that leave us desiring a superior measure.

The ℓ_p-norm d between two vectors (which herein represent LOS coefficients) $\mathbf{w}_j$ and $\mathbf{w}_k$, where $\mathbf{w}_j, \mathbf{w}_k \in \mathcal{R}^N$, is

$$d(\mathbf{w}_j, \mathbf{w}_k) = \left(\sum_{i=1}^{N} (w_{ji} - w_{ki})^p\right)^{\frac{1}{p}}, \tag{2.3}$$

which can be converted into a similarity[3] (if desired) via

$$s(\mathbf{w}_j, \mathbf{w}_k) = \frac{\rho - d(\mathbf{w}_j, \mathbf{w}_k)}{\rho}, \tag{2.4}$$

where ρ is the maximum allowable distance.[4] Second, the ℓ_p-norm operates on a

[3]It is important to note that not all dissimilarity (or similarities) can be converted to their dual on such a simple premise.

[4]For example, $\rho = \sqrt{2}$ for the ℓ_2-norm, $w_i \geq 0$, and $(\sum_{i=1}^{N} w_i) = 1$.

per-bin basis, meaning only $\mathbf{w}_i^1$ and $\mathbf{w}_i^2$ are compared, with no interaction considered between $\mathbf{w}_i^1$ and $\mathbf{w}_j^2$, where $i \neq j$. This lack of cross-bin interaction is often acceptable for comparing LCSs if there is no implicit relation between bins. However, this is not the case for LOSs due to their sort, meaning information is lost in this comparison.

Consider our touchstones mentioned in Section I (T1-T3); the question is, which of these do the ℓ_p-norm satisfy? Without loss of generality, we use $N = 3$ in the following to demonstrate compactly our points.

(T1) The max and min operators $\mathbf{max} = (1, 0, 0)$ and $\mathbf{min} = (0, 0, 1)$, result in $s(\mathbf{max}, \mathbf{min}) = 0$ meaning these are as dissimilar as possible. However, note that we get a maximal distance in many cases where there is no overlap in the non-zero coefficients.

(T2) The similarity $s(\mathbf{max},\mathbf{max})$ is 1, meaning that the ℓ_p proximity measure has a satisfying upper bound.

(T3) The similarity $s(\mathbf{max},\mathbf{median})$ and $s(\mathbf{min},\mathbf{median})$ are both 0, which is not what we would expect. This means that the ℓ_p based similarity does not enforce the same ordering we relate to the max, median, and min. This is the core issue that leads us to finding a superior measure.

In summation, the ℓ_p-norm behaves favorably when the LOS weight vectors overlap. However, it falls short of many semantic expectations due to the fact it ignores the interactions across bins, a scenario we remedy in the next section.

2.2.2 Earth Mover's Distance on LOSs

In this section, we evaluate the EMD to remedy shortcomings identified in Section II. The EMD is a measure of divergence between two distributions.[5] It is based on

[5]We refer to these entities as histograms hereafter versus distributions or signatures or etc. Typically, the nomenclature depends on the application and/or properties, e.g., probabilistic for positivity and sum to one.

a solution to the well-known transportation problem, i.e., the Monge-Kantorovich problem. In [13], Rubner introduced the EMD in the context of *content based image retrieval* (CBIR) for unequal mass distributions. In [14], Levina and Bickel proved that the EMD, a parametric measure, is equivalent to the Mallows and Wasserstein distance for the case of two probability distributions and it is different when applied to unnormalized distributions of different masses (e.g., *signatures* in CBIR). We use the EMD herein as it enables cross-bin interactions during dissimilarity, a concept not possible with an ℓ_p-norm, Jaccard or Dice measures, etc.

Let $\mathbf{h}$ be a (one-dimensional) histogram of length L_1, $h_i \in \Re+$, $1 \leq i \leq L_1$; and let $\mathbf{b}$ be a second histogram of length L_2, where $b_i \in \Re^+$. $EMD(\mathbf{h}, \mathbf{b})$ is the EMD between $\mathbf{h}$ and $\mathbf{b}$. The goal is to find a flow $F = [f_{ij}]$, where f_{ij} is the *flow* between h_i and b_j, which minimizes

$$WORK(\mathbf{h}, \mathbf{b}, F) = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} d_{ij} f_{ij}, \tag{2.5}$$

subject to

$$f_{ij}, \qquad 1 \leq i \leq L_1, 1 \leq j \leq L_2, \tag{2.6a}$$

$$\sum_{j=1}^{L_2} f_{ij}, \leq h_i \qquad 1 \leq i \leq L_1, \tag{2.6b}$$

$$\sum_{i=1}^{L_1} f_{ij}, \leq b_j \qquad 1 \leq i \leq L_1, \tag{2.6c}$$

$$\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} f_{ij} = \min\left(\sum_{i=1}^{L_1} h_i, \sum_{j=1}^{L_2} b_j\right), \tag{2.6d}$$

where $D = [d_{ij}]$ is called the *ground distance*. Once the transportation problem is

solved—i.e., the optimal F^* is found—the resulting EMD is

$$EMD(h, g) = \frac{\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} f_{ij}^* d_{ij}}{\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} f_{ij}^*}. \tag{2.7}$$

Herein, d_{ij} is the distance between "bins" (indices), e.g., $d_{11} = 0$, $d_{12} = 1$, $d_{13} = 2$, $d_{1L} = L - 1$, etc. There is no cost to stay in the same bin, which increases by one per bin thereafter. The EMD can be converted into a similarity via

$$s(\mathbf{w}_j, \mathbf{w}_k) = \frac{\rho - EMD(\mathbf{w}_j, \mathbf{w}_k)}{\rho}, \tag{2.8}$$

where the maximal allowable distance $\rho = L - 1$.

The primary benefit of using the EMD is that it allows us to define a ground distance matrix in a way that mirrors the sort induced by LOSs. The ground distance matrix asserts that bins that are adjacent to each other are "closer" than bins that are non-adjacent, the same way values that are similar are sorted to be near each other. As a result, this ground distance matrix changes the distance topology in a way that makes the EMD satisfy each of our touchstones, which we now describe.

(T1) $EMD(\mathbf{max}, \mathbf{min})$ is the scenario in which the entire "mass" has to be moved across the largest possible number of bins. With the constraint that the sum of weights is 1, we are moving the largest possible mass the farthest possible distance; hence, the EMD is maximized and the similarity is 0.

(T2) $EMD(\mathbf{max}, \mathbf{max}) = 0$ as there is no change in the coefficients, no work is required. The corresponding similarity is 1, resulting in the upper bound.

(T3) $EMD(\mathbf{max}, \mathbf{median}) = 0.5, EMD(\mathbf{min}, \mathbf{median}) = 0.5$, as these cases result in the entire mass of their respective histograms being shifted half the maximum possible distance.

Based on the above criteria (extreme bounds), which could clearly be expanded to include other desirable properties like monotonicity, idempotency, etc., the EMD is a more suitable distance measure for LOSs than the ℓ_p-norm, or other bin-to-bin measures at that. What is the disadvantage of using it? First, the EMD is undefined for distributions with negative values. While LOSs are usually constrained to have non-negative values, they do pop up in cases of regression such as in [15]. Additionally, the calculation of the EMD is computationally more expensive than an ℓ_p-norm, with [16] running in $O(n^2)$, though approximation algorithms such as [17] can run in linear time.

The difference between the EMD and an ℓ_p-norm can be visualized by calculating the distance between a point and the set of possible LOSs that exist in 3-space, shown in Figure 2.1. If we treat each axis as a different weight, we obtain a triangular plane that describes the set of possible LOSs with weights that sum to one. If we color-map this plane based on the distance from specific points, we can see how the topology changes based on the distance measure. The plot in (b) shows how under an ℓ_p each extreme operator is equidistant from the **mean**. This is counter-intuitive as we more closely associate the **median** with the **mean** rather than the **max** or **min**. In the EMD plot, we can see that this is correctly modeled as the triangle vertex corresponding to the **median** is closer than the other two vertices. We can see something similar in plots (c) and (d) where the **max** operator appear equidistant from the other two vertices under an ℓ_p norm, but not the EMD.

2.3 EMD BETWEEN CHIS

In this section, the EMD is extended to integrals via the idea of decomposing the ChI into its underlying set of LCS and corresponding LOSs. To the best of our knowledge, the only prior work on distances between FMs is based on the Hellinger distance [18, 19]. In 1909 Ernst Hellinger introduced a distance measure for two

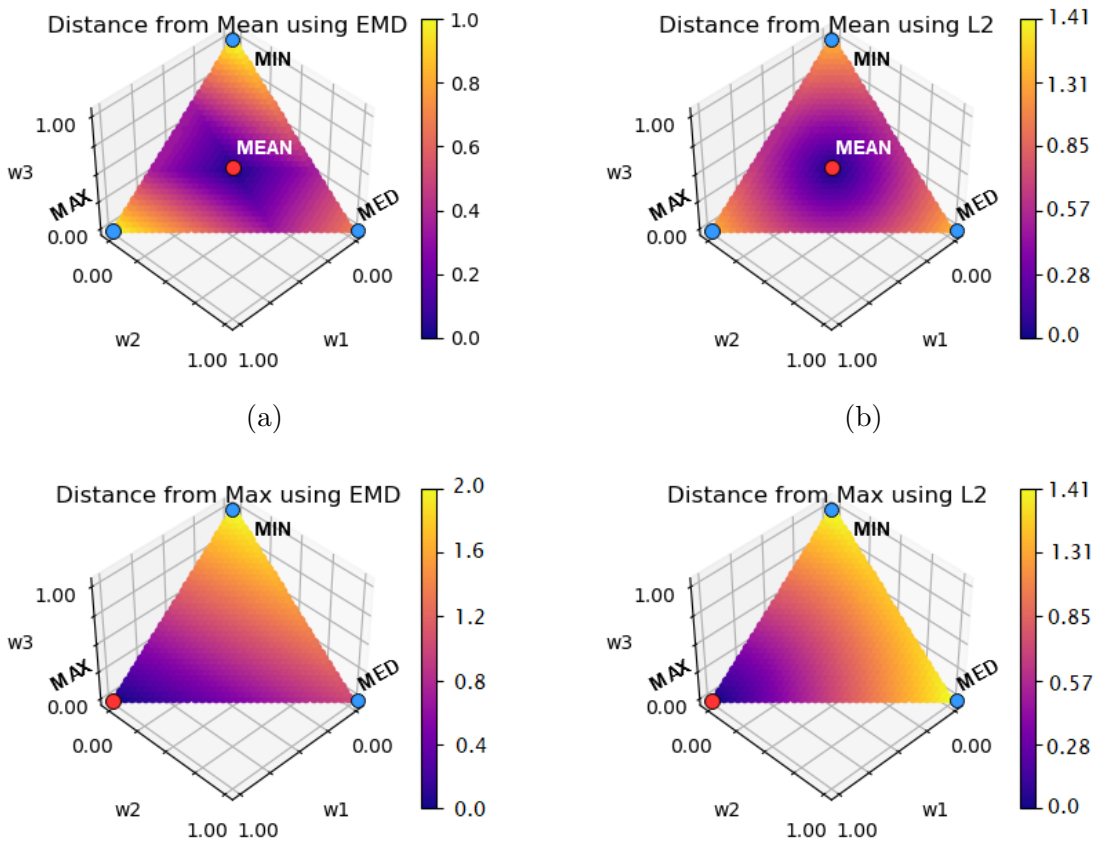


Figure 2.1: Visualizing the set of possible LOSs as a plane can help us understand what impact the EMD has on distance topology. In these graphs, each axis represents a variable and the color of a pixel on the plane represents the distance from a specified point to that location. In (b) and (d) we see that an ℓ_2 norm considers each axis to be equally distant from the others, where as in (a) and (c) it is more expensive to move from weight 1 to weight 3 than weight 1 to weight 2.

probability distributions (additive measures). In 2013, [18] Torra et al. explored its extension to FMs. While the authors introduce a variation of the Hellinger distance for non-additive measures, which relies on a Radon-Nikodym-type derivative for the FM, ultimately their distance measure works on a bin-to-bin basis and therefore it does not satisfy the semantic touchstones considered herein.

2.3.1 Choquet Integral

The FM, $g : 2^X \to R^+$, is a function with two properties[6]: (i) (boundary condition) $g(\emptyset) = 0$, and (ii) (monotonicity) if $A, B \subseteq X$, and $A \subseteq B$, then $g(A) \leq g(B)$. The ChI is

$$\int \mathbf{h} \circ g = C_g(\mathbf{h}) = \sum_{j=1}^{N} h_{\pi(j)}(g(A_{\pi(j)}) - g(A_{\pi(j-1)})), \tag{2.9}$$

for $A_{\pi(j)} = \{x_{\pi(1)}, \ldots, x_{\pi(j)}\}$, $g(A_{\pi(0)}) = 0$, and π (sort).

2.3.2 EMD on a Decomposed ChI

As already stated, a ChI on N variables has $N!$ underlying unique[7] LCSs (one for each possible sort, π). It is important to note that this sort is consistent across integrals. As such, distance can be measured and aggregated across the partitioned input space,

$$EMD_c^1(g_1, g_2) = \frac{1}{N!} \sum_{i=1}^{N!} EMD((g_1)_{\pi_i}, (g_2)_{\pi_i}), \tag{2.10}$$

where g_1, and g_2 are FMs, and $(g_k)_{\pi_i}$ is the ith sort of g (aka LCS). (2.10) is superscripted with a 1 to differentiate it from the data-derived variant detailed in (2.12). We elected to normalize the distance between all LCSs by the number of walks so the resulting value can be interpreted as an average distance between any two analogous walks. In summary, the idea is to produce an exhaustive and non-intersecting parti-

[6]For finite X, there is a third condition for continuous domains.

[7]We say unique with respect to the $N!$ sorts. LCSs are often duplicated (aka have the same weights) across sorts.

tioning of the input (and thus operator) space and to measure the average distance across all decomposed LCSs, with respect to already noted the EMD benefits.

Equation (2.10) is expressed in exhaustive—aka all $N!$ sorts—form. When both integrals have the same *underlying LOS structure*[8], (2.10) can be expressed as

$$EMD_c^1(g_1, g_2) = \sum_{i=1}^{M} \alpha_i EMD(\mathbf{w}_i^1, \mathbf{w}_i^2), \tag{2.11}$$

where M is the number of LOSs, $\mathbf{w}_i^k$ are the LOS weights for the kth FM, and α_i is the number of sorts in LOS i divided by $N!$. Otherwise, the EMD can be expressed as a combination of (2.10) and (2.11), where the $[0, 1]$ weights are $\frac{1}{N!}$ for individual LCS terms and the relative frequency terms outlined above for LOSs. The point is, the EMD between two ChIs can be expressed as an aggregation of the individual LCSs or its respective underlying LOSs.

2.3.3 EMD on Data-Derived Choquet Integrals

The ChI is often learned from data, e.g., [8, 7, 20, 21, 22, 23, 24], meaning we have additional information to aid in comparison. In [6], we showed a way to measure the frequency of observations per walk/context. A limitation of (2.10) is that it does not take data observably into account, which can lead to comparing walks that were not learned or have little data support. The following remedies this by weighting each walk,

$$EMD_c^2(g_1, g_2) = \sum_{i=1}^{|C|} \beta_i EMD((g_1)_{\pi_i}, (g_2)_{\pi_i}), \tag{2.12}$$

where $\beta_i = \frac{1}{2}\left(f([g^1]_{\pi_i}) + f([g^2]_{\pi_i})\right)$, C is the intersecting set of walks between the two datasets used to learn FMs g^1 and g^2, and f is the relative frequency of a walk with respect to our reduced scope. We choose to limit our region of interest to only

[8]Same structure here means that both integrals have the same number of underlying LOSs bound to the same sort sets.

walks that are observable in both sets of data. There are two reasons for this. First, as was shown in [5], variables in unobserved walks are imputed by whatever solving method was utilized, often defaulting to the initialization value or floor or ceiling imposed upon the variable by the monotonicity restraints. These values are often close enough for application uses, but we argue there is nothing truly to be learned about the ChI itself from these data-unsupported variables. Second, portions of ChIs are incomparable if one or both of them are not intended to operate on those specific sorts or partitions.

These frequency values act as a normalizing mechanism, where partitions of the data space that are more common in data—and thus, we argue, more important to the operator—are proportionately weighted. We treat these coefficients additively and rationalize that with the following three examples. If the f values for both $[g^1]_{\pi_i}$ and $[g^2]_{\pi_i}$ are near zero, the sort is very infrequent in either dataset, and therefore it matters less in the overall comparison of FMs. If both f values are high, then the sort in question is very common for both sets of data, meaning the difference between those LOSs is relatively more important. Finally, if one f value is high and the other is low, that sort is important in at least one of the data sets, meaning it should have some influence on the resultant distance. The inclusion of relative frequencies also takes care of the normalization, meaning we do not need to divide by $N!$ as in Eq. (2.10).

2.4 EMD ON A SINGLE CHI: STRUCTURE DISCOVERY

In this section, a second payoff of researching aggregation operator similarity is outlined. We show how measuring proximity between LCSs and LOSs can help us discover *underlying structure* in a single integral. This is useful as it relates to answering the XAI question, "what *logic* was learned?" As N gets larger, reporting $N!$ LCSs is intractable. Instead, it is more effective to summarize the $N!$ individual logics.

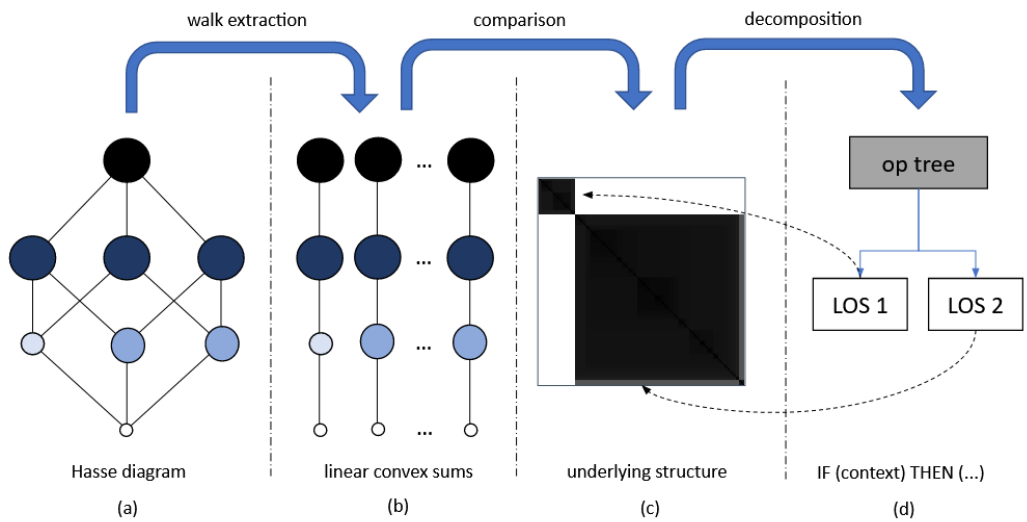


Figure 2.2: The proposed decomposition process starts with (a) a FM, which is (b) 'unrolled' into a set of LCSs (b). Next, these LCSs are compared against each other to (c) produce an iVAT image for structure discovery. Lastly, (d) clusters can be manually or automatically extracted into a smaller set of aggregation operators with additional context, e.g., LOSs.

Specifically, we are interested in a minimal, but still comprehensive, set of operators that capture the majority of aggregation information. See Fig. 2.2 for a flow diagram of the proposed idea and Algorithm 8 provides a formal description.

Algorithm 1: Discovery of underlying LOSs in a ChI

```
Data: g - input FM
for i to N! do
    for j to N! do
        D(i, j) = EMD(g_i,g_j)
D = IVAT(D)
if automatic clustering == True then
    return CLODD(D)
else
    return D
```

Herein, we use the *improved visual assessment of cluster tendency* (iVAT) algorithm [25] to highlight and recommend potential cluster structure in the LCSs. In the iVAT algorithm, similar data patterns generally appear as "dark rectangular blocks" along the Image (matrix) diagonal of the iVAT image. Once an iVAT image is ob-

tained, clusters can automatically be extracted using an algorithm like CLODD [26], if desired. We leave it to the reader to determine if a *human is in the loop* (HITL) or if an automatic procedure is needed.

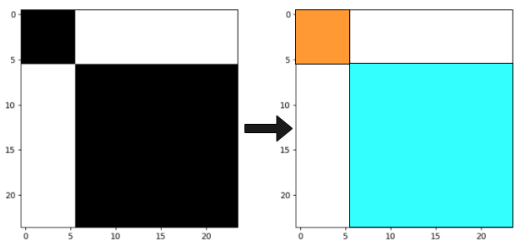
2.5 COLOR-CODED XAI VISUALIZATIONS

To further our goal of interpretability of the ChI, we devised a method to color the Hasse diagram in a way that helps the reader understand which subsets and walks up the diagram correspond to which iVAT/CLODD clusters. To this end, the walks which belong to each cluster are enumerated and all nodes that are a part of that walk receive a tally for contributing to that cluster — and by extension LOS. Each node in the diagram is then drawn as a pie chart, where the previous tallies can be used to show the proportions of how often each node is used by each LOS.

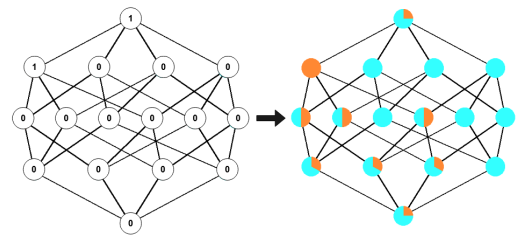
In summary, our iVAT and CLODD color-coded visualizations inform us about operator substructure and percentage of LCSs in each cluster. As a companion to the iVAT visualization, the color-coded Hasse diagram informs us about how each of the identified LCSs are associated with FM nodes and the interplay of the FM and ChI across all possible sorts.

2.6 EXPERIMENTS

In this section, we demonstrate the proposed techniques on two synthetic examples and a real-world example from the benchmark AID remote sensing dataset [27]. The synthetic cases are designed to test controlled scenarios where we know the answer and wish to demonstrate and validate the proposed approaches. The real-world experiment is a case study where we do not have the answer and no analytical solution exists.



(a) (left) iVAT image and (right) color coded CLODD results.



(b) (left) Hasse diagram with measure values and corresponding (right) color coded pie chart diagram.

Figure 2.3: iVAT, CLODD, and color coded Hasse diagram for the synthetic Experiment 1.

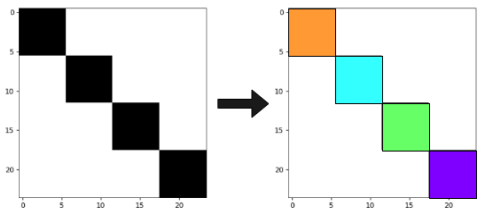
2.6.1 Synthetic Experiments

Experiments 1 and 2 are based on binary FMs, meaning $g(A) \in \{0, 1\}, \forall A \in 2^X$.[9] Though synthetic, many real-world data fusion problems can be best solved, or closely approximated via binary FMs/ChIs; see [29, 30, 31] for multi-sensor and multi-algorithm data fusion examples in hyperspectral image processing and remote sensing.

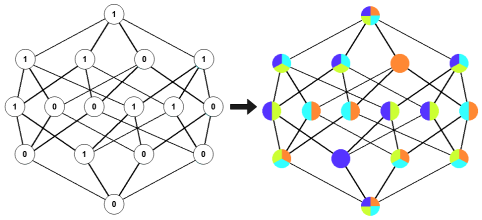
In Experiment 1, shown in Figure 2.3, we show the decomposition and discovery process on an FM such that $g(A) = 0, \forall A \in 2^X$, except $g(\{x_1, x_2, x_3\}) = 1$. Therefore, there are only two underlying LOSs, $\mathbf{w}^1 = (0, 0, 0, 1)^t$ and $\mathbf{w}^2 = (0, 0, 1, 0)^t$. In Experiment 1 we focus on setting a value at the top of the Hasse diagram to 1. In Experiment 2 (see Figure 2.4) we set a single value low in the Hasse to 1 (namely $g(\{x_2\}) = 1$). Due to the monotonicity constraints on the FM, the FM value for all subsets that contain source $\{x_2\}$ is therefore 1, and 0 otherwise. This results in 4 clearly separable LOSs; $\mathbf{w}^1 = (0, 0, 1, 0)^t$, $\mathbf{w}^2 = (0, 1, 0, 0)^t$, $\mathbf{w}^3 = (1, 0, 0, 0)^t$, and $\mathbf{w}^4 = (0, 0, 0, 1)^t$. It is interesting to note that this 'max-like' operator appears to be more complex than the above 'min-like' operator, despite their constructions being similar.

Figs. 2.3 and 2.4 show that despite there being 24 (4!) walks up the lattice for

[9] In [28] we proved that a binary ChI is a Sugeno integral.



(a) (left) iVAT image and (right) color coded CLODD results.



(b) (left) Hasse diagram with measure values and corresponding (right) color coded pie chart diagram.

Figure 2.4: iVAT, CLODD, and color coded Hasse diagram for the synthetic Experiment 2.

$N = 4$, iVAT and CLODD clearly highlight that there are two and four unique underlying LOSs respectively. In these figures, view (a) shows the iVAT result of the 24 underlying LCSs and its color-coded CLODD clustering result. The images in (b) are a new visualization technique proposed herein. The challenge is that the images shown in (a) help us understand underlying cluster structure. However, context is not preserved. In view (b), the left image is the Hasse diagram, where $g(\emptyset = 0)$ is on the bottom of the diagram, and $g(X) = 1$ set is on the top, and the monotonicity constraints are shown as edges. Each layer corresponds to a k-tuple, e.g., the layer above the empty set are the singletons, $\{g_1, g_2, g_3, g_4\}$.

The color-coded Hasse diagram corresponding to these FMs allow the reader to see at a glance which walks make up which LOS. For example, the top and bottom layers of the color-coded Hasse diagram in 2.3 reveal that approximately a quarter of the walks in this FM result in the orange LOS (0,0,1,0), while the rest result in the teal LOS (0,0,0,1). Additionally, due to the fact that the node corresponding to the singleton $\{g_4\}$ is solid teal, if that node is visited during a walk the resultant LOS is guaranteed to be the teal LOS.

2.6.2 Real-World Experiment

Unlike Section 2.6.1, the FMs in Experiment 3 are learned from real-world remote sensing data. In [32, 6, 8, 33, 34], we used the ChI to fuse a set of heterogeneous architecture deep convolutional neural networks (DCNNs) for object detection and land classification in remote sensing. The motivation of that work was that no single deep learning architecture nor trained model had been shown to be superior across all data sets and classification tasks. As such, the goal was to exploit the individual advantages of the networks and use the set to overcome their individual weaknesses to obtain a more accurate and reliable solution.[10]

While our previous publications focused on algorithm development and establishing the quantitative performance benefits of our ChI for fusing deep networks, we also explored different benchmark datasets, types, and numbers of architectures (see [32, 6, 8, 33, 34]). Herein, we restrict our analysis to the single case of fusing four DCNNs on the AID remote sensing dataset [27]. The goal here is to qualitatively explore the proposed XAI tools. The AID dataset contains $10,000$ images of 30 different aerial scene types. As outlined in [32, 6, 8, 33, 34], we trained both a "shared ChI", i.e., one FM shared across all 30 classes, and also one ChI per class. The benefit of the prior is that more data are available for training each class, whereas the latter has the benefit of learning class-specific fusion, generally at the cost of training data sparsity. We performed 5-fold cross validation with respect to neural learning and 2-fold CV for fusion. Rather than show all of the resulting 300 FMs, we summarize the findings next.

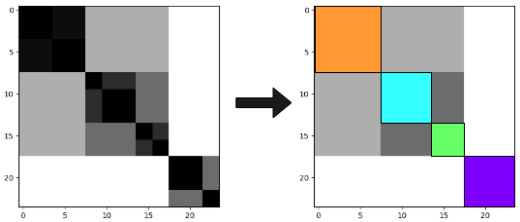
Overall, using the methods discussed herein, we discovered that nearly all of the learned ChIs were associated with a single operator, the minimum or a trimmed-minimum. This led us to further study our network outputs and we discovered that

[10] We showed in [8] that the ChI can be represented and trained as a neural net. Thus, the fusion of a set of heterogeneous nets is, therefore, merely a larger net with the benefit of explainability versus opaqueness [7].

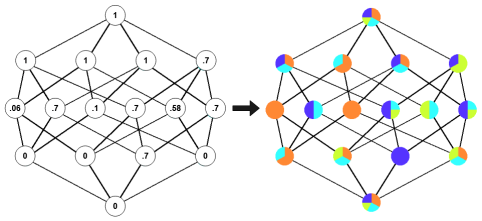
the deep networks are strong learners.[11] That is, the DCNNs were almost always certain—i.e., output values near 0 or 1—and they were frequently in agreement. The networks have not learned, nor where they informed to during learning, how to express uncertainty. They were trained to either output a 0 (not that class) or 1 (is a class). As such, it is logical to expect that the ChI learned to take a pessimistic stance, aka listen to the lowest confidence across the networks. In return, knowing this informs us that we should revisit the learning paradigm and, perhaps, take an ensemble of weak learners approach.

Fig. 2.5 shows the result of a typical experiment where the fusion of classifiers led to an increase in classification accuracy and algorithm robustness [32, 6, 8, 33, 34]. In general, much like iVAT on data for clustering, non-binary FMs result in less clearly separated and trivial LOS groupings. The distances shown in Figure 2.5(a) are normalized with respect to the min and max observed EMD distances. If the matrix was instead normalized with respect to the theoretical maximum possible EMD distance then we might be led to believe that there is instead a single cluster, a (trimmed) minimum. However, when normalized between min and max observed EMD value we see somewhere between two to five or perhaps seven clusters. The point is, it is up to the human visualizing these results, the set of CLODD parameters in the case of automatic cluster extraction, or some user-specified threshold governing approximation error between using the entire ChI and the iVAT/CLODD discovered number of LOSs. The focus of this paper is to introduce the tools and raise awareness of such questions. In future work, we will explore if there are answers to this question or if there are different truths for different contexts and applications.

[11] A scenario not the particularly ideal for fusion



(a) (left) iVAT image and (right) color-coded CLODD results.



(b) (left) Hasse diagram with measure values and corresponding (right) color-coded pie chart diagram

Figure 2.5: iVAT, CLODD, and color-coded Hasse diagram for the real-world Experiment 3.

2.7 CONCLUSION AND FUTURE WORK

In summary, the aim of this paper is to explore the role of the EMD as a measure to aid similarity analysis between and within FMs and ChI. We discovered that the bin-to-bin and ground matrix benefits of EMD improved our ability to recover semantic expectations of aggregation operator ranking. Furthermore, we showed how to compute distance between ChIs and considered how to take sampling frequency into account. We also presented a way to apply the measure within a single ChI, facilitating underlying operator discovery. These methods were illustrated via a combination of synthetic examples and a real-world dataset.

In future work, we will explore how to better express and study ground matrix selection and EMD in general relative to semantic ranking of a wider set of operators beyond the extreme cases of the max, median, mean, and min. We also plan to use the proposed measure on a set of learned ChIs across cross validations to understand if our remote sensing deep learner fusion is learning different logics. Other planned work involves looking at our most recent articles on visualization of the ChI [11, 10]. Namely, we have other ways to show data- and ChI-relevant information and information theoretic indices that likely would be beneficial to fold into the illustrations shown here. As already discussed, it is not clear, much like iVAT, CLODD, and clustering in general, what the underlying structure answer should be. We will further

explore this topic, most likely in the context of a specific goal or application to aid the specification process. Lastly, one of the major thrusts of the proposed article is XAI. We will look to combine the low-level methods proposed here into more useful high-level explanations for an end user.

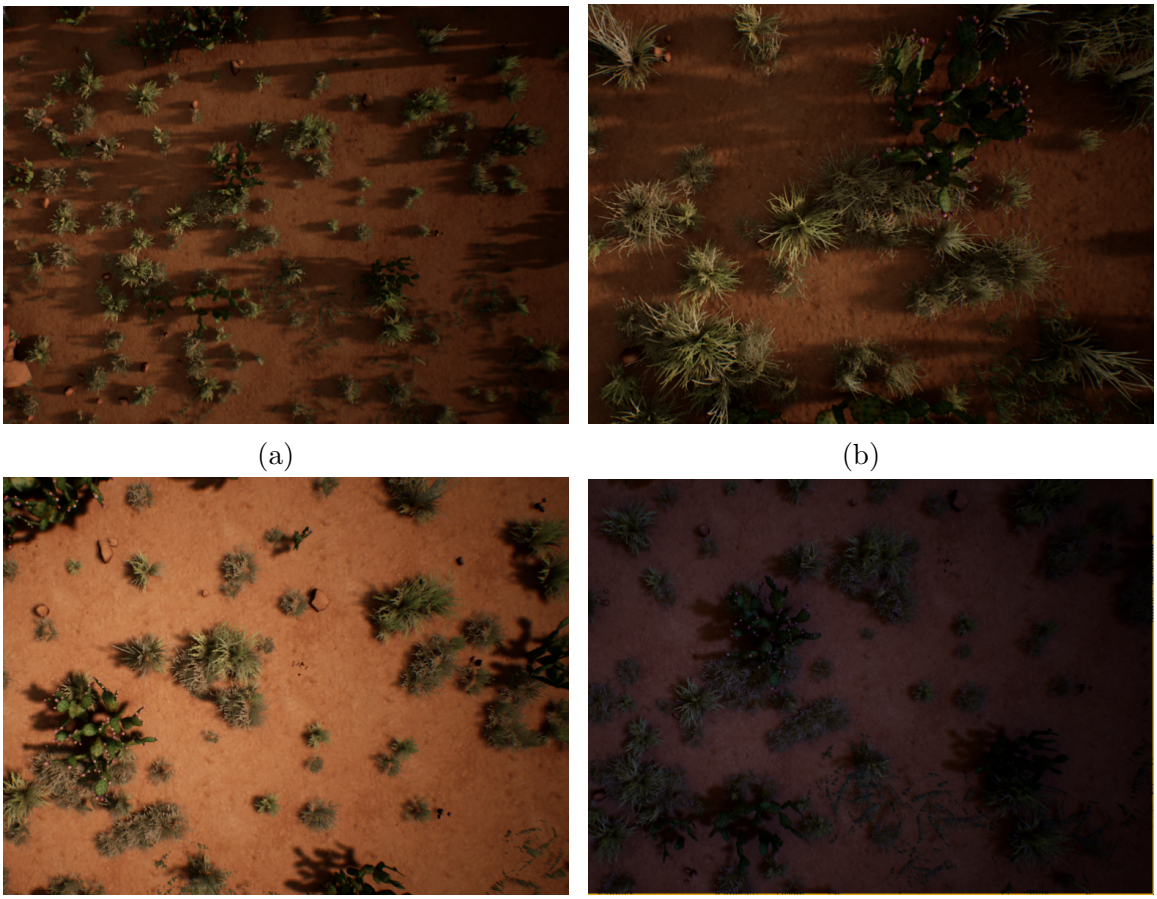
Chapter 3

METADATA ENABLED CONTEXTUAL SENSOR FUSION FOR UNMANNED AERIAL SYSTEM-BASED EXPLOSIVE HAZARD DETECTION

3.1 INTRODUCTION

The task of detecting and classifying explosive hazards (EH) from unmanned aerial systems (UAS) is a difficult one, in part due to the drastically varied environments and platform conditions one can expect to operate in/across. Detection in a hot desert at noon is a significantly different problem than detection in a frozen tundra at night. Detection from a UAS with a nadir sensing angle at 10 feet is different from a UAS with a sensor slant angle at 100 feet. Furthermore, the sensors used on UASs–e.g., RGB, IR, LiDAR, multi-spectral, etc.–all experience different sensor phenomenology depending on the environmental conditions and material properties of sensed objects. Figure 3.1 is an example that highlights environment and UAS (platform) variation. Herein, we refer to the above variations as *contexts*, as they are sources of information which we can use to enhance the performance of an underlying task like EH detection. In this paper, we propose an online and adaptive ensemble-based fusion scheme for EH detection that is driven by environment and platform metadata.

Before we delve into UAS-based EH detection (EHD), we briefly discuss related efforts. EHD technologies vary drastically. An early and well-known technique is the so-called "metal detector", which can be used to detect metal in the ground. However,



(a) (b) (c) (d)

Figure 3.1: Detection and localization algorithms are tasked with understanding objects in a variety of *contexts*, requiring robustness across factors like scale, color, illumination, and texture. Often, even a single location can look very different depending on platform altitude, look angle, time of day, etc. However, this information often goes unused in algorithms. The proposed contextual fusion scheme attempts to determine proper strategies based on metadata features which help inform context.

one limitation with this form of detection is that explosive threats that contain low amounts of metal may go undetected. Increasing the sensitivity of the device does not necessarily counteract this, as the number of false alarms would likely dramatically increase. To increase the robustness of detection, many different combinations of sensing methods have and are being explored, such as infrared (IR), ground penetrating radar (GPR), electromagnetic induction (EMI), and hyperspectral imaging (HSI), to name a few. The two predominant approaches to date for detecting ex-

plosives is vehicle-mounted detectors and hand-held detectors. While the latter is predominantly used in a downward looking fashion, the prior comes in a multitude of forms, e.g., forward looking[35], downward looking[36], and even side looking[37]. Herein, we focus on a UAS platform for EHD. Advantages of UAS, versus hand-held or ground vehicle deployment is it keeps humans at safer standoff distances and a UAS can in theory act like each of the above technologies. That is, it has the potential to search wide areas and dynamically interrogate regions of interest, likely through the use of a squad or swarm of UASs, with different sensors at different look angles. In this section, we limit our analysis to the use of a single UAS with multiple imaging and position sensors.

Adaptive fusion is not a new idea. For example, in Ref. [38], Frigui, Gader, et al. proposed a creative algorithm, context extraction for local fusion (CELF). We highlight and discuss this algorithm because we both use the Choquet integral (ChI). In particular, Frigui combined the fuzzy C-means (FCM) clustering algorithm and the ChI. They formulated a single joint optimization. Frigui partitions the input space based on the features to be fused. Our work differs as we initialize contexts based on otherwise unused metadata features such as altitude and temperature and learning independent operators from subsets of data. We also approximate the entire integral, aka all the underlying capacity variables, of which there are 2^N for N inputs. Frigui instead focuses on the "densities", i.e., the capacity defined on only the singletons, and an imputation strategy (the Sugeno λ fuzzy measure). We focus on learning the entire capacity because the tuples beyond the singletons capture interaction between sources. This is something we expect to occur and it can result in performance gain. Last, Frigui's fusion is driven by clustering. Herein, we exploit our recent transfer integral learning[5, 20] and data-driven eXplainable AI (XAI) methods[39, 40]. Specifically, the latter allow us to do things like identify what parts of a model (integral) were not approximated (sufficiently) from training data. The prior allows

us to transfer fusions, or parts of fusion, across integrals. The point is, our proposed method is more informed in the sense that when a fusion operator is being imputed we can better answer and resolve, "how similar is a new sample to our prior contexts" and "have we sufficiently approximated the parts of a fusion method that we would like to use." Each of these questions are important as we look to build, on the fly, the best response (aggregation operator). That is, something like we have seen before and have approximated from data. Last, we explore the use of the Unreal Engine [41] to create synthetic imagery. The graphical fidelity offered by these simulated environments proves as a useful surrogate for the otherwise difficult task of assembling large amounts of varied, UAS-captured data. Advantages include training real models from simulated data and rapid prototyping and experimenting with ideas that can later be transfered to real world experiments and solutions.

3.1.1 Machine Learning Models Derived from Limited Data Sample Sets

Data is king in modern machine learning. The performance of neural networks and other supervised learning models are intimately linked with the kind, quality, and diversity of training data provided. In a perfect world we could assume that good quality data can be obtained with enough time and patience, but this is rarely the case. It is in our interest to develop well performing classification models that have been exposed to only a limited amount of training data [20, 12]. This is especially relevant in the domain of UAS based vision as it can be difficult to obtain large amounts of appropriate aerial data.

The problem of limited training data is one which informs how the rest of this architecture is structured, and must be considered at every step. One known problem caused by small amounts of training data is *overfitting*. Overfitting occurs when a learning model memorises the solutions to training data but is unable to generalize to data it has not seen before. This is in part due to the training data lacking

adequate diversity, as the training data does not represent all possible variations of data that might be discovered. Our current method attempts to mitigate the overfitting problem by expecting models to only perform well on data that is similar to what has been seen before, and restricting their use to only when it is likely to perform well.

Another problem encountered due to limited training data is specifically tied to our fusion operator of choice, the ChI. As described later, the ChI partitions the input space based on the sorting of the inputs, where each unique sort results in a different method of combination. Because of this, we ideally would observe every possible sort in the training data so that an optimization algorithm is able to estimate all the values that are required. This is often not the case for a number of reasons. First, it can simply be difficult to encounter each sort through random chance. For N inputs there exist $N!$ possible sorts, meaning an adequately sized training set is required just to get one sample from each sort. A second reason the observed sorts are important is specifically tied to the domain of fusing *strong learners*. A strong learner is a learning model which usually produces only extreme (strong) values. For example, a strong classifier would only label detections as 0 (not the class) or 1 (is the class), but would rarely label something as 0.5 to denote uncertainty. Due to this, it is often the case that all models to be fused either agree on a class label of either 0 or 1. Thus, the only sort encountered is the default from when all values are the same. This heavily biases the ChI training procedure, as it is possible that nearly all observed data consists of only a few unique walks. If those unencountered sorts ever show up later in testing data, the operators will be poorly optimized to handle them. Our method attempts to mitigate this problem by transferring learned values from an integral that has observed the particular sort.

3.1.2 Ensemble of Neural Networks

A common technique to mitigate the reliance on a single black-box neural network is to train multiple networks which operate in parallel on the data. While this goes by different names in the community, we refer to it herein as an ensemble neural network. Each of the networks produces its own estimate of the target value, before each of the estimates are aggregated back into a single score. The precise method of aggregation depends on the architecture, though our method uses the ChI, in part due to it's capability of producing human-readable explanations of the fusion. This method creates an ensemble of networks with homogeneous architectures trained on varying subsets of data. A different popular technique in ensemble architectures is to vary the architecture of the individual networks (depth, number of parameters, etc.) but provide each network with complete data. The reader can refer to Refs. [8, 42, 34, 43] for our recent publications on ensembles of heterogeneous architecture neural networks for broad area scanning, land classification, and object detection in remote sensing. Figure 3.2 illustrates the flow of data and metadata in the ensemble architecture proposed herein. The pieces of this ensemble are described in greater detail in following sections.

3.2 METHODS

3.2.1 Fuzzy Measure and Fuzzy Integral

The fuzzy integral (FI) is a well studied tool in information fusion which defines a family of nonlinear operators. The integral is evaluated on a fuzzy measure (FM), $g : 2^X \to R^+$, which is a function that has two properties on finite X: (i) (boundary condition) $g(\emptyset) = 0$, and (ii) (monotonicity) if $A, B \subseteq X$, and $A \subseteq B$, then $g(A) \leq$

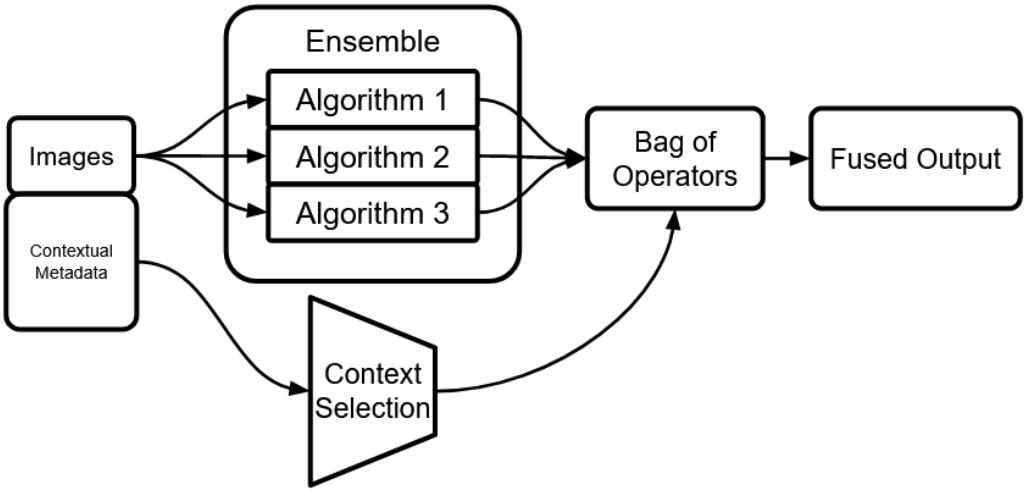


Figure 3.2: The general flow of images and metadata in our ensemble. Multiple algorithms are treated as sources of evidence to be fused together, while metadata such as altitude, temperature, and time of day inform the system how to construct the best possible aggregation operator.

$g(B)$. The Choquet integral (ChI) is a *type* of FI[3], given by

$$\int \mathbf{h} \circ g = C_g(\mathbf{h}) = \sum_{j=1}^{N} h_{\pi(j)}(g(A_{\pi(j)}) - g(A_{\pi(j-1)})), \tag{3.1}$$

where $\mathbf{h}$ is the integrand ($h(\{x_i\}) = h_i$ is the input from source i), $A_{\pi(j)} = \{x_{\pi(1)}, \ldots, x_{\pi(j)}\}$, $g(A_{\pi(0)}) = 0$, and π is a sort such that $h_{\pi(1)} \geq h_{\pi(2)} \geq ... \geq h_{\pi(N)}$. In our case, $h_{\pi(j)}$ is the jth largest return out of all algorithms, and $a \subseteq A$, $g(a)$ denotes the "worth" of a subset of algorithms. Thus, the ChI fuses evidence from each source based on the worth of a subset of sources.

It is relevant to note that the ChI is an operator which can be learned from data using various solvers. For example, in Ref. [44] we use quadratic programming (QP), in Ref. [8] we proposed constraint free full FM gradient descent optimization for supervised neural networks, and in Ref. [45] we proposed an evolutionary algorithm for efficient genetic operators on non-convex optimization surfaces. In this paper, we use the QP to learn a number of ChIs, trained on subsets of the data, which can be selected from based on what we believe the context to be.

A convenient way to visualize the ChI is in the form of its underlying Hasse diagram, where nodes in the diagram represent the g values of the power set of A in lexicographic order from bottom to top, left to right. For example, if $X = \{1, 2, 3\}$ the lexicographic ordering of the power set $P(X)$ is

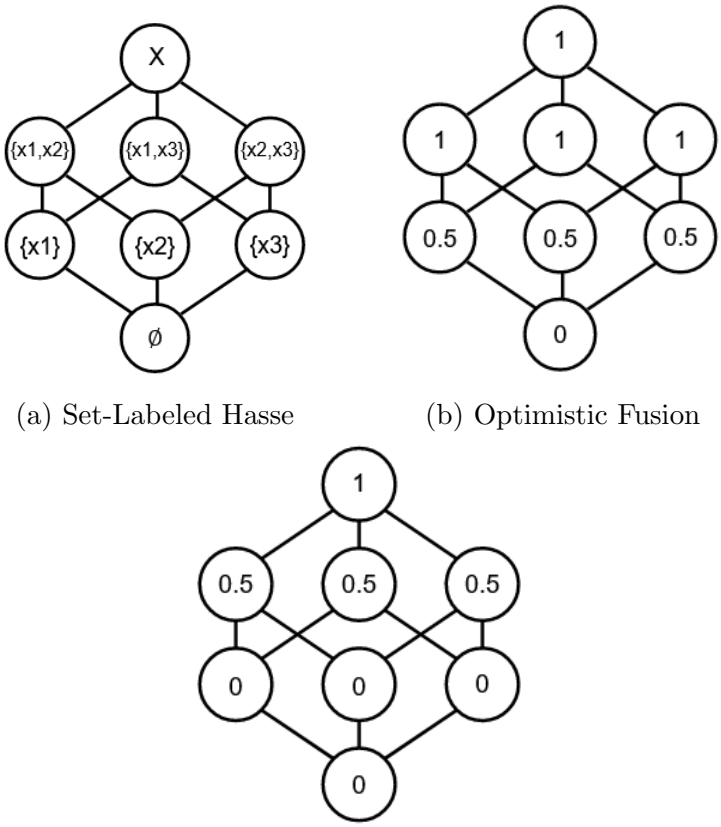
$$P(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

The edges in the diagram represent monotonicity constraints, meaning nodes in the upper layers are greater than or equal to the nodes connected below them. A *walk* up the Hasse diagram refers to a given sort on $\mathbf{h}$ and the resulting path taken from the bottom of the diagram to the top. Therefore, each walk defines a unique fusion operation the ChI is capable of. Figure 3.3 depicts example diagrams which are possible fusion strategies for three sources ($N = 3$).

3.2.2 Context Matters

The discrete ChI described above partitions the input space based on the sorting of inputs and each partition results in a different fusion operator. One way to think of these partitions is that they provide *context* as to what operator is most appropriate. An example interpretation of this for our current paper on UAS-based detection of EHs using multiple neural net algorithms is: "if Algorithm 1 has the greatest return, listen primarily to it. Otherwise, take the average of all the algorithms". We call this kind of context the *internal context*, as it is based solely on the data that is directly being fused. Specifically, Equation 3.2.1 informs us that each internal operator context is a linear convex sum (LCS) function, when $g(\emptyset) = 0$ and $g(X) = 1$.

However, there is more than just internal context in problems such as ours. Consider the task of EHD from a UAS. There are wildly different conditions in which the UAS might be flown, such as high altitudes, low altitudes, bright days, or dark nights. These are all normal operating conditions for such a system, yet the sensory



(a) Set-Labeled Hasse

(b) Optimistic Fusion

(c) Pessimistic Fusion

Figure 3.3: Hasse diagrams depicting different strategies of fusion for $N = 3$ inputs. An optimistic fusion like the one depicted in 3.3b averages the two largest input values. A pessimistic operator 3.3c averages the two smallest values. In algorithm fusion it is common to see pessimistic operators due to their redundancy as all algorithms must agree on a high value, i.e., unanimous consent.

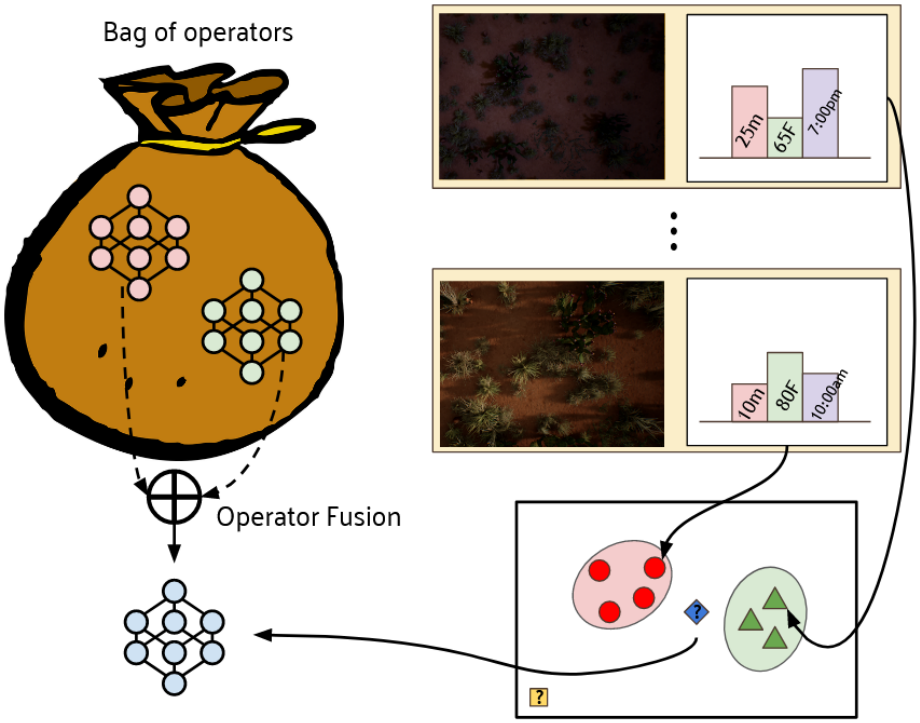


Figure 3.4: Illustration of the propose methodology. Metadata feature vectors are generated from training data and they are clustered to define initial contexts (red circles and green triangles). In this diagram we show an example image and the prototype per cluster. Next, a different ChI that combines a set of neural network classifiers is built per context (red and green Hasse diagrams). When a new sample (observation) belongs to a known context, the appropriate ChI is used. However, if a new sample, e.g., blue diamond, does not belong to a known context but it is similar to known contexts, then a new operator is built on the fly. In the event that a new sample is extremely different from anything that we have seen before, e.g., the yellow box outlier, then the system can decide to take no action or an operator can be built if the system is expected to always operate.

feedback in each of these conditions will be distinct. As a result, the algorithms that we use on this data must be robust to these variations. This system aims to better handle this kind of context, what we call the *external context*, of our fusion problem. Our method attempts to identify these unique external contexts by clustering the metadata obtained from the UAS (platform) and environment. Specifically, we use the GPS reported altitudes, recorded temperatures, and time of day as initial features to identify unique external contexts. Figure 3.4 illustrates our scheme of clustering metadata and using them to train unique fusion operators.

3.2.3 Metadata Feature Encoding

As the following metadata features will be used to inform the system of which context to associate the data with, it is important to consider their encoding. A common problem resulting from the use of disparate feature types is that one feature can dominate the space, e.g., have notably higher magnitudes. In our case, if we assume that the range of observed temperatures is on average larger than the range of observed altitudes, then the distance between temperatures will predominantly drive the distance measure. While there are other ways to handle this using techniques such as categorical encodings, a simple solution is to normalize the values (denoted as z below) on a scale of [0,1] based on minimum and maximum observed values,

$$z_{scaled} = \frac{z - z_{min}}{z_{max} - z_{min}}. \tag{3.2}$$

Special attention should also be paid to how the time of day is encoded, as it is a cyclical feature. Consider what happens if time of day was encoded as a scalar value in the range 0 to 23 hours (12am to 11pm). If we measure the distance from $t = 23$ to $t = 1$ ($11:00$pm to $1:00$am), the Euclidean distance is 22, though clearly those two time periods are only two hours apart. A simple yet clever way to avoid this problem is to split the time feature into two values given by

$$t_{sin} = \sin\left(\frac{2\pi t}{23}\right), t_{cos} = \cos\left(\frac{2\pi t}{23}\right). \tag{3.3}$$

When these two values are plotted as (x, y) pairs in the range $[0, 23]$, the result is a circle. This makes it a more appropriate encoding for use with Euclidean distance, as it now mimics distance on an actual clock, i.e., $t = 0$ and $t = 23$ are adjacent, while any two values offset by 12 hours maximize the distance. Note, in this paper we explore a few metadata. In future work we will investigate the inclusion of more

metadata and their respective pleasing semantic conditioning.

3.2.4 Determining Initial Contexts Through Clustering

As already discussed, our ensemble of neural networks is driven by context. To this end, we cluster the training metadata features into an initial set of contexts via the possibilistic c-means (PCM) algorithm[46]. The PCM is a mode seeking method that operates on a finite set of M samples $Z = \{\mathbf{z}_1, ..., \mathbf{z}_M\}$ relative to a specified number of c clusters. Unlike the k-means clustering algorithm, which is a crisp partitioning technique (i.e., every sample belongs to one, and one only, cluster), the PCM is a mode seeking algorithm. The PCM returns c clusters, which depending on the choice of underlying metric (e.g., Euclidean, Mahalanobis distance, GK metric, etc.) results in c prototypes, e.g., $C = \{\mathbf{c}_1, ..., \mathbf{c}_c\}$, and a partition matrix $[U]_{ik} = u_{ik}, i = 1, ..., C, k = 1, ..., M$, where u_{ik} is the typicality of sample $\mathbf{z}_k$ to cluster i. Unlike the k-means algorithm, the PCM allows samples to belong fully to multiple clusters and outliers can now be represented and detected. The PCM typicality degrees are especially useful at evaluation time, as it gives us a degree to which we believe a new data point belongs to a known context (cluster). As described later, we adapt our fusion strategy for new data (UAS observations) based on how similar it is to what has been seen before. In our implementation of the PCM, we use Euclidean distance and we initialize the cluster centers with the fuzzy c-means algorithm output because it helps us estimate PCM bandwidth parameters and it provides robustness over random initialization.

A problem ever present in all clustering algorithms is determining an optimal value for c. Herein, we use the fuzzy partition coefficient[47] (PC), an internal cluster validity index. The PC attempts to measure how well a set of data was partitioned

based on the membership values of each class given by

$$F_c(U) = \frac{tr(U * U^T)}{M}, \tag{3.4}$$

where U is the fuzzy partition matrix segregated into c classes, $*$ is matrix multiplications and $tr()$ is the trace, or sum of squared diagonals. A desired c can be selected from an index like the PC by looking for the maximum (or minimum) index value, or a trend (e.g., elbow) in the c plot. While the PC is used herein, it should be noted that there are more sophisticated internal (Xie and Beni index, Dunn, DBI, etc.) and external (Rand, etc.) cluster validity measures in the community, e.g., see Ref. [48]. If the reader desires to implement and use the methodologies contained herein, we recommend that a more robust cluster validity index be used.

3.2.5 Realtime Fusion

The above sections describe a set of offline computations on training data. The result is a set of contexts, neural classifiers (one per context), and subsequent aggregation operators (one ChI per context). This section outlines an online (aka runtime) selection mechanism to determine what contexts a new sample belongs to based on the typicality values provided by the PCM algorithm.

The selection process we developed breaks down into three distinct cases. The first case is when the data to be evaluated is highly typical of one and only one existing context, meaning we believe we have an appropriate fusion operator to use. The second case occurs when the data to be evaluated is highly atypical compared to all known contexts, meaning we are operating in an unknown context and will subsequently resort to using a default fusion operator. Herein, we explore the idea of a system taking an action, but a user could instead take no action because we are unable to predict how the system will respond. The third and final case occurs when the data to be evaluated belongs to more than one context. This is a case where

we will fuse multiple operators together to create a more appropriate adaptive fusion scheme.

The selection process above can be defined for data point z_i, where u_{ij} is the typicality of z_i in cluster j, and α and β are user defined upper and lower typicality thresholds respectively. The selection function is

$$\mathbf{g} = \begin{cases} g_k & u_{ik} \geq \alpha \text{ and } \forall j, j \neq k, u_{ij} < \alpha \\ g_{\text{default}} & \forall j, u_{ij} < \beta \\ \mathit{combine}(\mathbf{u}, g_1, ..., g_N) & \text{else,} \end{cases}$$

where condition one says pick a single FM/ChI when we are in a known context. Condition two is how we respond to a metadata outlier and condition three outlines the fusion of our fusions from metadata. The above scheme still leaves a few questions. First, what operator do we choose in case two? This is the case where the system is exposed to what appears to be a thus-far unseen context. We explore multiple methods, including a simple mean average of the network confidences and averaging the operators from all previously observed contexts. The simple mean is a natural place to start, as it credits equal worth to each of the sources to be fused. This is useful as it makes no assumptions about the worth of individual sources in unseen contexts, though this is also the method's weakness. If there is a clear pattern in the fusion strategies consistent across all contexts then the simple average will disregard this, throwing away information that could be useful as a default fusion strategy. The second method explored attempts to handle this problem by averaging the set of trained ChI operators. Here we define the average of a set of operators to mean calculating the average value on a per-node basis in the Hasse diagram. This produces a fusion scheme which retains any dominant fusion strategies common across contexts, while maintaining the monotonicity constraints required by the ChI. If there

is no obvious fusion scheme across all contexts (such as being generally optimistic or favoring a particular source), this averaged operator produces an operator that is in a way smoothed, and pulled closer to an operator that resembles the mean[1].

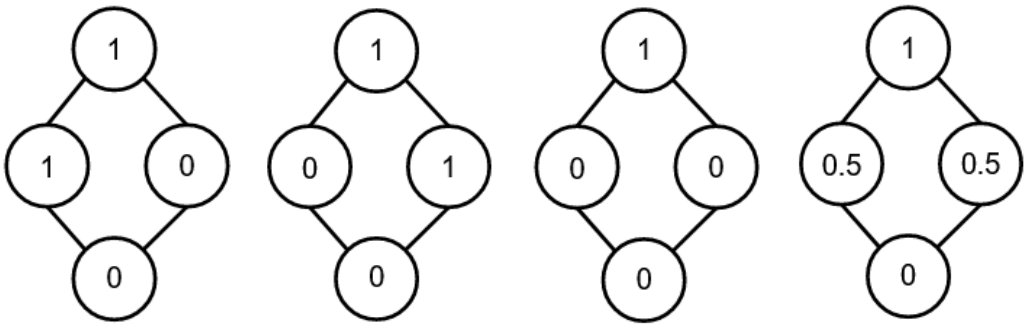
This solution inspires the combination method we use for case three (see Figure 3.5). In this case, our clustering is tight enough that a new data point can reasonably be considered to be in one of multiple contexts. To resolve the ambiguity, the operators in question are combined through a weighted average where the weight is determined by the relative strength of the typicality values. The $combine(\mathbf{u}, g_1, ..., g_N)$ function above is

$$\mathbf{g} = \sum_{i=1}^{N} \frac{u_{ik}}{T(\mathbf{u})} g_i, \tag{3.5}$$

where γg ($\gamma \in [0, 1]$) is defined as $\gamma g(A), \forall A \in 2^X$, $g = g_i + g_j$ is defined as $g(A) = g_i(A) + g_j(A), \forall A \in 2^X$, $T(\mathbf{u})$ is the sum total of typicalities for sample k, $T(\mathbf{u}) = \sum_{i=1}^{N} u_{ik}$. Other possible ways to aggregate FMs include the operators we outlined in Ref. [49] relative to evolutionary optimization, a simple t-norm like the minimum or product on each variable, or a set of t-norms and t-conorms outlined in Ref. [50] by Yager.

On a final note, we wish to comment that this is an initial study. That is, the above method only exploits metadata cluster membership values. This helps us build a new operator on-the-fly based on how similar the sample is to our past contexts. In future work we will also look at the internal context in each ChI (runs in the Hasse diagram) and combine it with our probabilistic estimate of how well that operator was supported[40, 39]. The idea being, there is no point relying on an operator that has not been sufficiently learned from data. Instead, a method like our ChI transfer learning[5] or similar should be engaged to derive an appropriate data informed internal operator. Last, these two disparate concepts need be combined.

[1] Assuming uniformly random FM g values



(a) First operator (b) Second operator (c) Minimum combine (d) Average combine

Figure 3.5: What is a reasonable scheme to combine FMs? 3.5a and 3.5b signify fusion schemes which listen entirely to a single source, a result that is likely to happen in our system if a given algorithm performs especially well in a certain context. If we combine based on a minimum operator or allow the quadratic solver to recompute on all data, the result is 3.5c. This operator is very pessimistic and will require both algorithms to agree on an answer, something that may be unlikely to happen. 3.5d is the result of a node-wise average, and maintains a degree of worth for individual algorithms.

3.3 PRELIMINARY EXPERIMENTS AND ANALYSIS

In this section we explore our proposed methods on a set of synthetic imagery meant to imitate changes in sensor phenomenology we might expect from use in different UAS environments. Imagery was generated using the Unreal Engine, as it allows automatic data-labeling and complete control of environment parameters. This provides us needed flexibility to explore ideas like adaptive fusion. That is, our methods are not bottle necked by real world factors like time and ultimately expense of collecting and labeling EH data. In our experiments we use the You Only Look Once version 5 (YOLOv5) network architecture [51] for EH object detection and localization, as it provides estimates of bounding boxes and confidences to fuse across, with a well documented implementation for easy training. We compare our method against a general model which has been exposed to all training data and is not a part of an ensemble. We examine what happens when the system is exposed to contexts that are not present in the training data, as well as good strategies for combining existing

operators when the metadata is ambiguous between multiple existing contexts.

It should be noted that we are intentionally not disclosing which environments, EH targets, and EH emplacement strategies we simulated. The targets and environments were determined in conjunction with our US Army Night Vision and Electronic Sensors Directorate (NVESD) collaborators. The targets are above ground objects (versus buried), they have moderate-to-low clutter (e.g., are often partially obscured by natural objects like a bush), we use a generic (aka similar to what you would find on the commercial market) RGB camera, and we believe that objects have enough pixels on target for detection. The goal was not to push the system to extreme breaking points, i.e., camera spectra being insufficient to detect an object or too few of pixels to even have an object that can be detected and discriminated from clutter. The point is to create a challenging and real world achievable problem that we can push to the point of failure and to compare the different avenues outlined herein. Furthermore, the UAS platform conditions were nadir (looking straight down) and a few arbitrary altitude variations were explored. In this section we do not consider all common scenarios, e.g., varying factors like aircraft speed and subsequent motion blur that would result. The point is, exact altitudes, environments, camera parameters, and etc. are just a surrogate herein to test the proposed algorithms. These experiments and environments are not real, but they are set up to mimic similar conditions to data that we have seen to date; meaning they are not overly simple and unrealistic. This paper is not a documentation of YOLOv5 for EHD on specific use cases. We would not report that information due to the real-world EH threat. In summary, what the reader can take away from the following experiments is the relative performance of the algorithms, their variations, and sensitivities.

3.3.1 Metadata Enabled Fusion versus Single Model

We start by evaluating the proposed algorithms on six sets of synthetic training data, where each dataset represents a different context that a UAS could experience during normal operation. As mentioned above, our goal is generality and diverse training data, not specific operational experiments. Specifically, the training runs consist of "high" and "low" altitude variants at solar noon (aka no shadows and ideal radiance conditions), afternoon (long shadows, darker), and night (very dark, most difficult) data. Below, these training datasets are referred to as "data set 1", "data set 2", and etc., and test datasets are simply referred to as "test 1" (Test1), "test 2" (Test2), etc. The test data sets contain otherwise unseen data (i.e., not resubstituion) with targets that are under heavy occlusion, shadows, and extreme angles so that they should be sufficiently difficult tests to evaluate our method. We feel like there is no loss of generality in our paper, as one can see performance in context, out of context, and with respect to outlier observations. As each training run has associated higher dimensional metadata (four dimensions herein) that we visualize using the dimensionality reduction techniques t-distributed stochastic neighbor embedding[52] (TSNE). Figure 3.6 shows the metadata from the six training runs, along with PCM assigned clusterings.

While the clusters are clearly separable in our experiments, when transitioning to the real world we do expect the data to be noisier. This can lead to an larger bandwidth parameter in the PCM algorithm, which will ultimately cause typicalities to increase across the board as the algorithm becomes more relaxed in what it considers a cluster. In short, these experiments are less likely produce the third case in the context selection procedure, as it is unlikely for a given point to have high typicalities across multiple clusters. It can also be noted that the reason the clusters manifest as rope-like in the projection is due to the time of day feature increasing linearly through time while the other features are pulled from a normal distribution. Furthermore, we

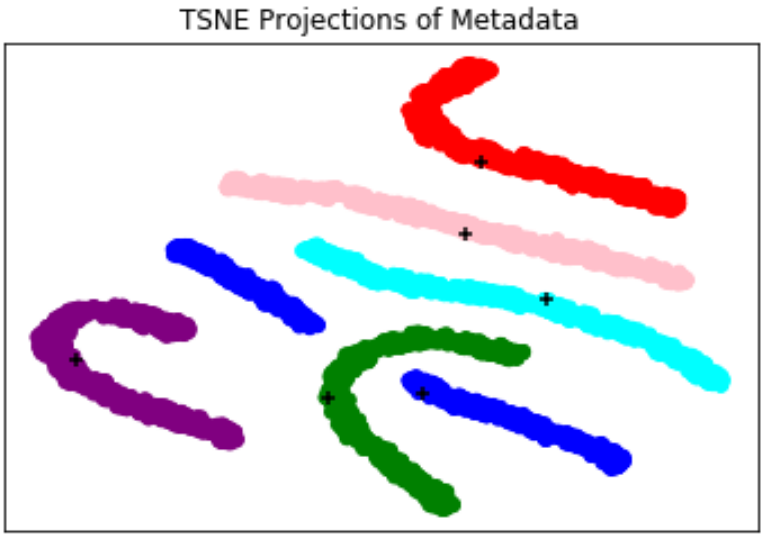


Figure 3.6: The metadata of our training sets reduced from four to two dimensions by TSNE. Color-coding is provided by PCM assigned clusters. Cluster centers are marked with a plus.

study this separable problem because it mimics the way that many real world collections occur. That is, data is often collected for a short number of consecutive days in a specific geographic area. We would not expect to have data from twenty four hours a day at all geographic locations. Last, it is our strong belief that if the proposed algorithms do not work for the scenarios explored herein, it is unlikely that they will work for the more challenging scenarios. And as stated above, an advantage of the Unreal Engine is we can generate a lot of data, of which all attributes are known. This is rarely the case in the real world as labeling can be sparse and error prone and documentation is never complete nor perfectly accurate (e.g., amount of cloud cover, temperature at each geospatial location, etc.).

Figure 3.7 shows the relative performance of our ensemble network (solid lines) compared to a single, out-of-the-box YOLOv5 model (dotted lines). Ideally, a good receiver operating characteristic (ROC) curve, where the x-axis is mistakes and y-axis is the positive detection rate, is the "zero FAR, one PD" (which is almost always achievable in real datasets). Most often, people look for "quick rises" (increase in

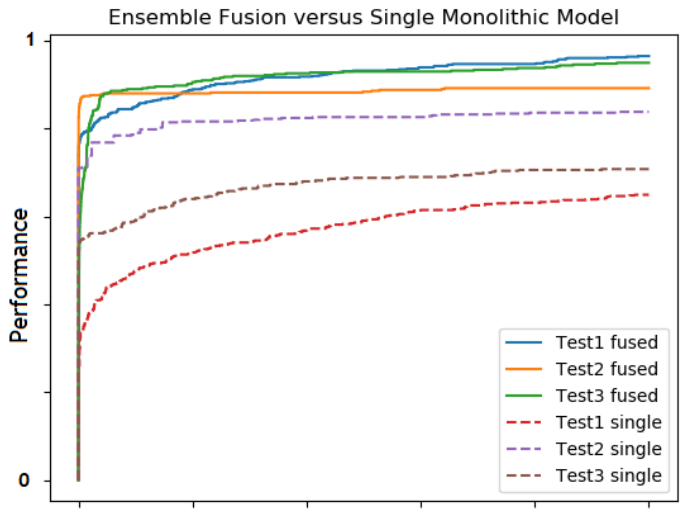


Figure 3.7: The proposed adaptive fusion scenario compared to a basic YOLOv5 architecture on three test scenarios. Test sets comprised of seen and unseen contexts.

PD with little to no mistakes) versus plateaus (no detections but more false alarms) or "linear climbs" (aka you have a 50% chance of calling something target or a false alarm). Thus, on the three test contexts evaluated, our ensemble method performs better than the standard model in all cases. This is similar to what we observed relative to fusing, with a fixed versus adaptive strategy, a set of heterogeneous neural networks for land classification and object detection in remote sensing[43, 34, 42, 39]. It should be noted that the total amount of data to train the ensemble is the same as the single model, though the single model was responsible for learning solutions across all of that data, while the ensemble was free to optimize a smaller subset of that data. While unproven, we believe this experimentally highlights the need to strike a balance between generalizeable models that perform well on all sorts of data and models that are experts in a more limited domain.

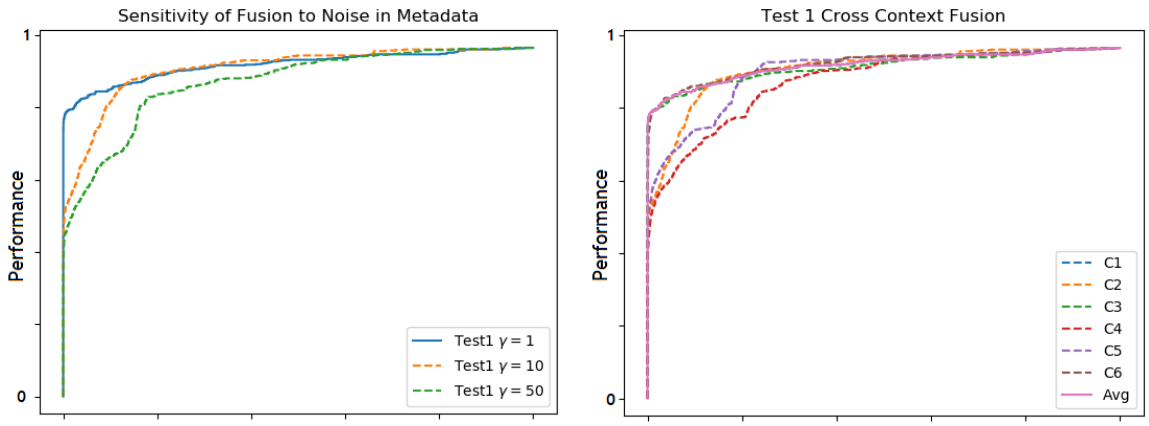
3.3.2 Sensitivity to Noise in Metadata

The above experiment is useful in the regard that it helps us understand operation in ideal scenarios. However, our method is reliant on additional data (metadata) provided by the UAS platform and/or environmental metadata. A benefit of using

simulation is that we have complete control over the fidelity of this data. That is, we can simulate noise and other errors, which are likely to appear when the algorithm is used in the real world. To better understand the robustness of our method to such errors, we construct the folowing two sensitivity experiments.

Figure 3.8a depicts the degradation of fusion performance as noise is introduced to the associated metadata. Again, we are not disclosing which metadata (altitude, time of day, etc.) lead to the biggest degradation due to the sensitive nature of EHD. Specifically, our metadata was generated based on normal distributions with varying levels of standard deviation. The gamma variables in figure 3.8a are scalar multipliers to the base standard deviation, resulting in more erratic (and less representative) metadata. Thus, $\gamma = 1$ is a single standard deviation (normal operating conditions), $\gamma = 10$ is 10 and $\gamma = 50$ is 50 times more noise, respectively. Semantically, the simulated types of errors lead to incorrect identification of current contexts, or relying on the generated default operator. It can be noted that the training clusters present in the synthetic experiments are nicely separable and spaced apart. It is unclear (future work) how these algorithms will work in the case of extremely close contexts. If the metadata is a good context identification scheme then contexts would be expected to be distinct and separate in space. However, if context, and or collected data are very close, e.g., model for 1pm and another model for 1:30pm, then we might expect that the trained classifiers and fusions should be similar. The point is, further analytical studies with performance characterization or experiments need to be performed in order to understand the impact of adaptive fusion for such scenarios. In summary, this experiment (Figure 3.8a) informs us that there is indeed an impact, but if our simulation is a close model to the operating conditions of a UAS for EHD, metadata noise is a concern but detection is not significantly impacted; we still get 40% detection with no error and do better than chance from that point on.

In Figure 3.8b, our second experiment, we consider what happens when the selec-



(a) Noise induced in metadata (b) All fusions compared on a single test run

Figure 3.8: Accurately capturing context data is important for the performance of our system. Excessive noise in the data can lead to incorrect operator selection as seen in 3.8a. The entire range of learned operators can also be evaluated on a single test run such as in 3.8b.

tion process consistently identifies an incorrect context. Whereas the last experiment showed us random controlled variations, this experiment tests extreme cases. This experiment is achieved by comparing the correct fusion operation to the rest of operators generated in the training process. These experiments results in six unique contexts being identified, thus, we can perform the fusion operation tailored to each of the six contexts. It should be no surprise that when the system is intentionally given incorrect contextual information, the performance of the fused algorithm is lowered, e.g., a morning model is used to detect EHs later in the day at a higher altitude. In general, we can see that the correct fusion operator (the generated weighted average) results in the best possible performance, though a few of the existing operators manage to perform equally well.

3.3.3 Explainability of Fusion

In this section we explore if there are any additional benefits of the proposed methodologies. One our underlying goals was to realize a trustworthy system versus opaque model. Our aim is to make a mathematical system that can be extended in future

work, e.g., factor in additional advanced EHD or physics knowledge. The system outlined herein is really "level 1 visual intelligence". That is, the system processes imagery and tries to learn robust low-level spatial and spectral features. Ideally, the system outlined herein is not the entire system, but one stage in the detection and understanding pipeline. The end goal being a "higher level environmental understanding" algorithm. To that end, one of the primary benefits of using the proposed ChI fusion strategy, as opposed to a black-box neural network is that the ChI can be opened up and examined after the fact to determine what fusion strategies were prevalent. That is, the ChI is a centralized and explicit model versus a distributed and implicit neural network. However, we remark that in Ref. [8] we put forth a way to encode and optimize a full ChI as a neural network, without loosing interoperability. In this section we take a look at what was learned in the previous experiments. We only report a subset of explanations. The reader can refer to our past works[39, 40, 8, 53] for our wider set of XAI fusion tools that generate statistical, graphical, local, and linguistic explenations.

In Murray et al. [39, 40], data-centric indices were proposed as a way of evaluating the kind and quality of data that was used to train the ChI. Of particular note is the walk-visitation calculation which describes what part of the Hasse diagram is well supported by data, i.e., how many (and which) internal contexts has a trained ChI observed and approximated. This can be used to identify "missing data", or perhaps more appropriately labeled "missing model variables." The trends reporeted herein are consistent across contexts, therefore we select and focus on the arbitrary Context 1. In this context, only 19% of the total possible walks received even a single piece of support. With this being a six source fusion, $6! = 720$ sorts are possible on the data, though only 137 of those were seen. Therefore, the integral is only approximately 20% approximated, which is not good. However, as we discuss in our fusion for remote sensing work[34, 42, 39, 40], most real world datasets do not have

sufficient volume nor diveristy. While many datasets claim to have both, our prior work showed that even for the higher volume datasets, their estimated fusion model values are frequently less than 20%. Meaning, our simulated scenario has arguably more diversity than we would encounter in practice. This makes sense to us, as the Unreal Engine lets us produce more data across different context.

Furthermore, of the walks that were taken, 65% of the time the data took the sort (1, 2, 3, 4, 6, 5), meaning that one walk almost completely dominates the operator. This is a common thing to encounter when training a ChI. That is, this is the default sort order. Usually that means that a bulk of the data is in agreement, is all saying the same thing. This is a typical behavior of strong learners, which we might expect from the YOLOv5 algorithm. Furthermore, this most prevalent walk corresponds to a **max** operator, meaning this particular operator tends to be optimistic. This is a difficult run in the Hasse to interpret. There is an fundamental entanglement that we cannot break apart. That is, this run is both the default sort order run and a valid case of what happens when algorithm 1 is more confident than algorithm 2, followed by 3, and so forth. As such, did we need the max to solve the latter or did it simply pick the max because it was an arbitrary selection when all algorithms say the same thing. Furthermore, the densities (values of the lowest level in the Hasse diagram) are $g(\{x_1\}) = 0.99, g(\{x_2\}) = 0.07, g(\{x_3\}) = 0.99, g(\{x_4\}) = 0.77, g(\{x_5\}) = 0.48, g(\{x_6\}) = 0$, showing that the fusion is often based entirely on the largest confidence value, as long as the largest confidence does not come from source two or six. Further analysis[39, 40, 8, 53] would be required to separate these variables to determine which are supported by data and we should trust.

This trend continues for most of the other contexts. Due to specific algorithms performing very well in each context (as the training procedure is therefore resubstitution), the fusion operators in those contexts weight those sources heavily. In future work we will look to sample and study validation data to minimize this effect. How-

ever, this is not the case for the default operator that was learned, as it was exposed to all data and it did not perceive a clear superior in the sources. As described in section 3.2.5, we explored multiple methods to generate default operators including an average aggregation and retraining on all data. While it would be difficult to display the full Hasse diagrams here (visually and with respect to page count), the operator that was retrained on all data resembles a **min** operator, while the average aggregation resembles a **mean**. This means that it is nearly impossible for the retrained operator to produce a high output, as all six algorithms would need to agree on a detection with a high confidence (something that rarely happens.) While it may seem a bit counter-intuitive to do all of this machine learning only to end up with something similar to what could be guessed at from the start (using a mean to aggregate sources), we believe that encouraging a less pessimistic model generalizes better to unseen data. The reader needs to keep a few things in mind. First, this is not conclusive and it is not a proof. It is merely an observed behavior of our experiments and experimental setup; i.e., simulated scenes, trained YOLOv5 classifiers, quadratic solver, etc. Thinking beyond our experiments, it is reasonable to expect that a high quality model trained for a specific context could prove to be *optimal*; versus the unachievable single model trained on all possible data or its ensemble approximation. Furthermore, we might expect that a mean like operator is an, on average, least worse strategy for outlier metadata scenarios. Last, when contexts truly overlap, something not explored yet, a mixture of models could prove to be a more robust approximation; similar to the performance gain we observed using fusion in Figure 3.7. Last, if we assume that each of these models and fusions are derived at least in part from data, then the models will always fundamentally have missing pieces. The point is, the above conclusion from our papers experiments are in no way conclusive. They are an experimental observation that we can use as intuition to set up the next and better approach.

3.4 CONCLUSION AND FUTURE WORK

This section proposes a metadata enabled adaptive fusion scheme for UAS-based EHD which attempts to discover the underlying contexts data was collected in to better inform the fusion operation. The offline determination of what constitutes a context is made by the possibilistic c-means (PCM) algorithm, a clustering algorithm which provides typicality values that describe to what degree a data point is typical of a given cluster. Once the training metadata is clustered, a set of context specific YOLOv5 location and detection classifiers are built, one per cluster. Finally, a Choquet integral (ChI) aggregation operator is trained for each context.

At evaluation time, the typicality values provided by the PCM allows the system to make an intelligent decision of whether or not an appropriate fusion scheme has already been trained. If the new data is sufficiently similar to an existing context then we are able to use the associated operator directly. If the new data is highly atypical from all previous contexts or similar to multiple contexts then the system uses a default strategy or it creates a new fusion scheme on the fly based on weighted interpolations of existing operators.

We evaluated the above methods on a set of synthetic imagery generated in the Unreal Engine, a process which allows us to circumvent the otherwise tedious process of obtaining large amounts of varied UAS data. Our results showed that there is benefit across the board in taking a metadata driven ensemble of our context dependent classifiers. Furthermore, we showed that while our system, as expected, is sensitive to metadata perturbation, the resultant ROC curve performance is still encouraging. Last, we showed additional sensitivity analysis experiments where we intentionally tried to destroy the algorithm. We note that this scenario is rare and might never be encountered in practice. However, the experiment reinforced our expected behavior of the system. That is, when out of context classifiers are used, performance is not ideal.

However, our metadata fused result remains resilient. In summary, these preliminary experiments are encouraging.

We have much future work to do, excluding EHD details that we omitted for sake of publication. For example, it would be interesting to see how well these simulator informed models transfer to real environments. Next, we developed a good amount of intuition through our setup and experiments. We will follow this publication up with in depth investigation for each component in a real UAS scenario, e.g., metadata, its similarity, context prediction, etc. for GPS, IMU, and environmental factors. Furthermore, we only achieved a first step of adaptive fusion herein. That is, we use clustering to inform the construction of an on the fly fusion operator. In future work we will advance this model to include the factors discussed above, like internal context and its degree of approximation in a ChI for a given context. That is, we want to advance this adaptive fusion mathematics and statistics to let us produce operators that are "as much like what we know before, but with respect to how well we know those solutions." This will likely lead us to processes like transferring solutions in and across models, and ultimately combining that with the metadata clustering typicalities. Last, our next goal is to move away, to some degree, from as heavy of experimentation and to rely on cases and analytical proofs, e.g., is it *optimal* to use a well trained model in a context versus an ensemble, what is the *optimal* operator for addressing outliers, etc. While preliminary experiments and the method are encouraging, there remains a great deal of future work.

Chapter 4

CONCLUSION

Working towards more explainable solutions like the ones presented in this thesis is an important step for the machine learning community. In this thesis, the goal was to show how the ChI can contribute to this ideal by engineering solutions with the intent to be more human understandable. Chapter 2 did this by considering a semantically satisfying similarity measure for LOSs and the ChI. By better measuring similarity in and across integrals we enable a decomposition procedure which can describe how many unique operators a given integral contains. Additionally, visualization techniques like the one presented in Chapter 2 provide tools to help us determine what was learned by an integral, and perhaps how to better engineer future solutions.

Chapter 3 furthered our goal by describing a fusion pipeline which is able to provide context-specific fusion operators to allow networks that were trained on limited data to act as an ensemble of more specifically trained experts. This system not only can tell the user what unique environments were encountered in training data, but is also capable of recognizing when new data is outside the bounds of what it was trained on, to alert the end user to either be wary of the an answer provided by this system, or to attempt a "best guess" fusion. Machine learning solutions that are aware of their own limits can do a great deal to better earn trust between an end user and the solution.

With the knowledge that engineering explainable solutions is possible, I hope that

it is made clear how important maintaining these principles are when dealing with the realm of big data machine learning. As we rely more on these solutions for tasks such as automatic target recognition[54], self-driving cars[55, 56], autonomous drones[57, 58], search engine algorithms[59], targeted advertisements and services[60], it is important we can not only strive for higher performance numbers and analytics, but must also remain informed as to how and why these algorithms are working.

Jaguar recognized this need for human-robot trust and tried implementing a user-facing solution for their autonomous driving cars by adding big googly eyes to their cars[61] that would watch and track pedestrians on the street to let them know what the car's autonomy system could see them. This ended up being deemed creepy, yet the need for that kind of system, where the human is made aware as to *how* a decision is being made, clearly exists. Similarly, in the domain of defense we will never (or should never) trust an algorithm so completely that we allow it to make decisions on who to shoot or where to send a missile without first being double checked by human engineers. The more contextual evidence these algorithms can provide, such as *why* that decision is being made and *how sure* the system is that this is appropriate, the easier time a human engineer will have interpreting the algorithm's advice to make the ultimate decision.

In regards to the above, one of the primary goals of building explainable solutions is to develop trust between the machine learning system and its user. In real-time systems especially we must trust that a machine is going to work correctly, and do what we expect it to do. This concept extends to human-human interaction, but despite the relative leap in complexity from a neural network to a human brain, it is much easier to trust and anticipate what actions a human may take compared to the black box that is a neural network. I believe that this divide between human and robot trust can be modeled by the concept of the uncanny valley. The uncanny valley is an idea borrowed from the field of aesthetics, where humans will show an increasing

affinity to a robot as it achieves more human-like qualities up until it reaches a point where it looks very close to human, but just different enough that our brains can tell the difference and find it almost frightening. I believe that there is a similar valley when it comes to trusting machine learning solutions. On the simple end, things like basic algebra, sorting algorithms, and optimization problems are trusted because it is obvious how they work, or at least they are provably working to some goal. However, as we approach complex machines which attempt to mimic human decision making, such as object detection[62], language translation[63], or game-playing[64] the mechanics of that system become too complex for a human to track. I believe that it is up to the ability of explainable solutions to dig us out of the trust valley by answering the same kinds of questions a human might be expected to answer when they make important decisions.

The problem of being able to fully trust machine learning solutions is one that will likely take many years before we are able to address it. In the meantime, allow me to humble myself by listing some concrete next steps to be considered following from the ideas presented in Chapters 2 and 3.

The EMD as a similarity measure was guided by an intuition. It was noticed that the few preconceptions we had about the order of common operators such as max, min, and median were not preserved when using an ℓ_p-norm. What we failed to discover was a concrete reason for why this was the case. A mathematically sound reasoning for our prescribed order hopefully exists, though it eluded us in this paper. Additionally, the motivating application for the similarity measure was its use in a decomposition process to determine unique operators within an integral. The basis for that decomposition procedure was defined, but there is work remaining to turn it into a useful tool. For example, ChI's have the problem of becoming intractable as the number of inputs grow, on order of $N!$. By decomposing an integral to its component parts, a significantly compressed version of the integral may able to be computed

while maintaining similar performance measures. Accomplishing this would require a clever data structure which is able to map a given sort down to a particular linear order statistic while maintaining a minimal set of parameters. Finally, it is desirable to come up with more succinct visualization techniques, as the proposed color-coded Hasse diagram is helpful if the user knows what they are looking at, but can still be difficult to interpret when the number of parameters grows.

The metadata enabled fusion system proposed in Chapter 3 is a prototype of a complex system, where improvements can be made at practically every step. One topic in particular to address is the effect low quantities of data has on training ChI's. By subdividing the training data as we did, a single integral is exposed to fewer data points, meaning it becomes increasingly unlikely that every possible variable and sort order will be observed in the training procedure. When this occurs, many values in the Hasse diagram must be imputed based on variables that were observed, despite there being little direct support. It is desirable that we develop a math-supported approach which can better impute these variables, perhaps by borrowing from similar integrals learned in nearby contexts. Furthermore, it is desirable that such a solution take into account the degree of data support versus simply seen or not seen. In addition, the clustering procedure outlined in this thesis was relatively simple, and should be expanded upon to best capture similarity in the metadata domain. Specifically, having semantically agreeable similarity measures such as the EMD are essential. The fact that each feature must be handled uniquely is unfortunate, but perhaps there is an ontological approach where properties of a given feature can inform the system as to how best represent that feature for clustering performance. This work also only considered an example metadata feature from the platform and environment domains. More metadata needs to be explored, and metrics relative to those metadata. We also explored a few strategies for dealing with operating in an unknown context. However, this still remains a big research question to tackle, as it will likely be a

big component of a real-world application like drones for explosive hazard detection. Last, the proposed methods exploited data in context. However, it is not always the case that we will have a large amount of collected data in a single context, e.g., early morning in a specific environment at altitude 20 and nadir look angle for a given subset of objects. Instead, it might be a good idea to identify similar contexts and utilize their training data or explore transfer learning.

In conclusion, machine learning is here to stay. As we apply these complex systems to more and more safety-critical problems, the ability to explain how and why a particular solution works is important for building trust between user and machine, as well as for determining when a system goes wrong. These are lofty goals, so we must focus on small but obtainable steps. The ChI is a data fusion tool for which we can construct explainable solutions through tools such as visualization, operator simplification, and context-aware fusion.

BIBLIOGRAPHY

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[2] J. Su, D. V. Vargas, and K. Sakurai. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, Oct. 2019. ISSN: 1941-0026. URL: http://dx.doi.org/10.1109/TEVC.2019.2890858.

[3] Y. Narukawa and T. Murofushi. Choquet integral and Sugeno integral as aggregation functions. In: *Information Fusion in Data Mining*. Ed. by V. Torra. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, 27–39. ISBN: 978-3-540-36519-8. URL: https://doi.org/10.1007/978-3-540-36519-8-3.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*. 2019. arXiv: 1910.10045 [cs.AI].

[5] B. Murray, M. A. Islam, A. Pinar, D. Anderson, G. Scott, T. Havens, F. Petry, and P. Elmore. Transfer Learning for the Choquet Integral. In: June 2019, 1–6.

[6] B. Murray, M. A. Islam, A. J. Pinar, T. C. Havens, D. T. Anderson, and G. Scott. Explainable AI for Understanding Decisions and Data-Driven Optimiza-

tion of the Choquet Integral. In: *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2018, 1–8.

[7] B. J. Murray, M. A. Islam, A. J. Pinar, D. T. Anderson, G. J. Scott, T. C. Havens, and J. M. Keller. Explainable AI for the Choquet Integral. *IEEE Transactions on Emerging Topics in Computational Intelligence*:1–10, 2020.

[8] M. Islam, D. T. Anderson, A. J. Pinar, T. C. Havens, G. Scott, and J. M. Keller. Enabling Explainable Fusion in Deep Learning With Fuzzy Integral Neural Networks. *IEEE Transactions on Fuzzy Systems*, 28(7):1291–1300, 2020.

[9] B. J. Murray, D. T. Anderson, T. C. Havens, T. Wilkin, and A. Wilbik. Information Fusion-2-Text: Explainable Aggregation via Linguistic Protoforms. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Ed. by M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager. Cham: Springer International Publishing, 2020, 114–127. ISBN: 978-3-030-50153-2.

[10] S. K. Kakula, A. Pinar, T. Havens, and D. Anderson. Visualization and Analysis Tools for Explainable Choquet Integral Regression. In: 2020.

[11] A. R. Buck, D. T. Anderson, J. M. Keller, T. Wilkin, and M. A. Islam. A Weighted Matrix Visualization for Fuzzy Measures and Integrals. In: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2020, 1–8.

[12] A. J. Pinar, T. C. Havens, M. A. Islam, and D. T. Anderson. Visualization and learning of the Choquet integral with limited training data. In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2017, 1–6.

[13] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, Jan. 2000.

[14] E. Levina and P. Bickel. The Earth Mover's distance is the Mallows distance: some insights from statistics. In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 2. 2001, 251–256 vol.2.

[15] S. K. Kakula, A. Pinar, T. Havens, and D. Anderson. Choquet Integral Ridge Regression. In: July 2020.

[16] R. Yossi, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, Nov. 2000.

[17] S. Shirdhonkar and D. W. Jacobs. Approximate earth movers distance in linear time. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] V. Torra, Y. Narukawa, M. Sugeno, and M. Carlson. Hellinger distance for fuzzy measures. *8th conference of the European Society for Fuzzy Logic and Technology*, Aug. 2013.

[19] H. Agahi. A generalized Hellinger distance for Choquet integral. *Fuzzy Sets and Systems*, 396:42–50, 2020. Generalized Measures and Integrals. ISSN: 0165-0114.

[20] S. K. Kakula, A. Pinar, M. A. Islam, D. Anderson, and T. Havens. Novel Regularization for Learning the Fuzzy Choquet Integral with Limited Training Data. *IEEE Transactions on Fuzzy Systems*:1–1, 2020.

[21] G. Beliakov. Construction of aggregation functions from data using linear programming. *Fuzzy Sets and Systems*, 160(1):65–75, 2009.

[22] M. Grabisch, H. T. Nguyen, and E. A. Walker. *Fundamentals of uncertainty calculi with applications to fuzzy inference*. Vol. 30. Springer Science & Business Media, 2013.

[23] M. A. Islam, D. Anderson, F. Petry, and P. Elmore. An Efficient Evolutionary Algorithm to Optimize the Choquet Integral. *International Journal of Intelligent Systems*, Sept. 2018.

[24] A. Mendez-Vazquez and P. Gader. Sparsity promotion models for the choquet integral. In: *Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on*. IEEE. 2007, 454–459.

[25] L. Wang, U. Nguyen, J. Bezdek, C. Leckie, and K. Ramamohanarao. iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment. In: vol. 6118. June 2010, 16–27. ISBN: 978-3-642-13656-6.

[26] T. C. Havens, J. C. Bezdek, J. M. Keller, and M. Popescu. Clustering in Ordered Dissimilarity Data. In: *International Journal of Intelligent Systems*. Vol. 24. 2009, 504–528.

[27] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, July 2017. ISSN: 0196-2892.

[28] D. T. Anderson, M. A. Islam, R. King, N. H. Younan, J. R. Fairley, S. Howington, F. Petry, P. Elmore, and A. Zare. Binary fuzzy measures and Choquet integration for multi-source fusion. In: *2017 International Conference on Military Technologies (ICMT)*. 2017, 676–681.

[29] X. Du, A. Zare, and D. T. Anderson. Multiple Instance Choquet Integral with Binary Fuzzy Measures for Remote Sensing Classifier Fusion with Imprecise Labels. In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2019, 1154–1162.

[30] X. Du, A. Zare, J. M. Keller, and D. T. Anderson. Multiple Instance Choquet integral for classifier fusion. In: *2016 IEEE Congress on Evolutionary Computation (CEC)*. 2016, 1054–1061.

[31] X. Du and A. Zare. Multiple Instance Choquet Integral Classifier Fusion and Regression for Remote Sensing Applications. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2741–2753, 2019.

[32] G. J. Scott, K. C. Hagan, R. A. Marcum, J. A. Hurt, D. T. Anderson, and C. H. Davis. Enhanced Fusion of Deep Neural Networks for Classification of Benchmark High-Resolution Image Data Sets. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1451–1455, Sept. 2018. ISSN: 1545-598X.

[33] J. A. Hurt, G. J. Scott, D. T. Anderson, and C. H. Davis. Benchmark Meta-Dataset of High-Resolution Remote Sensing Imagery for Training Robust Deep Learning Models in Machine-Assisted Visual Analytics. In: *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. 2018, 1–9.

[34] G. J. Scott, J. A. Hurt, R. A. Marcum, D. T. Anderson, and C. H. Davis. Aggregating Deep Convolutional Neural Network Scans of Broad-Area High-Resolution Remote Sensing Imagery. In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. 2018, 665–668.

[35] D. T. Anderson, K. E. Stone, J. M. Keller, and C. J. Spain. Combination of Anomaly Algorithms and Image Features for Explosive Hazard Detection in Forward Looking Infrared Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):313–323, 2012.

[36] P. D. Gader, M. Mystkowski, and Yunxin Zhao. Landmine detection with ground penetrating radar using hidden Markov models. *IEEE Transactions on Geoscience and Remote Sensing*, 39(6):1231–1244, 2001.

[37] J. Dowdy, B. Brockner, D. T. Anderson, K. Williams, R. H. Luke, and D. Sheen. Voxel-space radar signal processing for side attack explosive ballistic detection. In: *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXII*. Ed. by S. S. Bishop and J. C. Isaacs. Vol. 10182. International Society

for Optics and Photonics. SPIE, 2017, 421–435. URL: https://doi.org/10.1117/12.2262659.

[38] A. C. B. Abdallah, H. Frigui, and P. Gader. Adaptive Local Fusion With Fuzzy Integrals. *IEEE Transactions on Fuzzy Systems*, 20(5):849–864, 2012.

[39] B. Murray, M. A. Islam, A. Pinar, D. Anderson, G. Scott, T. Havens, and J. Keller. Explainable AI for the Choquet Integral. *IEEE Transactions on Emerging Topics in Computational Intelligence*, PP:1–10, July 2020.

[40] B. Murray, M. A. Islam, A. J. Pinar, T. C. Havens, D. T. Anderson, and G. Scott. Explainable AI for Understanding Decisions and Data-Driven Optimization of the Choquet Integral. In: *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2018, 1–8.

[41] Epic Games. *Unreal Engine*. Version 4.22.1. Apr. 25, 2019. URL: https://www.unrealengine.com.

[42] G. J. Scott, K. C. Hagan, R. A. Marcum, J. A. Hurt, D. T. Anderson, and C. H. Davis. Enhanced Fusion of Deep Neural Networks for Classification of Benchmark High-Resolution Image Data Sets. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1451–1455, 2018.

[43] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis. Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4):549–553, 2017.

[44] M. A. Islam, D. T. Anderson, A. J. Pinar, and T. C. Havens. Data-Driven Compression and Efficient Learning of the Choquet Integral. *IEEE Transactions on Fuzzy Systems*, 26(4):1908–1922, 2018.

[45] M. A. Islam, D. T. Anderson, F. Petry, and P. Elmore. An efficient evolutionary algorithm to optimize the Choquet integral. *International Journal of Intelligent Systems*, 34(3):366–385, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.22056. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22056.

[46] R. Krishnapuram and J. M. Keller. The possibilistic C-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3):385–393, 1996.

[47] T. J. Ross. Fuzzy C-Means Algorithm. In: *Fuzzy Logic with Engineering Applications*. Wiley, 1995, 358. ISBN: 978-0470743768.

[48] M. Moshtaghi, J. C. Bezdek, S. M. Erfani, C. Leckie, and J. Bailey. Online cluster validity indices for performance monitoring of streaming data clustering. *International Journal of Intelligent Systems*, 34(4):541–563, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.22064. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22064.

[49] M. A. Islam, D. Anderson, F. Petry, and P. Elmore. An Efficient Evolutionary Algorithm to Optimize the Choquet Integral. *International Journal of Intelligent Systems*, 34, Sept. 2018.

[50] R. R. Yager. A Measure Based Approach to the Fusion of Possibilistic and Probabilistic Uncertainty. *Fuzzy Optimization and Decision Making*, 10(2):91–113, June 2011. ISSN: 1568-4539. URL: https://doi.org/10.1007/s10700-011-9098-1.

[51] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, tkianai, yxNONG, A. Hogan, lorenzomammana, AlexWang1900, A. Chaurasia, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Durgesh, F. Ingham, Frederik, Guilhen, A. Colmagro, H. Ye, Jacobsolawetz, J. Poznanski, J.

Fang, J. Kim, K. Doan, and L. Yu. *ultralytics/yolov5: v4.0 - nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration*. Version v4.0. Jan. 2021. URL: https://doi.org/10.5281/zenodo.4418161.

[52] L. van der Maaten and G. Hinton. Viualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, Nov. 2008.

[53] B. J. Murray, D. T. Anderson, T. C. Havens, T. Wilkin, and A. Wilbik. Information Fusion-2-Text: Explainable Aggregation via Linguistic Protoforms. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Ed. by M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager. Cham: Springer International Publishing, 2020, 114–127. ISBN: 978-3-030-50153-2.

[54] E. Blasch, U. K. Majumder, T. Rovito, P. Zulch, and V. J. Velten. Automatic machine learning for target recognition. In: *Automatic Target Recognition XXIX*. Ed. by R. I. Hammoud and T. L. Overman. Vol. 10988. International Society for Optics and Photonics. SPIE, 2019, 119–130. URL: https://doi.org/10.1117/12.2519221.

[55] R. Kulkarni, S. Dhavalikar, and S. Bangar. Traffic Light Detection and Recognition for Self Driving Cars Using Deep Learning. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. 2018, 1–4.

[56] M. Martinez, C. Sitawarin, K. Finch, L. Meincke, A. Yablonski, and A. Kornhauser. *Beyond Grand Theft Auto V for Training, Testing and Enhancing Deep Learning in Self Driving Cars*. 2017. arXiv: 1712.01397 [cs.CV].

[57] A. Dhawale, X. Yang, and N. Michael. Reactive Collision Avoidance Using Real-Time Local Gaussian Mixture Model Maps. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, 3545–3550.

[58] W. Tabib, K. Goel, J. Yao, C. Boirum, and N. Michael. *Autonomous Cave Surveying with an Aerial Robot*. 2020. arXiv: 2003.13883 [cs.RO].

[59] D. Sullivan. *FAQ: All about the Google RankBrain algorithm*. June 2016. URL: https://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440.

[60] J.-A. Choi and K. Lim. Identifying machine learning techniques for classification of target advertising. *ICT Express*, 6(3):175–180, 2020. ISSN: 2405-9595. URL: https://www.sciencedirect.com/science/article/pii/S2405959520301090.

[61] R. S. Aouf. *Jaguar Land Rover's prototype driverless car makes eye contact with pedestrians*. Nov. 2018. URL: https://www.dezeen.com/2018/09/04/jaguar-land-rovers-prototype-driverless-car-makes-eye-contact-pedestrians-transport/.

[62] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. *Object Detection with Deep Learning: A Review*. 2019. arXiv: 1807.05511 [cs.CV].

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[64] N. Justesen, P. Bontrager, J. Togelius, and S. Risi. Deep Learning for Video Game Playing. *IEEE Transactions on Games*, 12(1):1–20, 2020.