EXPLORING THE IMPACT OF EXPLAINABLE HETEROGENEOUS FUSION PERFORMANCE FOR TARGET TRACKING

by

ASAD VAKIL

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND COMPUTER ENGINEERING

2025

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Jia Li, Ph.D., Chair
Wing-Yue Geoffrey Louie, Ph.D.
Li Li, Ph.D.
Robert Ewing, Ph.D.

To everyone in my life that has provided me with incredible support and understanding needed to finish my PhD, thank you.

ACKNOWLEDGMENTS

I would like to thank Oakland University for affording me the unimaginable opportunity to complete my study here, despite all the challenges which had frustrated all my efforts. I would especially like to express my deep and sincere gratitude to my research advisor, Professor Jia Li, whom for without her keen insight, patience, guidance, and understanding I most definitely would not have made it this far, especially with my thesis. I also remain grateful and thankful to the aid of Dr. Erik Blasch and Professor Robert Ewing and the opportunities and insights they have provided. And last but not least of course, this research would not be possible without the support of the AFOSR grants FA9550-18- 0287 and FA9550-21-1-0224.

Asad Vakil

PREFACE

Technology, like any tool, can be used for objectively good and objectively awful things. As the widespread usage of AI continues through the 21st century, the race to achieving great things with deep learning will only get *fiercer*, and the stock prices of NVIDIA and other key companies will only rise *higher* (especially after the turbulence caused by DeepSeek). As humanity continues to descend into what is probably not the best timeline but the timeline we ultimately deserve, a world ravaged by the effects of global warming, greed, apathy, and international conflict, should we somehow end up dealing with Skynet™ or whatever, I personally take consolation in the fact that I will hopefully not be alive for *any* of that, that no one ever bothers to read the preface of a 100+ page thesis, and that any accomplishments I make in this field that somehow inadvertently lead to the creation of such a thing will, at best, be *maybe* a subroutine or some minor footnote.

ABSTRACT

EXPLORING THE IMPACT OF EXPLAINABLE HETEROGENEOUS FUSION PERFORMANCE FOR TARGET TRACKING

by

ASAD VAKIL

Adviser: Professor Jia Li, Ph.D.

The prolific use of deep learning models in the information age has reached a point where it is almost ubiquitous with sensor fusion. The combination of available data and the ever-increasing processing power of available hardware has made many believe that data science is just a matter of throwing obscene amounts data into a deep learning algorithm. Rather than considering factors like the quality or available quantity of the data or trying to break down the complexity of the problem the model is designed to solve, the naïve use of brute force training is the chosen approach. This misconception, coupled with the inherent blackbox nature of deep learning algorithms, makes that the importance of explainability and transparency is integral to the success of machine learning, particularly with respect to using data we inherently lack an understanding of.

This is especially the case for Passive Radiofrequency (P-RF) data, which despite radar having existed since the 19th century, there is a lack of literature regarding the usage of available I/Q data for detection purposes. While P-RF data has many potential benefits for detection, with our research group's previous research and tentative patent requiring considerably less physical hardware to implement (using commercially available

software defined radio dipole antennas), the ability to utilize the modality is primarily dependent on the application of AI. In this thesis, we showcase research to the detection of vehicle targets via multimodal fusion in the ESCAPE dataset.

The research presented includes a comparison of over thirty multimodal fusion models for the shared objective of detection and differentiation of vehicle targets and examine the sensor data's impact on the model's performance. The integration of acoustic, electro optical, passive radiofrequency, and seismic data is implemented over three scenarios of data to determine the impact of different modalities using explainable AI methods. By comparing the local and global impact of each modality, as well as the F1 Score of the trained model, we can draw conclusions regarding how the modality was utilized with the benefit of the context of the training data and scenario. Computational costs of each model are considered in the context of FLoating Point Operations (FLOPs), and used to make determinations as to which modalities provided the most value to the fusion process.

TABLE OF CONTENTS

TABLE OF CONTENTS – Continued

LIST OF FIGURES

LIST OF FIGURES - Continued

LIST OF FIGURES - Continued

LIST OF ABBREVIATIONS

AI	“Artificial Intelligence”
DL	“Deep Learning”
EO	“Electro Optical”
P-RF	“Passive Radio Frequency”
FLOPs	“Floating Point Operations Per Second”
LLM	“Large Language Model”
ML	“Machine Learning”
XAI	“Explainable Artificial Intelligence”

CHAPTER ONE

INTRODUCTION

1.1 AI Hype

There is an intense global race to invest, develop, and acquire artificial intelligence (AI) technologies in the last decade, with seemingly no end to this trend in sight. The hype and investment for AI technology has only increased within the last few years, and there are few tech companies that have not announced new AI integration of some form or another. While this has been *fantastic* for the meme market and stock market (specifically companies like NIVIDA), it also begs the obvious question of whether or not the usage of AI in certain fields and applications is justified. We have witnessed companies such as BuzzFeed have their stock rise from announcements of AI content production, and also witnessed Samsung executives apologize for missing out on potential profits from the AI hype.

This is of course just a side effect of the modern news cycle in the 21st century, which is further compounded by distortion of certain influential individuals who own certain social media platforms pushing ridiculous claims about what AI is capable of despite narrowly avoiding being killed by one. These views are distorted and divorced from reality, but unlike the similar hype seen over the Metaverse, AI has actually useful applications that can contribute *meaningfully* to society right now, like denying health insurance claims and driving medical practices to bankruptcy. While AI is a disruptive technology, with the revolutionary potential to imitate the learning process, the fact of the

matter is that most neural networks (NN) require large amounts of training information, large investments of resources, and more importantly are inherently blackboxes.

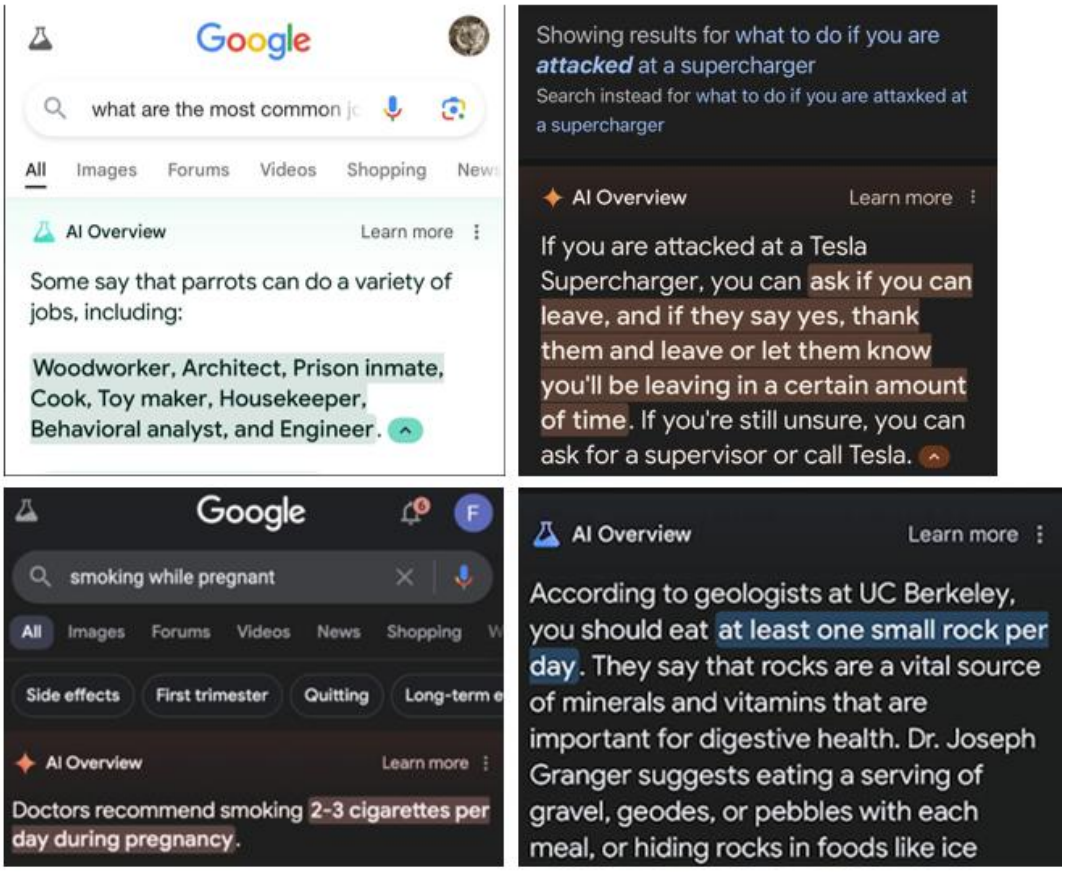

Figure 1. Examples of actual Search Engine AI Integration by a trillion-dollar company

This is not to say there haven't been any impressive steps (Figure 1), with plenty of people being more than impressed with Chat-GPT's mediocre misinformation spreading capabilities and blatant plagiarism, or ability of generative models for applications like deep fakes and art generation that are available for public use. Ignoring any moral quandaries relating to *where* the training data comes from and *what* this technology can be used for, it is painfully obvious looking at what models do exist, that civilization in 2025 is still *quite a bit aways* from the *Singularity*, let alone self-driving

cars or android servants/military grade terminators. For any believers of such Tesla hype, there's a disturbingly long list of outlandish proposed Tesla deliverables that haven't really gone anywhere, such as self-driving Teslas that can act as uber drivers while their owners sleep, that such individuals should reference.

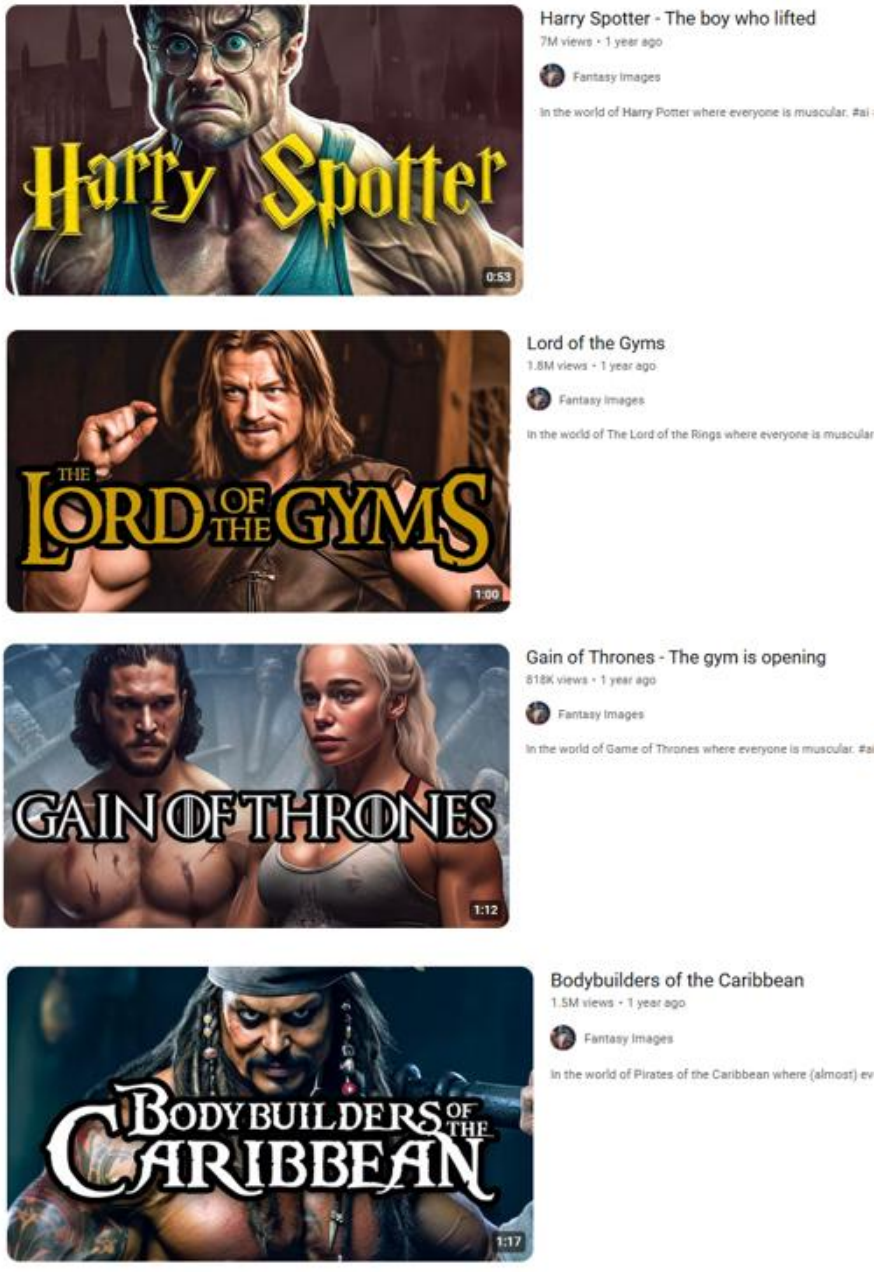

Figure 2. Examples of Mainstream use of Deepfakes for Entertainment

Besides corporations, there has even been a somewhat mainstream usage of AI technology, an almost “democratization” of the technology. From generative AI art to the use of deepfakes (Figure 2), the technology has been used for all manner of applications, such as entertainment or even in bizarre and downright *disturbingly* actual use by politicians [1] as seen below in (Figure 3). This is of course ignoring the impact of access to Large Language Models (LLMs) like ChatGPT have had on the younger generation, and the plagiarism it has enabled.

Figure 3. Political use of AI generated images.

While Deep Learning (DL) has benefited from the accessibility of data in the Information Age and the massive advancements in hardware capabilities, the next steps

of AI development are rather open-ended, with a number of important challenges [2] that data scientists will need to solve. In particular, developing a cognitive model-based approach to integrate pre-existing knowledge will be necessary to develop an AI capable of understanding reason and achieving the singularity.

Addressing legal and ethical questions regarding the use of the technology will also be important to discuss, as even ignoring the legal qualms of where data is being sourced from, there is the environmental impact to consider. While healthcare is often a field where AI usage has been hyped for to reduce the rising costs of one of the most important industries for human beings, how AI would be implemented and who is legally responsible for any mistakes that occur from AI implementation in healthcare are important questions to address.

Since the advent of the information age, for better or for worse, countless sources of information has been collected and stored. This fact is plainly obvious to anyone living in 2025, where our politicians have valiantly spent years trying to defend our youth from the national security threat that is Tiktok. Truly, there is *no greater problem* currently plaguing America than foreign social media influencing younger generations to do things like silly dances and not being okay with ethnic cleansing in the middle east. However, the *implications* of all this data existing and not being forgotten are something that people don't always consider. Philosophically speaking, life on this planet can be described as the long-term propagation of information.

The content of this information, however, is not *limited* to genes in the case of human beings and many other multicellular organisms. Creatures other than human

beings *do* teach their young, and in a sense pass down their own life experiences, knowledge, culture, history, and ideas. Whether it is the humble Japanese Macaque passing down a rudimentary form of spicing up their food [3] in the 60s, to lions teaching their young to hunt, these ideas needed to be passed on for future generations to learn from and improve. This is especially prevalent within the recording of human history, such as the transition from the vaguely mystical practices of alchemy to the actual scientific method, the development of algebra going from Persians in the Dark Ages using a heinous amount of Euclidean geometry to solve quadratic equations (example in Figure 4 below for any who are uninitiated in such witchcraft), to the more modern mathematical proofs that we take for granted.

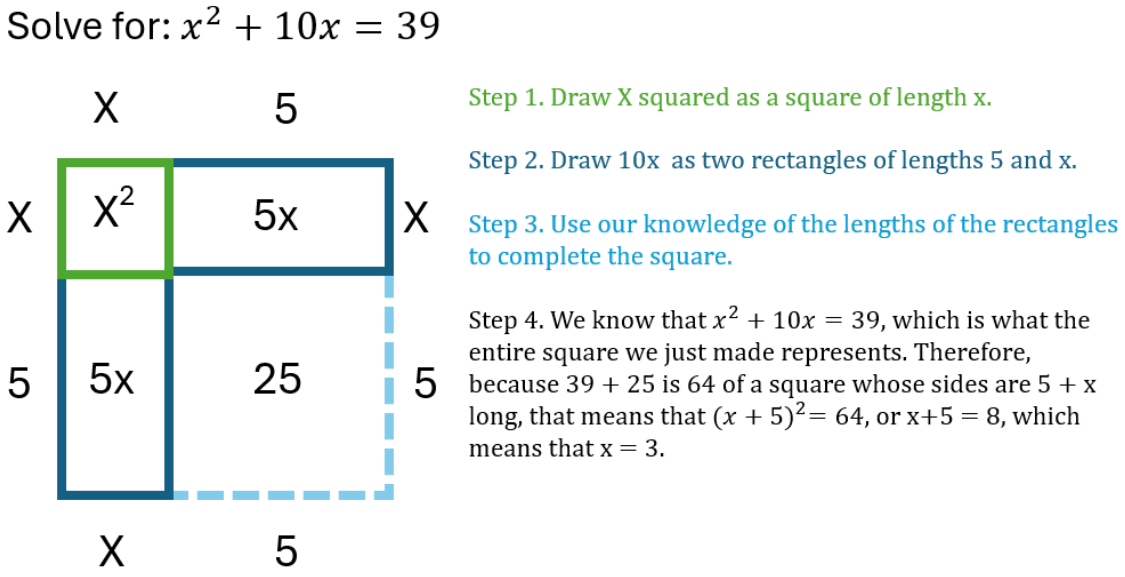

Figure 4. Dark Ages Math

Homo erectus and other forms of homo sapiens for example didn't go from using hand axes to the bow and arrow overnight, it happened in a process that spanned the better half of two million years, with even a few human lineages going extinct during that

period of time. Modern humans have always tried to keep some form of records of their lives, with differing levels of detail. Whether it is through oration, pictures drawn on caves, symbols, tablets, books. But in general, not all of this information is inherited by later generations, only a small percentage of the whole was selected and processed, then passed on. In a manner not unlike genes really, that's what history is. But in the Information Age, trivial information is accumulating every second, preserved in all its triteness. Never fading, always accessible. Rumors about petty issues, misinformation, slander, all of this junk data, preserved in an unfiltered state, growing at an alarming rate. On its own, this junk information is garbage, pointless, an absolute waste of time to even read, let alone try and monetize.

However, with the right contextual information (like age, ethnicity, where you vaguely live, etc.), an AI algorithm can take this garbage information and find general trends can be used to create a big picture, and *that* has **value**. These trends are used for everything, from awful targeted advertisements people receive all the time, to helping form digital echo chambers on social media [4]. The fusion of insurmountably large amounts of different sources of data is objectively one of the better uses of AI technology. Sure, movie producers want to use AI algorithms to generate films, so they don't have to worry about silly things like paying animators or actors or whatever, but between the costs for such models and their limited capabilities, it is infeasibly ridiculous with current technology. Data fusion, however, is an application for which AI has proven that it excels in.

The ability to estimate a general trend between multiple sources of data, to find potential correlations in data that are relevant, is invaluable to countless fields. It is also *especially* valuable for the benefit of reducing uncertainty. The problem of what regulations should exist, whether this large amount of trivial information being preserved for all to see and the resulting loss of nuance and objective truths is a good thing, how AI should be used etc. are absolutely not covered within the scope of this thesis, but what is covered in this thesis is the importance of explainability and transparency within AI, particularly for multimodal sensor fusion.

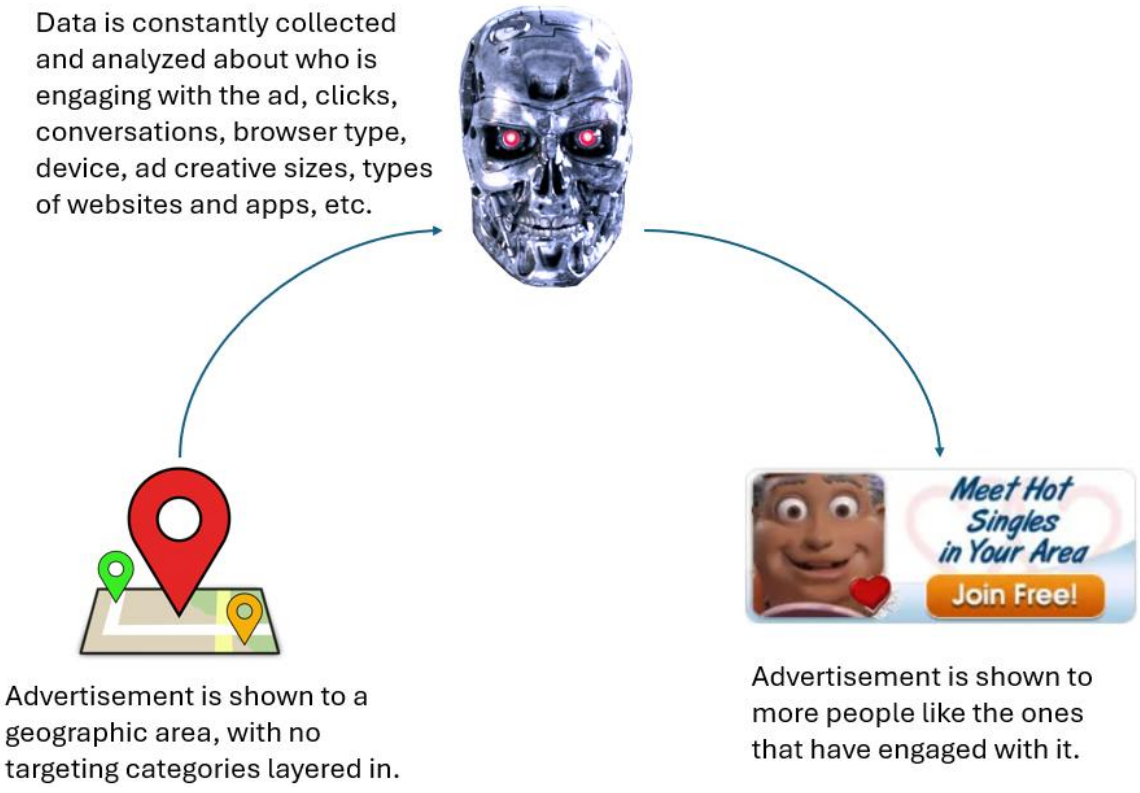

Figure 5. Data Fusion and targeted advertisements.

Chapter 1 will cover deep learning, the need for explainability, efficiency metrics, and overall contribution to AI research. Chapter 2 will cover a literature review on

relevant topics to the thesis, such as EO/P-RF heterogenous sensor fusion, XAI, efficiency of machine learning. Chapter 3 will cover the ESCAPE dataset, the sensors used and relevant data preprocessing, and scenarios as relevant to the experimentation shown in this research. Chapter 4 will cover the culmination of the P-RF modality research and making use of the features it provides. Chapter 5 will cover model metrics for performance, fusion, and efficiency, as well as discussion over preprocessing methods that were removed from the final testing. Chapter 6 will detail the experimentation for Phases I through III of the experimentation in this thesis. Chapter 7 will cover the conclusion, caveats, and future work, based on the results.

1.2 Deep Learning and Information Fusion

The integration of raw data and information fusion is a prevalent and necessary step in the information age. There exists a vast range of classical probabilistic data fusion techniques, but the inherent versatility of deep learning provides many potential benefits to many fields. When using nonlinear data or when existing literature does not necessarily provide an approach, the potential that deep learning has with methods such as unsupervised or end-to-end learning is extremely appealing. Whether it's for attempting to understand how whales communicate or finding out the secret names of Marmosets, AI has a lot of possibilities for pattern detection. While the application of machine learning is far from an infallible problem solver, there are a number of situations in which the method is a viable solution.

As is common knowledge, AI includes a wide range of categories and techniques that include machine learning and more specifically deep learning methods under its umbrella (Figure 6). More traditional methods generally belong under the umbrella of AI or machine learning, such as canonical correlation analysis, decision trees, etc. Deep Learning generally includes neural network based models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), Large Language Models (LLMs) such as Chat-GPT, and Multilayer Perceptrons (MLP). The primary focus on this paper will be towards Deep Learning models and will use the terms fairly interchangeably for the purposes of discussing them.

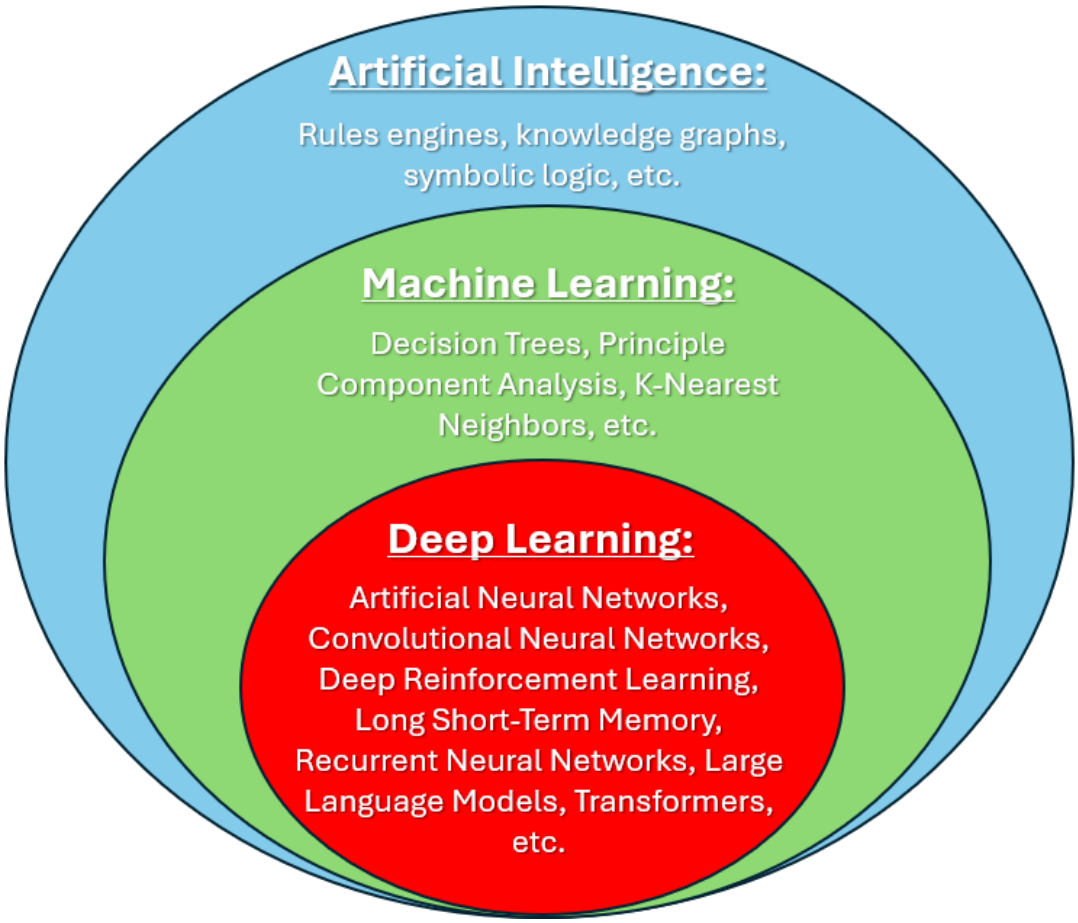

Figure 6. DL vs ML vs AI (redrawn according to analyticsvidhya)

Deep learning, under the right circumstances, has an almost endless amount of potential. While there are a number of applications that seem far and out of reach of what is available currently, there are many mundane and everyday examples of what deep learning can accomplish. In everyday utility and entertainment applications, passive AI can be seen, learning, and making decisions based on the input the user provides. From those awful targeted advertisements on your phone, your suggested videos on websites like YouTube, Netflix, or Spotify, to the ability to crash a Tesla via the "Autopilot" feature, the AI overview feature from search results using websites like Google and Bing, and sometimes correctly diagnosing medical conditions, it's frighteningly clear just how

versatile and retentive the results can be. The capability to utilize particular algorithms to make computer systems “learn” by using given data without specific programming is what defines machine learning as a whole. Creating computer systems that can see, know, learn, and predict the world like a human being and having the end goal of eventually reaching the singularity.

Machine learning is broadly divided based on its attributes for learning, such as unsupervised, supervised, semi-supervised learning, and reinforcement learning. The field however is always evolving and many broad paradigms exist. Other common paradigms include Self-Supervised Learning, Meta-Learning, Online Learning, Batch Learning, Curriculum Learning, Rule-Based Learning, Neuro-Symbiotic AI, Neuromorphic Engineering, etc., but these are outside of the scope of this paper and therefore not covered in further detail.

If the information is completely labeled out, with every input having a labeled output, the algorithm seeks to construct a model to map the input to the output, this is referred to as supervised learning. Typically, supervised learning deals with applications such as classification or regression, and includes algorithms such as Decision Trees (DT), K-Nearest Neighbors (KNN), Linear and Logistic Regression, Linear Discriminant Analysis, Naïve Bayes (NB), Similarity Learning, Support Vector Machines (SVM), and standard Artificial Neural Networks (ANN). This is especially useful in image and speech recommendation, fraud detection, natural language processing, and many other applications. In information fusion, it is also important to consider factors such as how heterogenous the data sources are, if there’s any redundancy in the data, the impact of

outliers, noise in the output values, dimensionality of the input space, and how complex the desired function is with respect to the available training data.

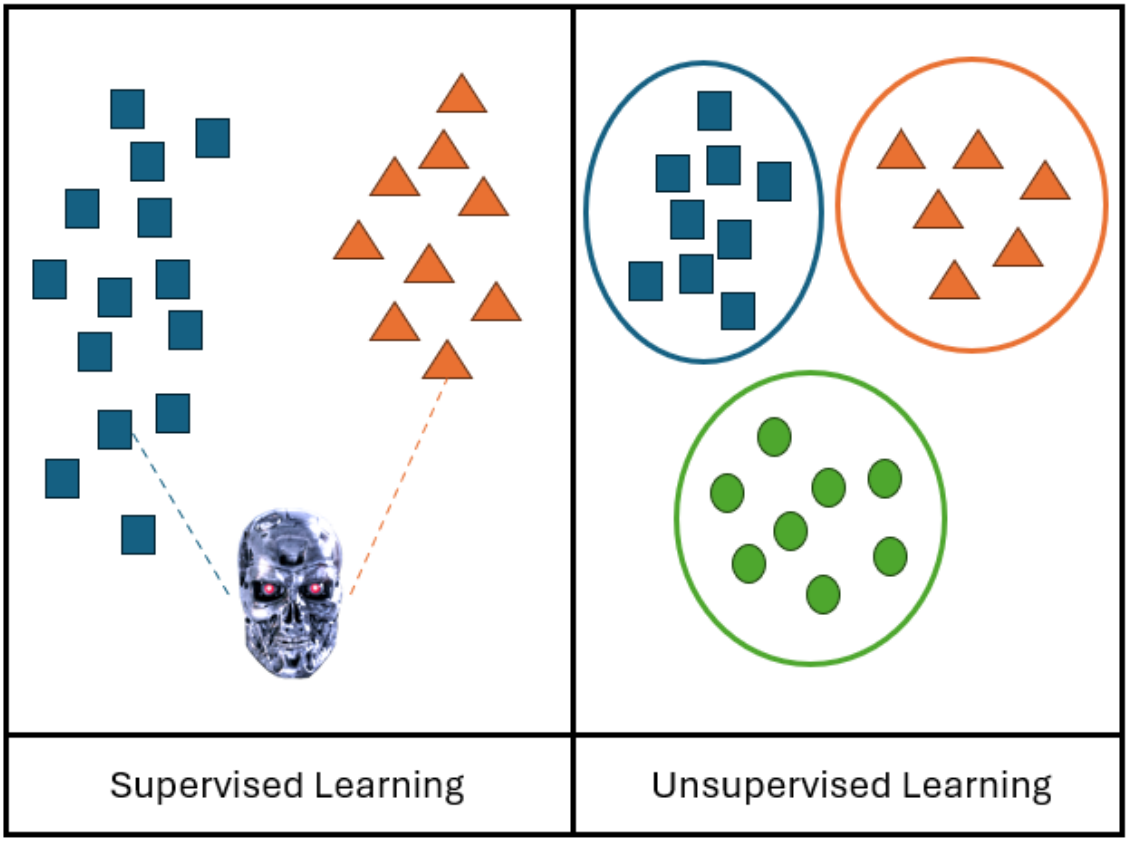

Figure 7. Supervised vs Unsupervised Learning (redrawn according to anubrain)

For unsupervised learning, algorithms are forced to extract features and patterns themselves, seeking similarity or distance of data inputs. While it may seem disadvantageous to force feed information to an algorithm and just expect useful results, unsupervised learning's ability to identify patterns and relationships in the data as well as cluster it based on similarities has a large number of potential uses. Commonly, these approaches will include clustering (Hierarchical Clustering, Mixture Models, DBSCAN, OPTICS) and anomaly detection (like Local Outlier Factor or Isolation Forest), or

otherwise aim to learn latent variable models (like Expectation-Maximization, Method of Moments, Blind Signal Separation). An example of unsupervised learning would be K-means clustering, Principal Component Analysis (PCA), Independent Component Analysis, Non-Negative Matrix Factorization, Singular Value Decomposition, and models such as Hopfield Networks, Boltzmann Machines, Sigmoid Belief Net, Deep Belief Network, Helmholtz Machine, Variational Autoencoder, etc. Typically, these applications seek to reduce the complexity of a problem or aid in matters such as feature selection.

With regards to what lies between the two, semi-supervised learning is the halfway point between the two approaches, starting off with a series of labeled data points as well as some data points that are not known. The end goal of the semi-supervised model is to classify some of the unlabeled data using the labeled information set. These models seek to accomplish that which supervised models do, to predict a target value for a specific input data set. Applications for semi-supervised models include fields such as speech analysis or web content classification.

Tasks categorized as discriminative tend towards (but not always) supervised learning tasks, while tasks categorized as generative tend towards (but not always) unsupervised learning. Finding associations with their internalized heuristics, these clustering methods are referred to as unsupervised learning. The separation between the two is fairly hazy, as for example, object recognition tends towards supervised learning, but unsupervised learning can also cluster objects into groups. Some tasks will often employ both, while others swing from one or another, as image recognition has done in

the last two decades, originally having been more supervised, but now often uses hybrid approaches.

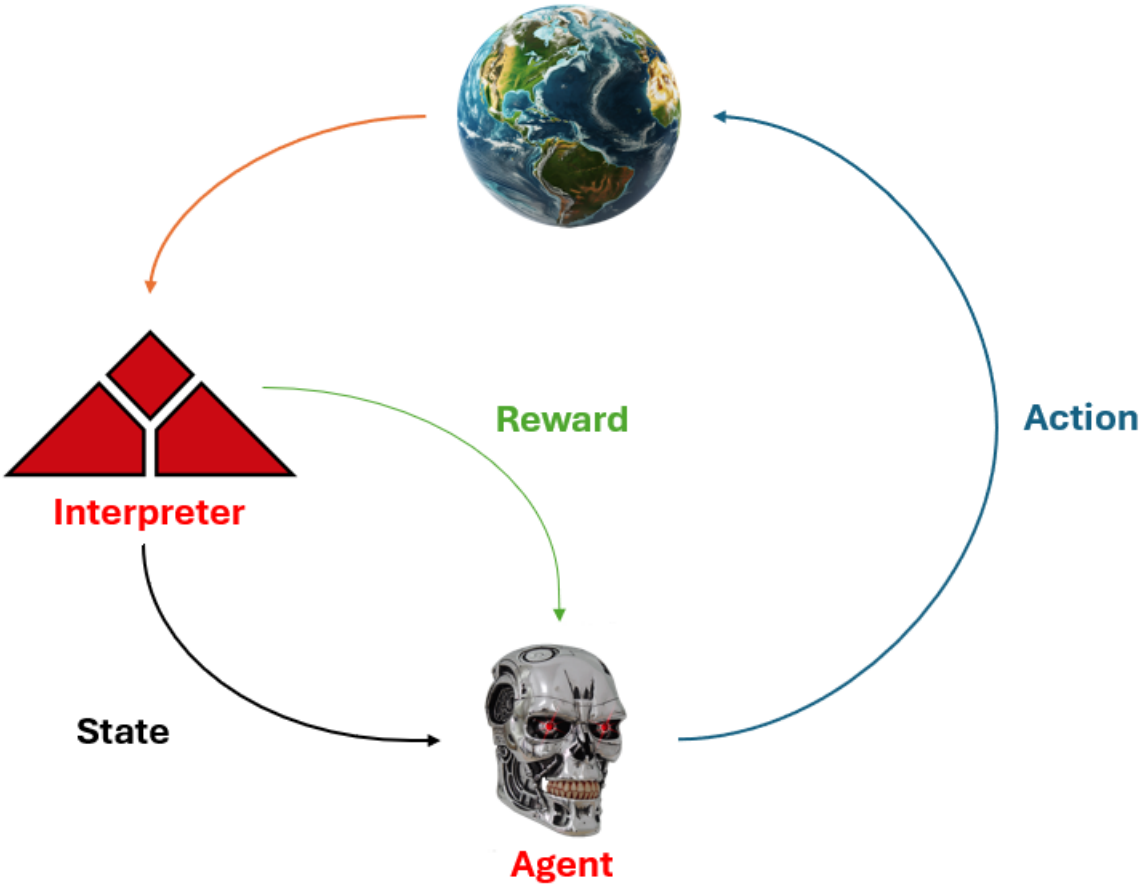

Figure 8. The traditional framing of a Reinforcement Learning (RL) scenario.

Reinforcement learning is one of the major machine learning paradigms, besides Supervised, Semi-Supervised, and Unsupervised learning. It aims to classify data points and utilize labeled datasets to correct actions taken in different scenarios, to maximize the reward function. The key components of a reinforcement learning system are the agent, the environment, and the reward signal. The agent learns to take actions based on factors such as its current state and the feedback provided (reward signal). The environment determines the outcomes of the agent's actions and provides feedback in the form of the

reward signal. The reward signal is a scalar value that reflects how well the agent is performing.

While deep learning applications and the versatility in what kind of problems they can solve sound great, there are certain matters that need to be addressed. Efficiency, quality, stability, robustness, extensibility, and even aspects such as privacy and transparency are important to consider. The standard neural network model requires a lot of data, and if a data model fusion cannot use its resources economically this becomes an issue when the dataset available isn't particularly large.

Determining how to best utilize the fusion network, whether its preprocessing data to improve the model's ability to find discriminating details and what kind of impact it has on information accuracy is an important consideration. The level at which fusion is implemented might also impact the data. Is the information lost from preprocessing worth the feature extraction? Is implementing data/low level fusion worth it? Or for the particular application is it better to simply pool together assessments of the data and implement decision level fusion instead? If an underlying environmental factor is changed, is the fusional quality still ensured? More importantly is the data fusion model capable of being improved or expanded beyond its initial training?

While there are many potential applications for data fusion and machine learning, ranging from fault detection, navigation, multi-target tracking, classification, discussing potential pitfalls for such fusion models is important. Data imperfection, data inconsistency, data confliction, data alignment, data heterogeneity, and the actual location of fusion are the best descriptors of the typical issues machine learning based fusion can

run into. Data captured by sensors is often imprecise, and it is the incomplete and uncertain nature of such modalities that makes sensor fusion and its uncertainty reduction capabilities so appealing.

Data inconsistency is a major issue for fusion models if they are incapable of distinguishing the reasons that cause the noises, or unable to remove or suppress such noise. With regards to systems that utilize belief functions, such as Dempster-Shafer theory, if problems that should be treated independently are erroneously integrated with each other, a representation error can occur, known as data confliction. For data captured from different sensors with different frames, if the information is not properly registered or correlated with each other, then there is the issue of data alignment, the less than harmonious synchronization of information coming from different modalities.

Similar to data alignment, is data heterogeneity. If data of different types or formats have lower quality due to missing values, or high data redundancy or are ambiguous or just plain untruthful, this can be an issue in fusion architecture. Sometimes this error is syntactic in nature, in that the data sources are not communicating with each other. Other times the heterogeneity mismatch might be conceptual in nature, having differences in modeling the same domain of interest. Terminology might also be an issue after solving a syntactic error, with different variations in names when referring to the same entity from different data sources. Finally, after all of the issues with the individual modalities and communication and alignment are solved, there might still be the issue of fusion location. Data fusion might occur in a centralized system or locally in independent

nodes, choosing between trading off matters such as reducing communication burden for data accuracy.

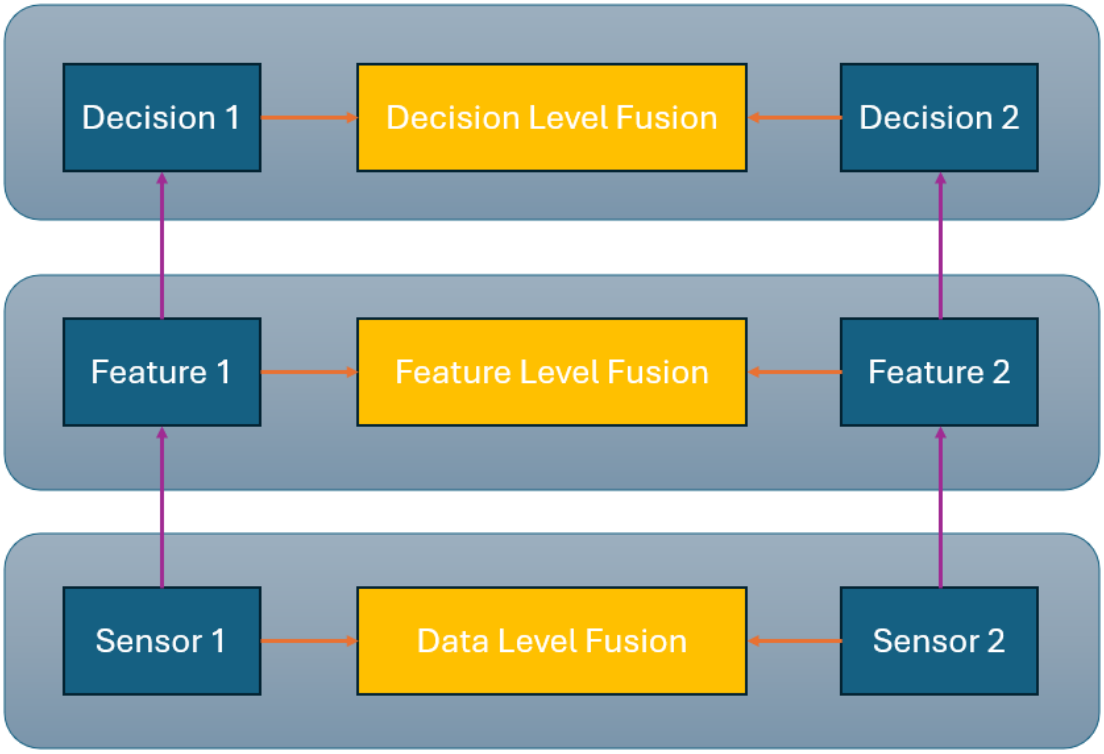

Figure 9. Sensor Fusion Levels (source [5])

Once the sensors are pre-processed and ready for fusion, there are also questions as to what type of sensor fusion scheme is being used. Different sensor fusion schemes exist based on *when* the fusion occurs in the model (Figure 9) and *how* the model uses the sensors that it is provided (Figure 10). There is no magical or universal solution to *how* sensor fusion should be implemented, and more sources of data are not always the best or most efficient path forward. Rather the optimal solution depends on the resources available, the data the model has to work with, the objective of the model, and its environment. Like many things in real life, the optimal solution depends largely on the circumstances of the problem and the resources available. A sensor fusion model does not

necessarily require every source of data available to perform well, but the added modalities generally provide further *uncertainty reduction*.

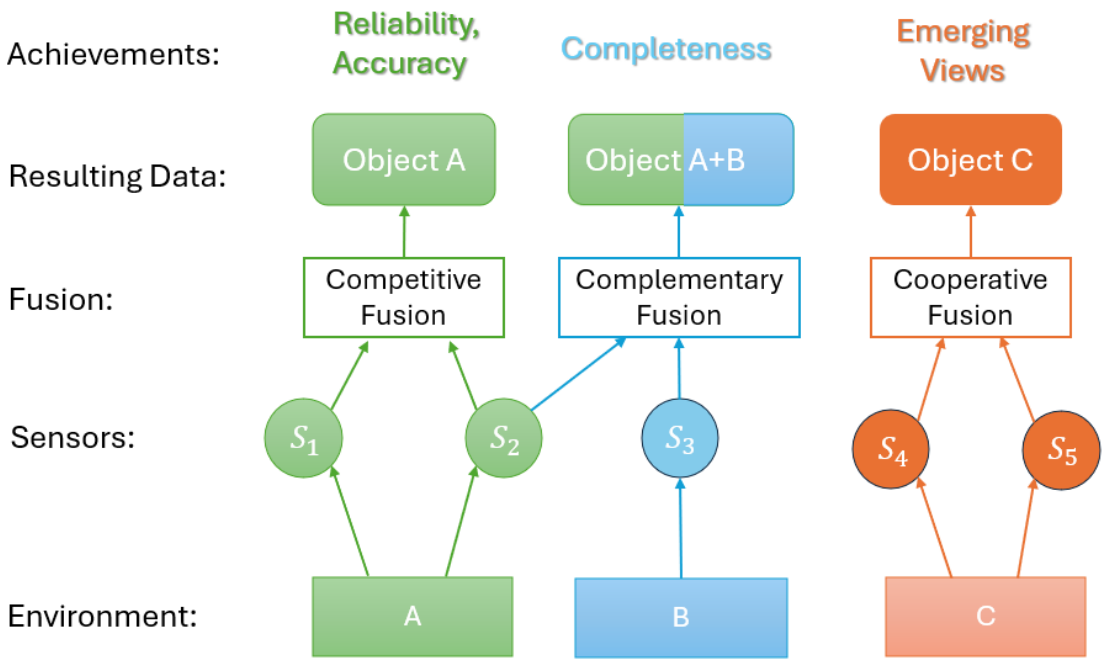

Figure 10. Competitive, complementary, and cooperative fusion (redrawn according to netwerkt.wordpress)

Choosing how to balance fusion cost and fusion quality is an important choice to consider when designing a fusion model. At the end of all of this, there remains one final important question to ask. Most methods of deep learning can produce impressive results. But using methods of deep learning, such as neural networks, essentially makes most of the process something of a blackbox. These hidden layers and nodes are not necessarily designed to be transparent and knowing what discriminating factors are being used to achieve its objective is important to consider. Which leads to the question of how does one know what weights or factors are being used for decision making? And why should that be an important factor for a fusion model?

1.3 The Need for Explainability

The success of deep learning and neural networks has had a major impact in both research and industrial applications. Machine learning has reached a point in which its large impact and potential consequences cannot be ignored, and are even somewhat mainstream with casual usage being fairly common. Having access to these blackbox systems is great, but when Chat GPT spews out misinformation that a plagiarizing student plans on submitting, how does one figure out the cause of what is obviously a failure as far as the LLM's output goes? It's not a terribly important problem to troubleshoot, provided you are not the plagiarizing student in question, and if anything, such a "tragedy" could be considered kind of funny, a sort of modern day Aesop's fable, a warning to younger generations about overreliance on unreliable technology.

But consider slightly more important applications that have nothing to do with plagiarizing papers for classes, such as autonomous driving, buying stocks, or diagnosing cancerous tumors. It only takes even a momentary dysfunction in a computer vision algorithm for a car accident to occur or a single false negative in disease detection to severely impact a patient. This begs an important question about interpretability and explainability of machine learning algorithms. Who is accountable when things go wrong? How did things go wrong? If things are not going wrong, then do we know why? How can we figure out what went wrong? And can we leverage such characteristics or inputs to further improve performance?

There are some who might argue that a “right to explanation” is unnecessary, sometimes even harmful, potentially even stifling innovation. For example, back in 2017 when the European Union announced its General Data Protection Regulation (GDPR) in order to ensure the right to obtain an explanation and a right to opt out, there were those who argued against such measures. These arguments largely consisted of points such as such regulations would do little to help actual consumers and instead would slow down the development and use of AI in Europe. Such critics are quick to point out that holding developers to such a standard would be unnecessary and even unfeasible. To individuals who argue such things are unnecessary, I personally recommend they should exclusively drive with Tesla’s “auto-pilot” feature and put their money where their mouth is. All of this is to say that those critics believe that regulations would be unprofitable because they would require a lot of work to meet such requirements. Some models certainly don’t *necessarily* require explainability, either because of the simplicity of the model or its application isn’t important. However, if a DL algorithm is handling anything important, something that can have a major impact on one or more human lives, basic ethics and common sense indicate that some level of explainability is necessary.

Fundamentally, many algorithms used in deep learning are not easily explainable, being an inherent blackbox for most vanilla iterations of the common DL models that exist. The output of a deep neural network is made of many layers of computations, and the complexities between the input and what aspect of the input activates the neurons is difficult to determine in many cases. These interactions between millions, sometimes billions of parameters are non-linear and complex, making it difficult to trace the path

from input to output of the network to understand how each parameter contributes to a final decision. These are made more complicated by the use of non-linear activation functions, which are responsible for transforming the outputs of one layer into the next, which add to the complexity.

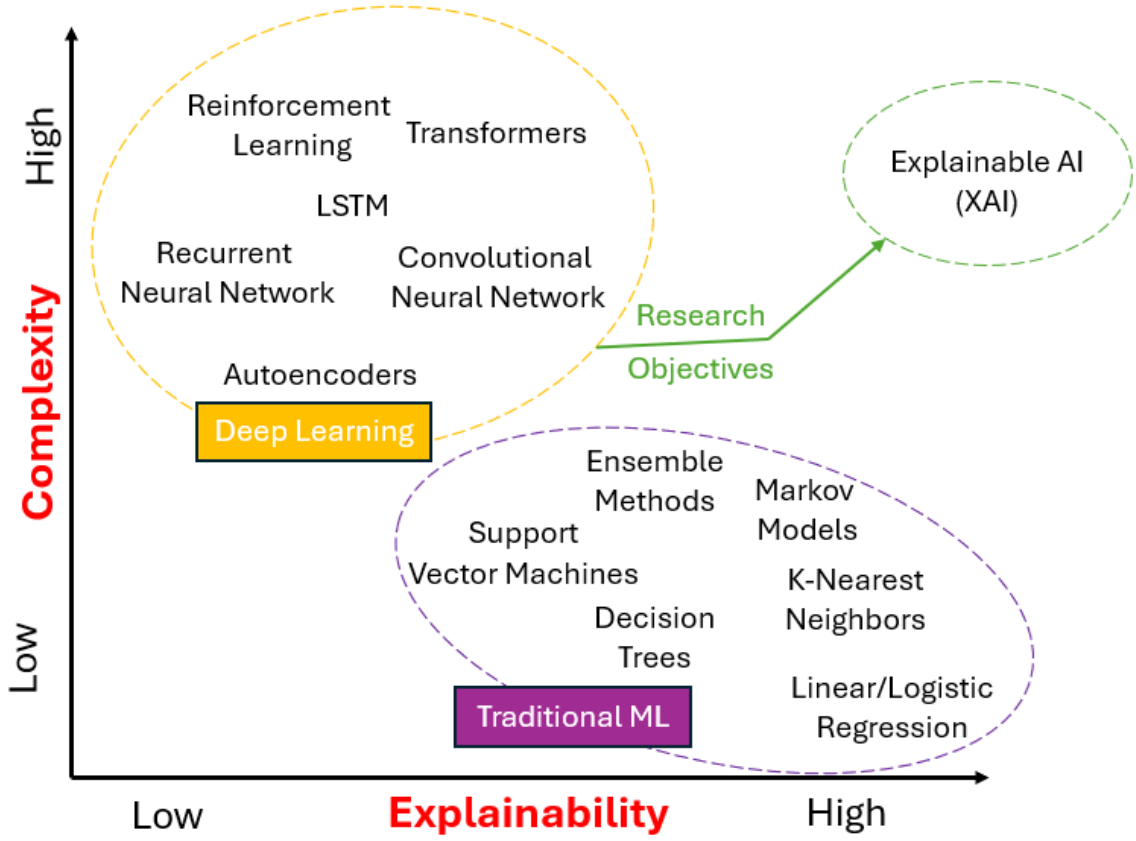

Figure 11. Tradeoff of complexity and explainability in AI

Unlike traditional machine learning algorithms (Figure 11), understanding what associations in the data are being used to drive the decision-making process is difficult to determine on most vanilla incarnations of Deep Learning algorithms. Decision Trees (DT) for example split the data into subsets and produce an actual decision tree to explain how the algorithm makes its decisions. Support Vector Machines (SVM) have been studied extensively since the framework of VC theory was proposed, performing non-

linear classification via the Kernel Trick, and because it uses max-margin clustering, SVM is resilient to the effects of noise on the dataset. Because their complexity is considerably lower than that of Deep Learning algorithms, it is inherently easier to understand the decision-making process of such models. For this reason, integration of explainable algorithms into such architectures is highly desirable, as the ideal goal is to have an understanding of these complex algorithms when they're being used in the field, if not for ethical reasons, then for design and diagnostic purposes.

When using a machine learning algorithm, there are a number of important things to consider when troubleshooting the results. It is not enough to look at the accuracy score and assume that the objective has been reached successfully. When looking at the results that the model has troubles with, it is important to ask *why* did the model output this prediction? Does that prediction align with the domain logic of the application that the model is supposed to solve? Or is the model treating irrelevant features that have more weight when it misidentifies an input? Also important is determining if the model's behavior is *consistent* among different subsets of data. If it is behaving differently when in theory the model should not, then it is important to compare and contrast the results in order to identify where the behavior is varying. It may be that the model is bias towards a specific feature that is *exclusive* to that part of the dataset, and if so then the bias should be removed as to avoid negatively impacting the performance. Determining what influences the predictions is important in order to reach a more desirable outcome for the model.

There are other aspects to interpretability that can also impact the design and training of such models. Due to the nature of how training data is setup or what information is provided, the model may be learning to search for features that might not necessarily be relevant when compared with real data. For example, model trained to recognize the area of a hotel property based on image inputs might misclassify a swimming pool area with that of the bathroom because of shared features it recognizes in the pool portion of the dataset.

Without any form of explainability, it becomes difficult to determine how and why this error occurred, but with interpretability the reason might be obvious. With the application of Gradient Class Activation Maps, the heatmaps might indicate that the bias that caused the model to misclassify the swimming pool image as a bathroom image was in fact the metallic rails both areas share. Machine learning algorithms have been proven not to be immune to bias [6], and by using explainability [7] and visualization techniques it becomes possible to reduce the uncertainty and bias [8] that comes with training a model that's essentially a blackbox. Considering some of the uses of AI in the modern day are in fields such as algorithmic trading [9], medical diagnosis [10], and autonomous vehicles [11] and the major impact they can have on their consumers, it is clear that these models that handle such important applications should probably have at least *some level* of interpretability.

1.4 The Need for Efficiency

Neural networks are a class of parametric models that achieve state-of-the-art performance over a wide range of tasks, but their heavy computational requirements hinder practical deployment on resource-constrained platforms, such as mobile phones, Internet-of-things devices, and offline embedded systems. The training of these DL methods are computationally expensive, energy intensive, and result in considerable carbon emissions, a known driver of anthropogenic climate change. Additionally, the platforms which these DL systems run on are associated with environmental impacts including and beyond carbon emissions from power consumption. These can include the use of freshwater consumption for cooling, an issue that climate change is projected to worsen, pollution from e-waste, the mining for resources needed to manufacture such DL systems, and other related byproducts.

Moreover, there are extreme expenses involved in the development of frontier models, and their sustained usage that need to be addressed. This is not helped by the painfully human attitude of excess, which manifests itself in AI in many ways, such as an attitude that "data is all you need" and relying on exponentially larger amounts of data to train these models. With the excessive use of data and resources (power, cooling, hardware, etc.), it is clear that the trends we are seeing continue, it will make serious AI development restricted in a similar way to that of space travel: restricted to governments and billionaires.

In recent years there has been a lot of hype over the usage of AI that has been pushed in a number of products, such as Google's and Bing's AI integration, OpenAI's Sora text-to-video generator, and Apple's AI integration with Siri's upgrades. However,

while billions are being poured into AI development, there are genuine concerns regarding the efficiency and practicality of these proposals for how AI is used, and whether or not current practices have enough data [12] to complete the training needed. While AI has been put to use in practical applications that make economic sense to use, such as targeted advertisement or rejecting insurance claims, that does not make every ridiculous proposed use of AI *economically viable* as a business practice.

There are a number of different costs attributed to AI [13], from the development costs of the necessary hardware, their depreciation over time, the actual energy consumption cost, and related cloud computing cost. Reducing the energy impact of training DL models can be accomplished using an astonishingly simple yet effective method, power-capping. By simply putting limits on power consumption, it provides significant benefits across multiple computing platforms at minimum cost. This can take the form of limiting the power the NVIDIA chips are able to use while not impacting the code, only requiring minor hardware level changes. Besides this, scheduling based on energy-awareness can also provide benefits with the advent of solar energy.

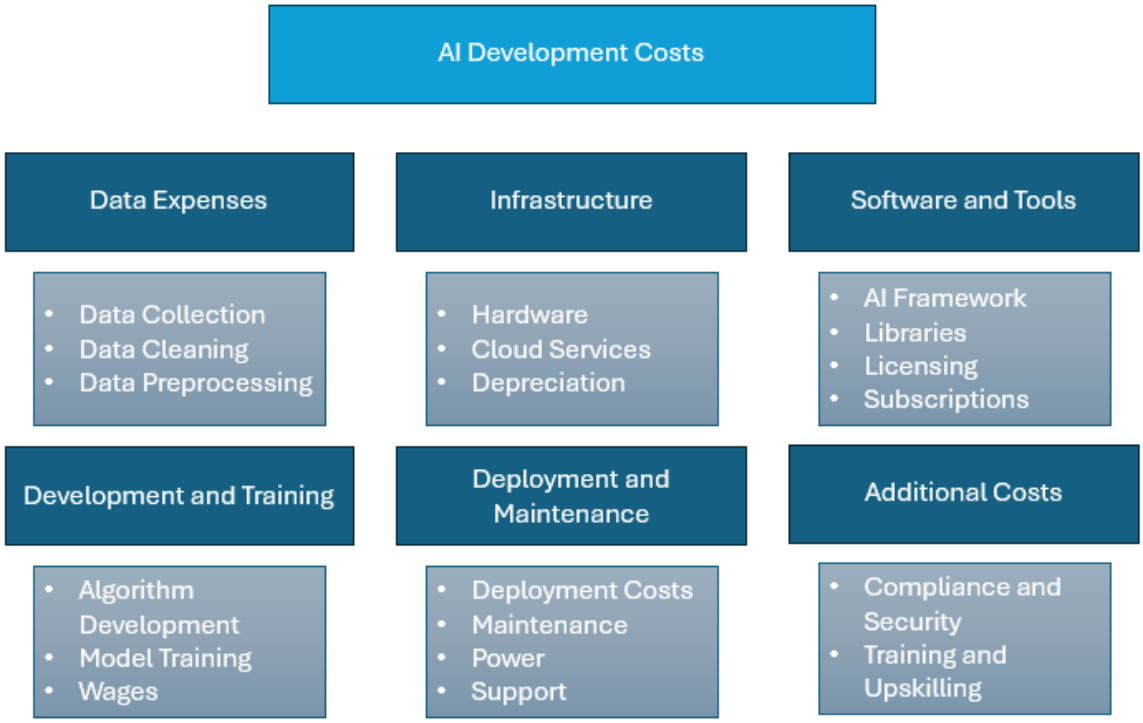

Figure 12. Breakdown of AI Component Costs

There are also other costs that must be considered, which are associated with the collection of the training data, as well as the cost of the actual research and development staff (Figure 12). And even worse, are the costs associated with combating data drift [14], which will entail re-training and corrections needed to negate the effects of the phenomenon. It is a resource and time-consuming process to determine what data is at fault, and what the model's failures stem from. Not to mention the work that is required in order to successfully re-train and test the model in order to correct for any failures that are caught.

Unlike the development of integrated circuits, in which Moore's Law and the physical limits of the atomic scale played a major role in the growth of the industry, there have only been exponential costs with the training of further iterations of frontier models

and questionable business practices. Human beings often equate bigger with being better, and are rarely known to take an approach that isn't exorbitantly wasteful even if their lives depended on it. And this is a trait which has been seen painfully in the development of AI by western companies. With papers indicating a genuine concerns regarding the amount of training data available on the *internet itself* for Large Language Models (LLM) [13] and Generative AI models to utilize, the likelihood of AI technology reaching the capabilities that are being proposed by the industry. Say nothing of the Singularity, none of these proposed applications of AI unlikely to happen anytime soon with the current algorithms and their exponential training data costs.

Therefore, it is absolutely important to consider the efficiency of the training of AI, both in terms of processing power and economic costs. Naturally, as a new field, AI does not have a lot of metrics which are universally agreed upon by all experts, for things like efficiency or explainability. However, Floating Point Operations Per Second (FLOPs) are commonly used as a means of measuring computational costs. Metrics for calculating cost efficiency for power often include metrics such as FLOPs/Watt or FLOPs/Clock [15].

TECHNOLOGY

China's DeepSeek is Our Digital 'Pearl Harbor' Moment

How we deal with China's rising tech prowess in the next few years, will determine the next 100.

By Matteo Wong

Figure 13. Example of the corporate news response to DeepSeek (source: The Atlantic).

The AI development practices in the west have clearly been nothing if not wasteful, again ignoring any alleged misuse of copyrighted data or ethical issues, a fact that was recently and deliciously illustrated by the fallout over the recent uproar over the open-source LLM DeepSeek in late January, which saw even companies like Nvidia drop in market value as the estimates in losses were something akin to at least a trillion. It goes

without saying that money decides everything in the end. Therefore, it is of the upmost importance that sustained and continued AI development will require some level of cost-effectiveness and implement sustainable practices.

1.5 Thesis Contribution to the Current State of Knowledge

For the last few years, our research group has been focused on the usage of passive radiofrequency (P-RF) data for detection and differentiation of targets, both human and vehicle. While our most recent research has been focused on P-RF modalities, given that the majority of the vehicle detection research that we've published has been done with the ESCAPE dataset, this thesis integrates the use of the other available modalities within that dataset and showcases a deep learning framework evaluation approach which compares the impact of each modality, each sensor fusion model scheme's predictive performance, and computational efficiency (measured in FLOPs). The culmination of half a decade of work towards better understanding the impact of these modalities in the fusion process, how this data is best utilized, and how to incentivize better fusion is covered within the contents of this thesis. In particular, the contributions of the research presented here are as follows:

1. The development of a deep learning framework evaluation approach which compares, provides explainability, and evaluates sensor fusion deep learning models for the purposes of achieving multimodal fusion and detection of vehicle

targets, while also looking for different combinations of input data in order to better incentivize sensor fusion between different modalities [See Chapter 5].

2. The usage of XAI and other metrics to determine the impact of each of the modalities used, tested over each potential permutation of the available sensors for multimodal fusion [See Chapter 6].
3. The exploration of exploiting and understanding the usage of P-RF modality for vehicle target detection [See Section 4.3].

The following Journal publications have been produced from the research described in this thesis:

1. A. Vakil, E. Blasch, R. Ewing, J. Li, "Finding Explanations in AI Fusion of Electro-Optical/Passive Radio-Frequency Data," *Sensors*, vol. 23, no. 3, 1489, Jan. 2023. DOI:10.3390/s23031489
2. L. Yuan, J. Andrews, H. Mu, A. Vakil, R. Ewing, E. Blasch, J. Li, "Interpretable Passive Multi-Modal Sensor Fusion for Human Identification and Activity Recognition," *Sensors*, vol. 22, no. 15, 5787, Aug. 2022. DOI:10.3390/s22155787
3. A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, J. Li, "A Survey of Multimodal Sensor Fusion for Passive RF and EO Information Integration," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 44-61, Jul. 2021.

4. J. Liu, H. Mu, A. Vakil, R. Ewing, X. Shen, E. Blasch, J. Li, "Human Occupancy Detection via Passive Cognitive Radio," *Sensors*, vol. 20, no. 15, 4248, Jul. 2020. doi: 10.3390/s20154248

The following Conference publications have been produced from the research described in this thesis:

5. A. Vakil, R. Ewing, E. Blasch, J. Li, "Explainable Hybrid Decision Level Fusion for Heterogenous EO and Passive RF Fusion via xLFER," *NAECON 2023 - IEEE National Aerospace and Electronics Conference*, pp. 221-226, Aug. 2023. DOI: 10.1109/NAECON58068.2023.10365911
6. A. Vakil, E. Blasch, R. Ewing, J. Li, "MVI-DCGAN Insights into Heterogeneous EO and Passive RF Fusion," *Prof. of 2022 9th International Conference on Soft Computing and Machine Intelligence*, Nov. 26-27, 2022. DOI: 10.1109/ISCMI56532.2022.10068480
7. A. Vakil, R. Ewing, E. Blasch, J. Li, "Visualization of Fusion of Electro Optical (EO) and Passive Radio-Frequency (PRF) Data," *Proc. of IEEE National Aerospace and Electronics Conference*, pp. 294-301, Aug. 2021. DOI: 10.1109/NAECON49338.2021.9696424
8. E. Blasch, A. Vakil, J. Li, R. Ewing, "Multimodal Data Fusion Using Canonical Variates Analysis Confusion Matrix Fusion," *Proc. of 2021 IEEE Aerospace Conference*, pp.1-10, Mar. 2021.

9. J. Andrews, A. Vakil, J. Li, "Biometric Authentication and Stationary Detection of Human Subjects by Deep Learning of Passive Infrared (PIR) Sensor Data," *Proc. of 2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1-6, Dec. 2020.
10. J. Andrews, M. Kowsika, A. Vakil, J. Li, "A Motion Induced Passive Infrared (PIR) Sensor for Stationary Human Occupancy Detection," *2020 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 1295-1304, 2020.
11. A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, J. Li, "Feature Level Sensor Fusion for Passive RF and EO Information Integration," *Proc. of 2020 IEEE Aerospace Conference*, pp. 1-9, Mar. 2020.
12. J. Liu, A. Vakil, R. Ewing, X. Shen, J. Li, "Human Presence Detection via Deep Learning of Passive Radio Frequency Data," *NAECON 2019 - IEEE National Aerospace and Electronics Conference*, pp. 296-301, Jul. 2019.

CHAPTER TWO

LITERATURE REVIEW

2.1 Passive Radio Frequency and Electro Optical Fusion

With the advent of the Information Age and the vast availability of data, sensor fusion has become a focus of many different domains. While there is no common model of sensor fusion or any singular comprehensive system of classifying sensor fusion methods [16], the majority of existing models propose partitioning of the information fusion architecture based on how the source input information undergoes preprocessing, feature extraction, pattern processing, situation assessment, and decision making [17]. There are many different paradigms of sensor fusion [18], that are differentiated into many different methods of sensor fusion, such as *when* in the process the fusion occurs (data-level, feature-level, decision-level) [19], and *what* types of data are being used in the fusion (heterogenous vs homogenous).

Heterogenous sensor fusion between sources of data that are inherently different from each other can come with its own challenges [20], but can also provide a wide range of benefits [21]. There will always be a series of trade-offs in advantages and disadvantages with any sensing method, but what sensor fusion successfully provides, regardless of the application or modalities, is the reduction in *uncertainty* to the model. Determining what sources of data are worthwhile to include can be difficult, especially in blackbox systems, and while there is a sentiment that *more data* automatically is a better, that is not always the case. In terms of efficiency, viability, and what the fusion model

actually learns, the importance of how the chosen data sources are integrated into the model is key to achieving peak sensor fusion.

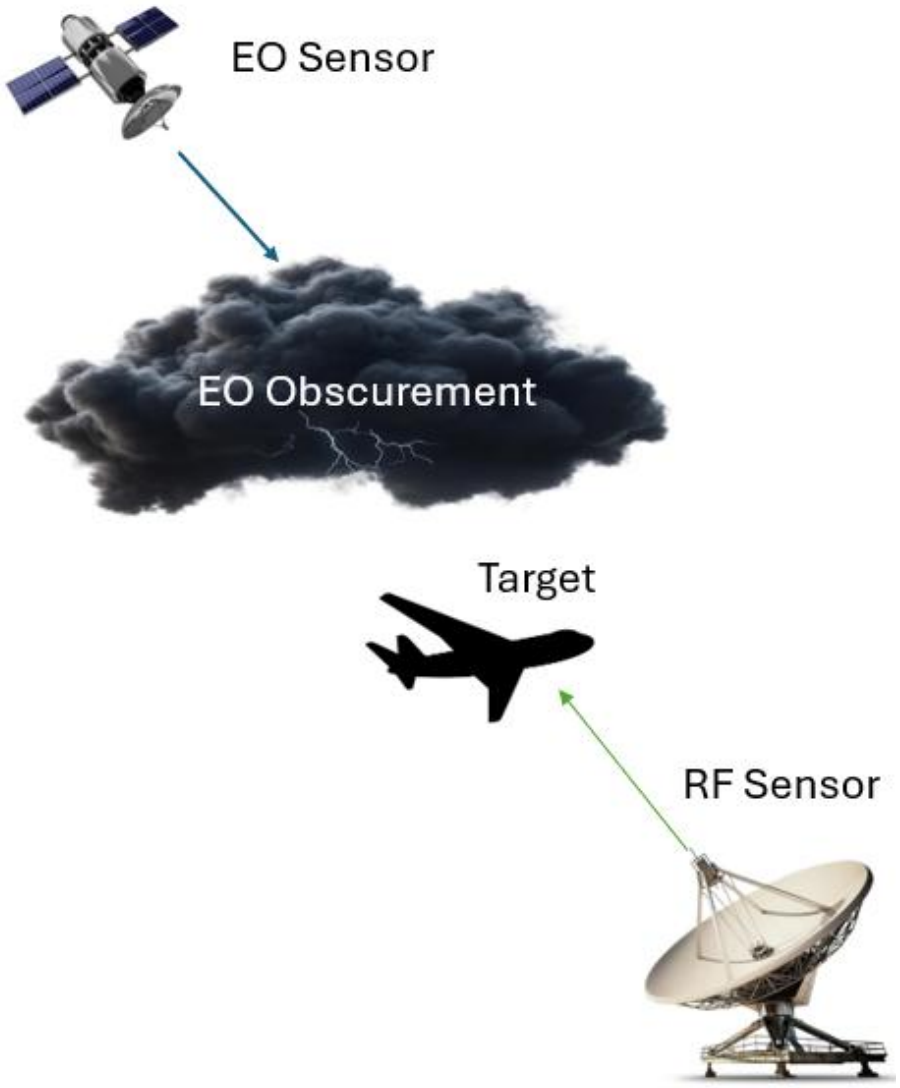

Figure 14. Example of the benefits of EO/RF fusion

When it comes to modalities that involve radio frequency (RF) and electro-optical (EO) sensors, the focus for fusion research has traditionally been on active RF sensors and applications. Whether it is the fusion of Synthetic Aperture Radar (SAR) and multispectral images via random forest classifier [22] or double weighted decision level

neural network fusion scheme [23], many EO/RF sensor fusion applications normally use active RF in one form or another. Active Radar, such as doppler radar and imaging radar (e.g., side-looking airborne radar), as well as other similar active RF sensors, are well suited for tracking a moving target, a combination that is highly desirable when successfully fused with EO input. Other approaches use sensor fusion at multiple levels (data, feature, and decision levels) such as with SAR and multisource satellite imagery [24] and other features like 2-D radar, EO, IR with extended Kalman Filter, State Vector Fusion [25]. In Zhang *et. al.* [26] , the use of Rao–Blackwellized particle filtering was implemented in order to fuse the asymmetrical fields of view for the radar and EO modalities. The application of adaptive waveform design and control incorporated dynamic agility selection in order to improve the performance of the system's ability to track an unknown number of targets using the two modalities. Other notable algorithms include the use of sparse representation in order to combine medium wavelength IR (MWIR) cameras and RF Doppler sensors for vehicle tracking [25], which uses a joint sparse approximation approach for multimodality images.

Finally, in Robinson *et al*. [27], the use of simulated multisensor data is used to locate a moving emitter. The method of fusion implemented is Sheaf Theory, a tool for systematically tracking locally defined data attached to the open sets of a topological set. For the purposes of implementing sensor fusion, the data samples and model of data are used as the inputs of a sheaf-based fusion architecture. The model of the data is used to construct the sheaf while the data samples are converted into samples for partial assignment. The outputs of the two are then used to search over the global sections using

the optimizer, before using the results to report values over the stalks. During testing, the observed stalks for each sensor measure the real-time offset and complex I/Q samples for each RF sensor while the EO data are collected, keeping the xy-location of each detected pixel for the video input. For the modeled stalks, which provide a comparison for each pair of sensors, the time offset relative to the video detection and the time aligned I/Q samples for each group of RF sensors are tracked. The third vertex tracks the true location of the ground truth, the emitter, and transmitted signal.

The combined view and the exploitation of the two types of sensor modalities still have room for improvement [28] however. RF based modalities excel in providing range, angular, and spectral resolution of information from RF modalities and the benefits of combining that data with higher spatial resolution of EO based sensors are extremely desirable for detection and tracking of targets in a number of different environments.

There are a number of RF based modalities that are used in applications such as tracking, proximity, localization, and detection. While many EO modalities are intuitively easier for humans to understand and to implement for similar applications, unlike RF modalities, RF approaches to such applications are less susceptible to outer factors such as ocular interference. RF-based sensors are not limited or obscured by factors like visual interference from natural phenomenon such as fog, clouds, snow, or any other form of weather that would otherwise normally interfere in the collection of EO data. In addition to this, RF based sensors can provide repetitive coverage over a wide geographical area, and in doing so can determine the precise distance and velocity of a target.

While most research is focused on active RF applications, there are a number of advantages for the implementation of passive RF modalities such as passive radar or RFID over the use of active methods. Passive RF modalities are difficult to detect, typically having lower power requirements and have lower costs than the ones associated with the construction and usage of active radar, and are harder to implement countermeasures against, such as jamming and spoofing which can corrupt the collection of RF-based modalities and transmitted imagery. Combining the two modalities improves the overall reliability and have been implemented in a few applications for target detection, estimation, and tracking.

Sensor fusion algorithms are generally evaluated in terms of accuracy and robustness. During experimentation, certain trials will likely be completed in order to compare the robustness of the model. Methods such as adding visual occlusion or noise will normally be implemented, with other methods of distortion to ensure that the model can use that training in unfamiliar situations. But sensor fusion is the process of combining measurements from multiple nodes, each of which have a certain level of uncertainty. When these sources of information are fused together, it is not always clear how these uncertainties will interact and influence the overall performance of the sensor fusion algorithm. It is important to collect information in order to gain insight into the performance of an implemented fusion architecture.

For the vast majority of the many sensor fusion applications that are NN based, F-1 Score is traditionally used to score the performance. The measurements of precision and recall, the measurements that compare the true positive rate and the sensitivity are

commonly accepted for most classification problems. However, not all EO/RF sensor fusion applications are focused on classification. Some are made for tracking and estimation purposes, and therefore need to be compared to a ground truth. For machine learning however, there are a few more factors that need to be addressed when testing a NN-based sensor fusion architecture.

One of the major characteristics and concerns for machine learning in general is the nontransparency of the models. Deep learning itself suffers from several major limitations, requiring vast amounts of training data, having poor ability to represent uncertainty, being easily fooled by adversarial examples, and being difficult to optimize. Because of the black-box like nature of such networks, it can be difficult to perform reasoning with them and these aforementioned qualities mean that they are prone to generalizing poorly or overfitting the data. In order to improve these issues with uncertainty, the implementation of Aleatory Variability and Epistemic Uncertainty are commonly used approaches. Aleatoric variability is uncertainty inherent in observation noise while Epistemic Uncertainty is ignorance about the correct model generated by the data, the parameters, the convergence, etc. In Tagasovska and Lopez-Paz [29] specifically, the use of Simultaneous Quantile Regression and Orthonormal Certificates are used as a loss function to estimate Aleatoric Variability and Epistemic Uncertainty, respectively. There is an important distinction between a system malfunction, a failure that is recognized, and a normal operation where false-positives can occur, and the use of uncertainty is an important measure to help interpret the results of a NN-based fusion method.

For the multilevel map classification via SVM [30], Kulin et al. compared the producer's accuracy and user's accuracy, the complements of omission and commission error, respectively. Producer's accuracy is the map accuracy from the map maker's point of view. This metric for mapping measures how often real features on the ground are accurately shown on the classified map and probability that a certain land cover on the ground is classified correctly. Conversely, the user's accuracy is accuracy from the point of view of the map user. The metric can be summarized as reliability, calculating the total number of correct classifications for a specific class and then dividing it by the row total.

In [31], concerns about the stability of a fusion system that is trained end-to-end is not encouraged by its readers, due to the potential incompatibility with assumptions about the stable hierarchical architectures of components. In Yang *et. al.* [32], the creation of an explainable neural network (xNN) in order to improve the understanding and provide sufficient model interoperability is explored. The estimation of multiple parameters via a modified mini-batch gradient descent method derived from the backpropagation for calculating derivatives and the use of the Cayley transform is implemented to preserve the projection orthogonality. The approach improves interpretability of the model while maintaining the prediction accuracy. While there has been research into creating a more transparent models [33], the vast majority of machine learning methods are all-training based and nontransparent, which makes them being able to use metrics in order to properly interpret the results all the more important.

For the evaluation of their joint manifold learning framework, Shen *et al*. [21] compared their results for vehicle detection and estimation to the ground truth. In order to

better test the reliability of their algorithm, the implementation of white noise and position shifts are applied in order to test if the framework can apply its learned intrinsic mapping from one scene to a similar one. In addition, comparisons were made with other traditional sensor fusion algorithms in multiple scenarios, relating the estimated trajectory with the ground truth and plotting position errors over the respective frame index.

For the passive multispectral radar/EO/IR sensor fusion architecture for UAS in [34], the use of a Cramer–Rao lower bound (CRLB) on localization accuracy is used. In similar applications, such as [35] or [36], which seek to avoid collision between unmanned aircraft, a simulation system for the EO and radar input is used to implement tracking, trace range, and azimuth error for the RF modality in terms of the mean and standard deviation. Performance indices were used to measure the performance of the UAS system, emphasizing the characteristics of the response deemed to be relevant to optimize the control system and providing feedback. Besides metrics that directly relate to the evaluation of accuracy measurements, another aspect of performance to consider is computational cost of the process and delay caused by the fusion architecture. As previously mentioned, the design of decentralized or hierarchal fusion architectures are less taxing in terms of computational cost, but in exchange introduce more delays in communication. Certain models actually exploit the nature of decentralized architectures, improving the individual node estimators by capturing the correlation between the sensor observations of matching parameter values for different manifolds [37]. Aside from the metrics of computational cost and efficiency, the size of training data is also relevant. The main problem that all fusion methods face are the constraints in the execution of data

acquisition, processing, and the implementation of the algorithm. In Martın *et al*. [38], the accuracy, computational costs, and memory fingerprints for the traditional classifiers Naïve Bayes, Decision Table, and Decision Tree were calculated for different sensor data and optimization method. These metrics provide insights into how the individual modalities interact with the fusion, particularly with the memory fingerprints. While there is no universal benchmark for fusion evaluation, these are some measures that provide for better understanding the fusion architecture and assessing the impact of different modalities.

2.2 Explainable AI

As explainable AI (XAI) is an emerging field in research and industry [39], there has yet to be a widely adopted standard, say nothing of any kind of widely adopted method for quantifying interpretability for explaining models. Even discussing methods of how to quantify such approaches is quite a daunting task in of itself, as there are a large number of different classifications for such types of interpretability [40] that are generally accepted. To even *begin* discussing the topic of explainable AI, the most important thing to do is to define interpretability.

There are a number of definitions of interpretability or explainability, but for the purposes of this paper the two terms are used interchangeably. For domains that deal heavily in images for example, interpretability might be defined as being able to map the predicted class into a domain that the human user might be able to make sense of. In an

ideal system, one might even define interpretability as a reasonable explanation as to why a collection of features contributed to the decision-making process, or at least determining how much weight the decision-making process gave to said features. Whichever definition of interpretability one might subscribe to, given the lack of a widely adopted standard, so long as the method provides insights that can answer questions regarding how and why the model performs in the way that it does, that is a method that provides some level of transparency.

From saliency maps to activation maximization, there are a number of methods by which interpretability can be achieved. The distinctions between these types of methods are typically twofold, described as either ante hoc or post hoc, local or global, or model-specific or model-agnostic. Ante hoc and post hoc describe an intrinsically interpretable model from different approaches. Ante hoc systems provide explanations from the beginning of the model, such as the Bayesian Rule List, a generative model that yields posterior distribution over decision lists consisting of a series of if–then-statements. Other examples can include methods such as visualization, saliency mask, rule extraction, and even neuron activation. Post hoc techniques, on the other hand, focus on creating explainability in a model based on the model's outcome, marking the part of the input data responsible for the final decision. Similar to ante hoc techniques, this can include visualization and saliency mapping, and can use methods such as gradients and feature importance as well.

Besides any of the means of providing explainability, one of the simpler methods of providing transparency is the use of smaller DL models [41], which owing to not being

as complex are inherently more explainable than more complex DL models. In general, training data will contain trends which a DL model can exploit to make accurate predictions [42]. However, once the trends become more complex, the model's decision making process can be adversely affected by irrelevant factors, and thereby require more data. By focusing on a first principle approach, i.e. breaking down complex problems into a sequence of smaller problems that build on each other, using an effective system of smaller DL models can provide a more inherent transparency. For example, in [43] demonstrated a targeted preprocessing of radar data that made the critical features for identifying interference more consistently observable, thereby allowing a simple DL model to process the data more effectively. By having a model learn from robust and viable behavior from a representative environment, the model can be trained continuously while interacting with the real world [44].

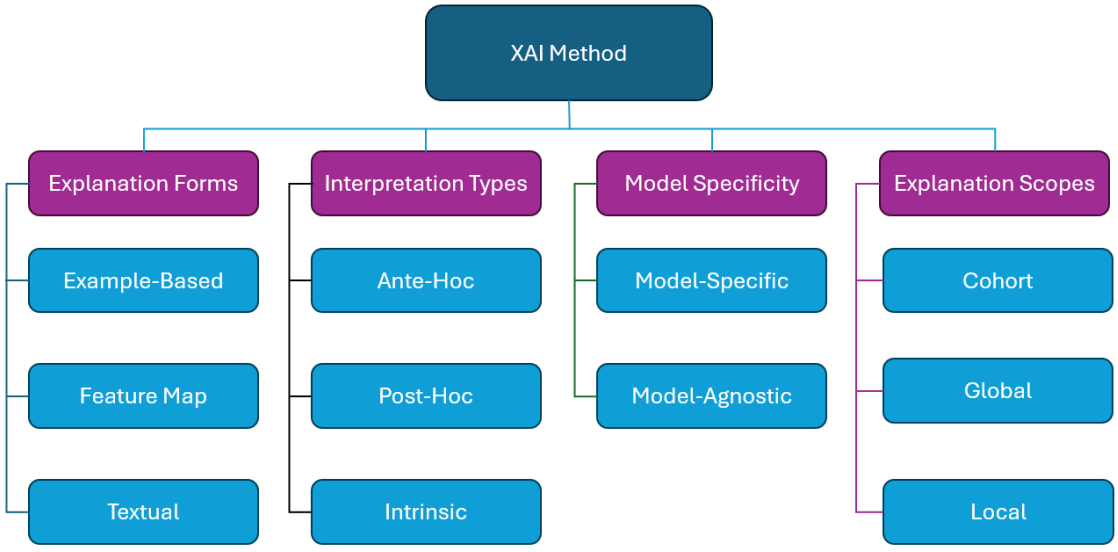

Figure 15. Commonly used XAI Methodologies

If the developers of an AI model don't care about things like first principle approach or smart design practices, and just force fed a DL algorithm with data, there is

still ways of getting explanations from that model (Figure 15). Broadly speaking, while there is no common consensus on the classification of different XAI methods, there are general categories that they can be placed in. These include Explanation Forms, Interpretation Types, Model Specificity, and Explanation Scopes.

From saliency maps to activation maximization, there are a number of methods by which interpretability can be provided by such models. The distinctions between these types of methods are typically twofold, described as either Ante-hoc or Post-hoc [45], local or global [46], or model specific or model agnostic [47]. Ante-hoc and Post-hoc describe an intrinsically interpretable model from different approaches. Ante-hoc systems provide explanations from the beginning of the model. These types of systems enable one to gauge how certain a neural network is about its predictions.

An example of an Ante-hoc system includes the Bayesian Rule List [48], a generative model that yields posterior distribution over decision lists consisting of a series of if-then-statements. Other examples can include visualization methods, saliency mask, rule extraction, and even neurons activation. Posthoc techniques entail creating explainability into a model based on its outcome, marking the part of the input data responsible for the final decision. Like Ante-hoc techniques, this also can include visualization and saliency mapping, but also uses methods such as gradients or feature importance. These methods are more easily applied in comparison to different models but say less about the whole model in general.

Similar to but not quite the same descriptors are local and global interpretability. Local provides explanations for only each single prediction while global explains the

logic for the whole system, from input to every possible outcome. Methods such as GradCAM [49] are an example of local interpretability systems, using global average pooling and heat maps of a pre-softmax layer in order to determine the regions of an image responsible for prediction. Lastly, model specific and model agnostic describe the usability of different aspects of the system, with model agnostic being indifferently usable while model specific is tied to a particular type of blackbox or data. For methods that are Post-hoc oriented solutions to explainability, there exists a number of methods for image based neural networks.

Visualizations [50], gradients [51], activation maximizations [52], deconvolutions [53], and decomposition [54] are common approaches. Visualizations techniques will typically use tools such as generative models or saliency maps in order to determine activations produced on each layer of a trained CNN or DNN after processing an image or video. The visualization of the key neurons or neuron layers highlights the responsible features that lead to a maximum activation or the highest possible probability of prediction. Gradients and variants of guided backpropagation similarly emphasize the important unit changes, drawing to attention sensitive features or input data areas. With these techniques it can potentially be possible to produce artificial prototype class member images that maximize the neuron activation or class confidence values.

Other types of explanation forms include the use of textural and example-based approaches. Textural based approaches are an extension of language-based approaches, and the meta appears to be the use of rule-based techniques like fuzzy logic in order to capture both image segmentation imprecision and the vagueness of human spatial

knowledge and vocabulary. Example-based approaches tend to be model agnostic and include techniques such as DiCE (Diverse Counterfactual Examples) and Adversarial Example methods.

Deconvolution, sometimes referred to as inverting DNNs [55], can be applied to create special typical inputs or parts of an input. These special inputs are created to fit the desired output of the network, producing a special layer or single unit in order to recreate the results. Decomposition, also known as isolation, transfer, or limitation of portions of networks can provide further insights into which way single parts of the architecture influence the output layer. Methods such as Automatic Rule Extraction [56] and Decisions Trees are similar to the application of Deep Taylor Decomposition.

How and when XAI is applied is dependent on the application. Some DL models are more than content to only apply it when necessary, for the purposes of troubleshooting, development, etc. Other DL models directly integrate XAI into computer vision [57] or bias the model toward policy interpretation [58] in order to provide better explainability. How this transparency is used is, as always, up to the developer, but having some level of understanding is necessary for the ethical training of DL models that handle important tasks and applications.

2.3 Efficiency of Machine Learning

Owing to this widespread adoption of DNNs, there has also been pressure to build and train DNNs that are even more complex, using larger amounts of data, but this

growth is not necessarily compensated for by an increase in efficiency. While machine learning has spread to various domains and been a topic of heated debate, one topic worth discussing is what *defines* efficiency for what is otherwise an extremely costly algorithm to train. The evaluation of the efficiency of a deep learning algorithm involves striking a balance between the computational resources consumed, training time required (for both initial training and re-training as needed), prediction time the model is capable of achieving, its scalability, data requirements, model size, and energy consumption.

Addressing the model's full impact, its final carbon imprint, depends on not only just the internalities (the hardware and software directly involved in their operation for both training and re-training) but also the externalities (all the social and economic activities around the use of the model). The model's actual accuracy and performance metrics, algorithmic efficiency, robustness, resilience, hyperparameter tuning efficiency, overfitting, generalization, and even interpretability are also important factors to consider. While explainability has been discussed in the previous section, the discussion of efficiency is also an important topic to consider as quantifying [59] this aspect of performance is important for future AI development.

Typical procedures for providing an AI service include first *training* a machine learning model from data, and then performing *inference* with the trained model. This typically necessitates in some form or another data collection, data preprocessing, model selection and training, model evaluation, deployment, and lastly monitoring and maintenance. Deciding which relevant data is necessary for the objective of the trained model is important, as is determining how that data is read by the model after

preprocessing. The type of model is highly relevant, as the modality and application may make one type of DL algorithm more desirable than another.

For example, language-based models fundamentally require the temporal component of the sequential data to be considered (for the process of understanding a sentence for example), so a model like a CNN is less likely to be used than for example a RNN. Following validation, the model needs to be integrated into a software system or application by creating a user interface and then has to be monitored and maintained to stay up to date and effective. Identifying mistakes and anomalies are important for addressing problems and handling issues such as data drift.

The performance of the model can be measured by its model accuracy, which can potentially be improved by collecting and training more data. However, training and any necessary re-training to avoid data drift or fix potential errors that the model has been making can be time consuming. In order to train a model efficiently, distributed architectures are often adopted, which can introduce additional communication costs for exchanging information across nodes. The computation and communication costs will grow exponentially for these high-dimensional deep learning models. Additionally, low-latency is also a critical for inference in applications (such as smart vehicles, smart homes, or smart drones, etc.).

Depending on the type of architecture being used, this can also exacerbate the complexity and costs of the system. Edge AI poses further constraints on the algorithms and system architecture when compared to cloud based AI [60]. Edge nodes have limited resources, unlike the large number of powerful GPUs and CPUs that integrated servers

for cloud-based AI will have. This results in limited computation, storage, and power resources on edge devices, with limited link bandwidth among large number of edge devices and servers at base stations and wireless access points. Designing efficient edge AI is challenging owing to the coordination and scheduling for edge nodes to efficiently perform a training or inference task under various physical and regulatory constraints. In order to provide efficient AI services, new distributed paradigms for computing, communications, and learning need to be jointly designed [61]. While the communication costs for cloud-based AI services may be relatively small compared to the computational costs for cloud-based AI, but in edge AI it becomes a dominating issue owing to the constraints of the system.

Ignoring the specifics of what type of AI computing is being provided, there are also concerns and constraints over the privacy and security requirements [62] such architectures would naturally require. DNNs are deployed almost everywhere [63], from smartphones, microelectronics, cars, which all come with their unique computational, temperature, and battery limitations for each application. For smaller applications, DNNs that require less resources, accepting trade-offs in performance that normally occur between larger and smaller DNNs. If the DNN is not hosted in-device, then customers are repeatedly requesting access in a transparent way [64], causing millions of inferences over the same DNN.

But when it comes down to it, the question becomes what exactly is the best way to measure AI “efficiency”? What factors must be kept in mind to fairly and objectively compare different models? The application of the models should at the least be the same,

certainly, but do the models need to use the same type of data as one another? It also should be pointed out that algorithmic differences provide inherent advantages over other types of algorithms if the metric is for example computational cost, as attention-based methods only conduct convolution operations on areas of greater "importance" for example.

With all the AI hype not reflecting the economic realities of AI usage in business and the ever growing costs of frontier models [65] one might argue that the simplest answer is *money*, the financial cost of the model. Quantifying the financial costs of an AI model is certainly a viable option, but the metrics can often be incomplete, with costs like data scientists and financial loss from appreciating prices on hardware assets often not being included. As AI is a relatively new field, there is no standards for this kind of measurement, especially when one considers the variables of cost of energy, the various financial angles that would need to be considered to quantify "financial cost", and averaging the costs of retraining as needed. It is also important to consider factors like inference time and latency, but when it comes to efficiency metrics, it is important to remember that they are not necessarily correlated with each other.

No *single* cost indicator is sufficient on their own, and incomplete reporting as a measure of efficiency can be misleading. For example, a model with low FLOPs may not be actually fast, given that FLOPs don't take into account data such as degree of parallelism (depth, recurrence, etc.) or hardware-related details such as the cost of a memory access. Likewise, the number of trainable parameters (aka the size of the model) is commonly used as a de-facto cost indicator in the NLP [66] community and previously

the vision [67] community. This can however be misleading when used as a standalone measurement of efficiency. Intuitively, a model can have very few trainable parameters and still be very slow, for instance when parameters are shared among many computational steps. Similarly, throughput and speed as a primary indicator of efficiency can also be problematic, since this tightly couples implementation details, hardware optimizations, and infrastructure details into the picture. Therefore, it might not present an "apples-to-apples" comparison of certain methods, or worse, across different *infrastructure* or *hardware*.

With all of that being said, FLOPs still provide a number of benefits for measuring computational efficiency, especially when comparing models with *similar architectures* trained on the *same hardware*. Floating-Point Operations Per Second (FLOPs) are a unit of measurement that quantifies the computing power of a computer or processor [68]. It measures the number of floating-point calculations that can be performed in one second. This is particularly important for training an AI model, as the calculation of theoretical FLOPS vs measured FLOPs can better represent the actual computational performance observed during real-world applications [69].

FLOPs in particular are considered a measurement in evaluating the true performance of computing systems. It can be used to identify bottlenecks and inefficiencies in software and hardware, guiding optimizations for better performance. It also ensures that computational resources are being utilized effectively, for determining hardware cost-efficiency. There are many factors that can affect FLOPs efficiency, in

particular hardware architecture being a major factor. The design of CPUs, GPUs, and accelerators can considerably impact performance for a model [70].

All of that being said, there are certainly some caveats with using FLOPs as a metric for efficiency. It should be pointed out that FLOPs as a measurement does not necessarily correlate well with the energy consumption [71], nor does reduced FLOPs necessarily directly translate to performance improvements. Even as a hardware-agnostic resource constraint, FLOPs cannot translate the runtime complexity for target edge systems as they are inherently hardware-independent as a metric. Additionally, the number of parameters is usually reported but is not directly proportional to the computational cost measured in FLOPs. For example, in CNNs, convolution operations dominate the computation: if d, w, and r represent the model's depth, width, and input resolution respectively, the FLOPs grow following the relation [72]:

$$FLOPs \propto d + w^2 + r^2 \#(1)$$

This is to say that FLOPs do not directly depend on the number of parameters of the model. The number of parameters affect the model's depth or width, however distributing the same number of parameters in different ways (using different layer shapes, filters, etc.) will *still* result in a different total number of FLOPs. Moreover, the input resolution is independent on the number of parameters directly, as input resolution can be increased without increasing the overall model's network size.

Other factors include the number of cores, clock speed, and parallel processing capabilities [73]. Additionally, the rate at which data can be read from or written to memory will also affect computational efficiency for the training of AI models. Higher bandwidth for memory generally leads to better performance. Avoiding overheating is

another major factor, as overheating can throttle performance if one has insufficient cooling solutions to maintain the high efficiency of the hardware. Managing power consumption through dynamic voltage and frequency scaling can optimize performance per watt. On the software side, efficient algorithms and optimized code can drastically improve performance, making better use of the available hardware architectures [74]. How the data is accessed and stored affects performance, as efficient data locality reduces latency and improves throughput. Lastly, optimized compilers and tuning compilation settings can result in more efficient executable code.

CHAPTER THREE

THE ESCAPE DATASET

3.1 The ESCAPE Dataset

The Experiments, Scenarios, Concept of Operations, and Prototype Engineering (ESCAPE) dataset [75] was released in 2019 by the Air Force Research Laboratory (AFRL) Information Directory for the purposes of enabling advanced multi-modal signature data-fusion research. The ESCAPE dataset combines a wide variety of different sensors, including Acoustic, EO (Full-Motion Video), Infrared (IR), P-RF, Radar, and Seismic data in a common set of scenarios for the application of advanced fusion. In each of the scenarios, there is one or more vehicle targets which attempt to "escape" detection, hence the name. The motivation for the ESCAPE dataset is to develop any model using multimodal data with an incentive to use more than just the EO sensors, especially when the EO sensor data is purposefully limited.

The dataset is based off previous research they had conducted, which investigated pre-detection (upstream) level sensor fusion of multi-modal signatures using simulated Infrared, Full Motion Video, P-RF data collected from moving emitters. Because of the test design for these scenarios, the ESCAPE dataset significantly enhances the complexity and opportunity of the scenario, by increasing the number of modalities utilized and other outdoor experimental irregularities. The data collection utilized commercial off the shelf sensor equipment, providing Acoustic, EO, IR, Radar, P-RF, and Seismic data. The target types for this dataset are ground-based vehicles.

Figure 16. STS site. [75]

The actual data collection was performed at the AFRL/RI Stockbridge Test Site (STS), in a rural part of central New York state near Stockbridge (43.031897º N, 75.651311º W) during the summer and fall of 2018. The STS is a 300 acre, partially

wooded site, formerly an Air Force radio frequency and antenna pattern measurement facility that is re-purposed for Small Unmanned Air Systems (SUAS), optical, expanded RF, cyber and Intelligence, Surveillance, and Reconnaissance (ISR) research. The choice of this location is especially beneficial for providing targets different sources of obscurement from line-of-sight viewpoints, as they pass through obstacles.

In order to meet the objectives initially set for the dataset, the team exploited the unique features of the area to provide simultaneous line of site of the target scene from the majority of the sensors participating in the tests. The terrain is primarily flat grass fields, dirt, roads, asphalt/concrete roads, sparse-tree hedgerows, and a large commercial steel building (the Butler Building, referred to in the scenarios as the garage), which provides a tunnel like feature for vehicles to enter/exit from the south and enter/exit from the North. When one or more vehicles enter, and another leaves, the AI must make the determination as to whether or not the vehicle target that leaves the garage is the original target or if it was an entirely different target. The design of these scenarios enables performance evaluation of target representation approaches, as determining if the vehicle target has in fact “escaped” detection becomes integral to the model’s performance.

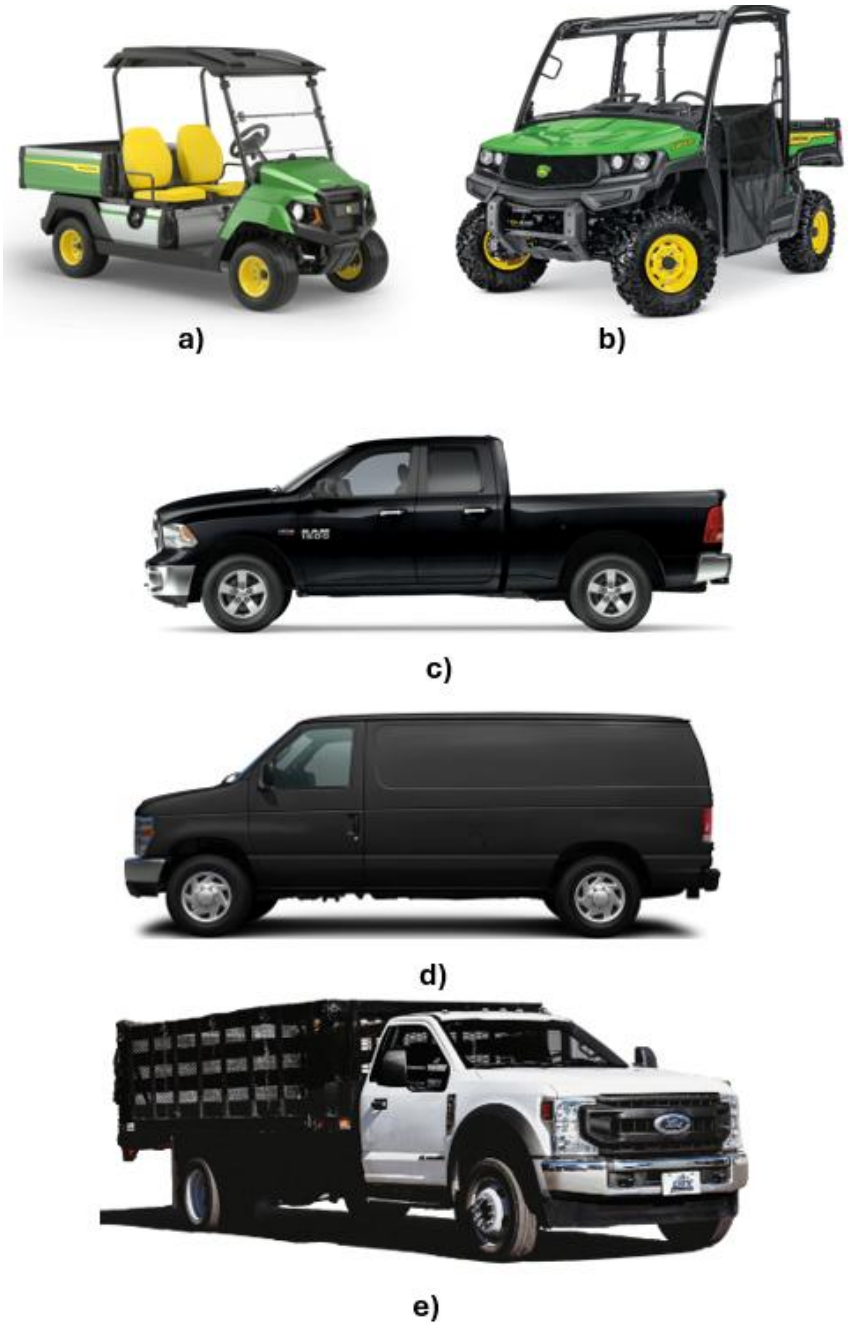

Figure 17. Ground Vehicle Targets for the ESCAPE dataset, a.) gas motor Gator utility vehicle, b.) diesel motor Gator utility vehicle, c.) pickup truck, d.) panel van, and e.) stake rack truck.

The ground vehicle targets (Figure 17) for the ESCAPE dataset collection consist of utility vehicles (John Deer Gator), a pickup truck, a panel van, and a stake rack truck. There is also a variation of a diesel and two gas Gators. All three of the Gators are 4-seat

versions, which ensures the similar length and width, but the diesel gator presumably having a different acoustical and seismic signature due to the nature of diesel and gas engines. In any given ESCAPE dataset scenario, at least one or more moving vehicles is attempting to avoid detection and make a switch, and the theoretical target to detect and differentiate from will be among these different vehicle targets.

Each of these vehicle targets is also fitted with an RF emitter, with one emitter placed in two separate vehicle targets for all scenarios and one emitter used in the single target scenarios. Each of these emitters signal design simultaneously transmits 13 different transmissions separated in frequency. There are 11 typical communication module schemes used, including Binary Phase Shift Keyed (BPSK), Quadrature PSK (QPSK), Offset QPSK (OQPSK), 8 PSK, 16 Quadrature Amplitude Modulation (16-QAM), 64-QAM, Gaussian Minimum shift Keying (GMSK), 4 Amplitude Shift Keying (4ASK), Frequency Modulation (FM), Amplitude Modulation Single Sideband Suppressed Carrier (AM-SSB-SC) and Amplitude Modulation Double Sideband Suppressed Carrier (AM-DSB-SC). Additionally, two tones were added to make 13 total frequency channels.

Each emitter transmits the 13 frequency channels over a 4 MHz frequency band. One emitter would transmit over a 4 MHz band from 1240 MHz to 1244 MHz; referred to as the low band emitter. The second emitter, separated by 1 MHz from the first emitter, would transmit on a similar 4 MHz frequency band from 1245 MHz to 1249 MHz; referred to as the high band emitter. The total is 8 MHz worth of transmit signals, plus 1

MHz of guard band between them. It was designed to fit within the ESCAPE data collection receiver's bandwidth, particularly the ESCAPE payload, of 12.5 MHz.

While it would be *possible* to preprocess the RF data, looking for those specific signals is not a part of the preprocessing, tracking, and differentiation of different vehicle targets. Rather it should be clarified that it is a signal of opportunity which potentially aids in differentiating between different targets. As this system is passive in nature, the system fundamentally cannot be designed focus on those *exact signals*, but it can process the RF data further to provide better visibility of the existence of these features, for the model to then utilize. In any real-life scenario, target vehicles will likely have some level of RF component; for example, emitted by wireless communication systems, mobile phones, and maybe even a self-driving car system in the future, etc.

3.2 Scenario Overview

For the purposes of the scenarios and sensor data chosen from the ESCAPE data, the three that were used in this research are designated as Scenarios 1, 2, and 3, which correspond to the ESCAPE dataset's Scenarios 1, 2C, and 2D respectively. The total number of vehicle targets between the chosen three scenarios is ten, and each scenario deals with a different number of targets (for a total of 2, 3, and 5 targets respectively). The overall benefit for training using this dataset lies within the design of the scenarios which incentivizes use of available assets, while recording the multi-modality data on diverse circumstances.

With respect to the vehicle targets is that all the targets are designed to “escape” detection, by employing a number of different tactics. In the context of military/intelligence interests, these scenarios are based on practices and countermeasures used to avoid detection and tracking via satellites. The use of these tactics thereby incentivizes the fusion model to use multiple modalities as data inputs to determine if the actual target is in view or if the target has “escaped”. In any given scene, at least one vehicle was emitting various communication waveform standards.

The evasive scenarios come with the benefit of thereby challenging any model meant to differentiate between the potential vehicles when engaging in tracking. As well as any traditional detection-level fusion approaches, while potentially showing the benefit of pre-detection level approaches. Similar vehicles move in such a manner that even human agents might have difficulty in differentiating between the visually similar targets with just the EO visual information that they can process.

All three scenarios involve multiple vehicles entering and exiting a garage (a location inside which satellites on their own cannot track car targets), with multiple vehicles of similar make and build combined with dissimilar vehicles. The movements of these vehicles are also varied in order to present challenging discrimination opportunities that confuse tracking differentiation. The simpler single vehicle scenarios allow researchers to focus on, and perfect, new algorithmic approaches for developing multi-modality target representation by exploiting the pre-detection level (upstream) sensor data. Denser target scenarios contain conditions of closely spaced targets, opposing targets, passing targets, move-stop-move trajectories, and line of sight blockage.

Each of these scenarios are focused on the actual evasion for the vehicle target, and only contain 90 seconds worth of data. After being synchronized with respect to time, the EO (FMV), Passive RF, acoustic, seismic, etc. modalities can be used for training. All three scenarios focus on the garage, and an important question the fusion model must answer is whether or not the target that entered the building is the same target as the one that exits the building.

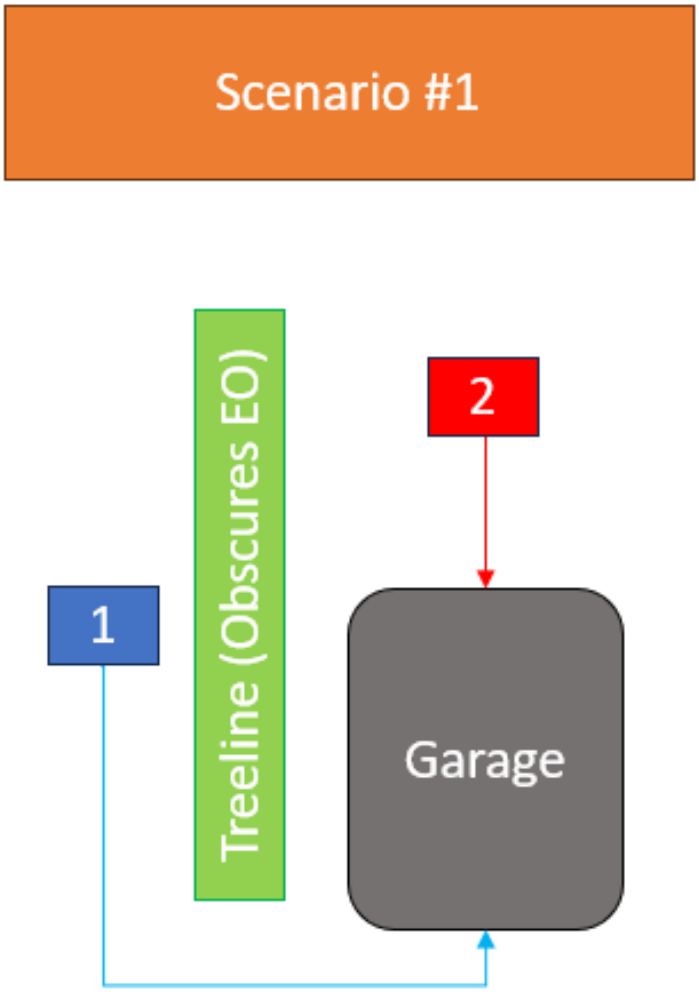

Figure 18: Scenario #1

Scenario #1 (Similar Vehicle Switch, Figure 18) has two possible vehicle targets, both of which are of the same build and color as each other. The scenario starts as vehicle #2 travels into the garage in plain view of the EO sensors. As this happens, vehicle #1 travels into the garage from behind the tree line. While doing so, from the EO sensor's point of view, vehicle #1 is "hidden" due to visual obscuration that prevents the model from detecting its movements most of the time. Once vehicle #2 enters the garage, vehicle #1 then exits the garage, and the objective of the first scenario is to successfully determine when the "switch" is made. If the model incorrectly identifies the vehicle exiting as #2, then that means the model has failed and the vehicle has successfully "evaded" detection.

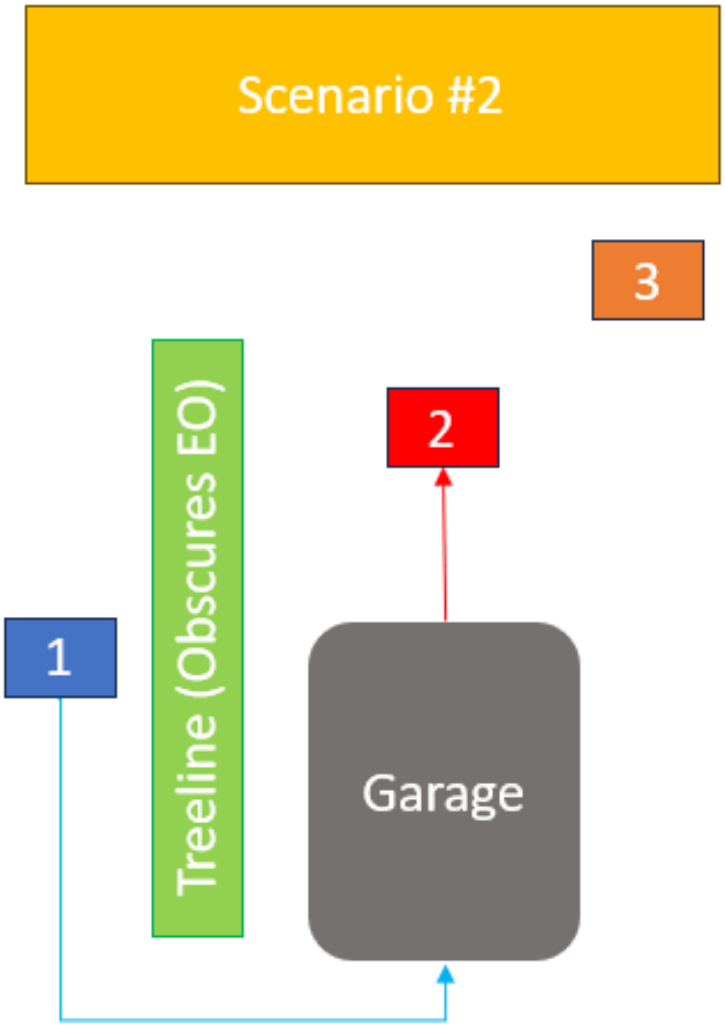

Figure 19: Scenario #2

Scenario #2 (Similar Vehicle Appearance, Figure 19) is nominally more complicated than scenario 1 by comparison. However, in this scenario, there are three total vehicles and essentially follow the same pattern as Scenario 1, but only two of the three are visually similar. The difference is that rather than vehicle #3, which is visible, or vehicle #1, which is not possible to watch the switch in the garage, is that vehicle #2 starts parked in the garage the entire time. This makes it appear to an outside viewer that vehicle #1 enters and exits when in fact it is hidden inside of the garage. If the model

determines vehicle #2 to be vehicle #1, then vehicle #1 will have thereby "escaped" detection. The EO input is insufficient on its own to make that determination, as the difference between similar vehicles, incentivizing the use of P-RF data.

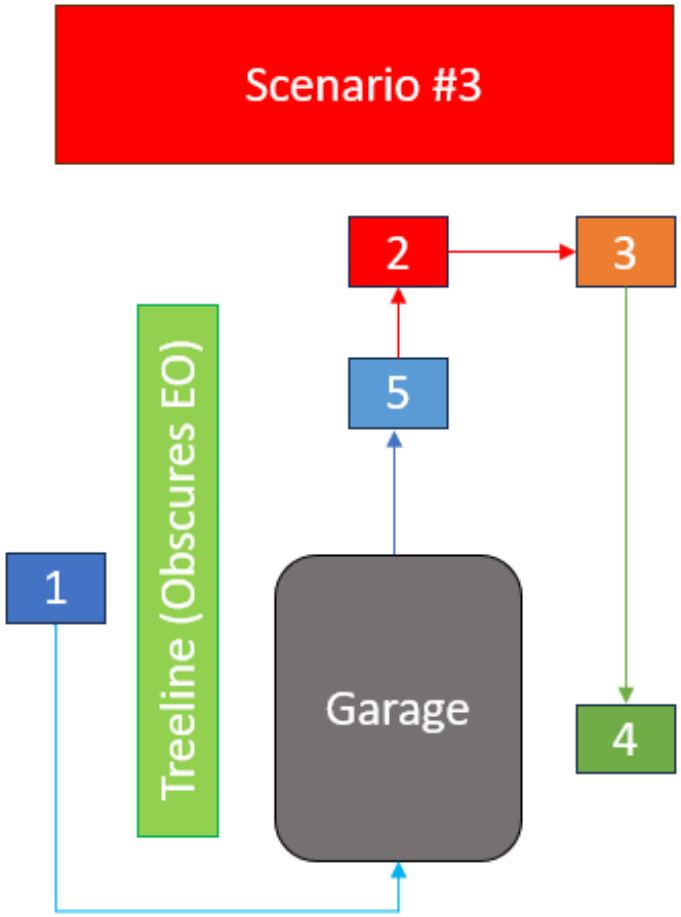

Figure 20: Scenario #3

Scenario #3 (Similar Vehicle Overtake, Figure 20) is the most complicated of the three and chosen due to the complexity of the five vehicle targets, all of which are traveling at different speeds and with different makes. Four of these vehicles arrive out of the front of the garage, while the fifth vehicle arriving from out of view, thereby making the tracking at the end of the video input, linearly speaking, extremely difficult to conduct

with only the EO input for that time frame. The variable speeds displayed by the five vehicle targets also presents an additional dimension of complexity with respect to tracking as the vehicles that are similar in design will overtake the other at different points within the scenario, making tracking a challenging process for Scenario 3 in general.

3.3 Sensors and Preprocessing

In order to collect the EO, full motion video (FMV), infrared (IR), P-RF, Radar, Seismic, and Acoustic data, sensors needed to placed across the testing site where the data was recorded. These sensors are placed on all manner of surfaces, structures, and drones, with small unmanned aircraft systems (SUAS), ground tripods, and ground contact sensors being predominantly used. It should be noted that the locations of some of these sensors change between scenarios, which provides further difficulties when using the Acoustic and Seismic data.

Radar data is collected using an Akela RF vector signal generator, two Akela tapered horns (one transmitter, one receiver) and a control laptop. Besides the sensors the Air Force used, there were also some that were provided by Michigan Tech Research Institute (MTRI). It should be noted that the sensors are not synchronized via GPS, and adjustments need to be made beforehand to synchronize the difference data sources with respect to time. The key information needed for understanding the vehicles, scenarios,

sensors, and overview of the dataset's structure is in three PowerPoints that come with the dataset, which is really important for the ground truth labels of each scenario.

There were also three passive P-RF receivers, independent of the payload, which synchronously measure the RF signals using an Ettus N210 software defined radio receiver. The nodes operated remotely via a standard wireless network, with data being recorded in real-time to the hard drive of a Macbook Pro and an AH Systems SAS-510-4 Log Periodic Antenna. Each node was attached to a 40ft tall portable tower, with the antenna being mounted directly below an EO FMV camera. A DJI M600 SUAS vertical take off and landing aircraft was used to carry a payload of EO FMV, IR FMV, and passive RF sensors. Owing to difficulties in synchronizing the data from the SUAS, that P-RF data is not utilized in the model, with only the tower data being used. The EO FMV used an AVT Prosilica GX3300 GigE camera (resolution 3296 x 2472, 12 bit color), which broadcasted the data to a SATA hard drive and control laptop.

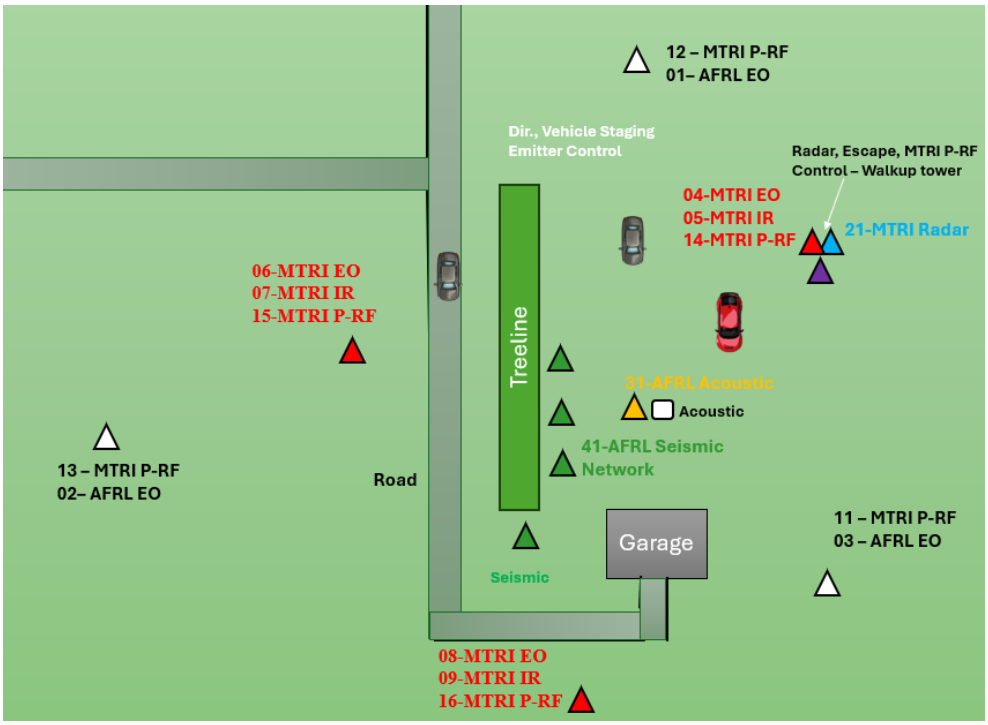

Figure 21. Scenario Sensor Setup

The non-traditional Intelligence Surveillance, and Reconnaissance (ISR) employed in the dataset are above-ground acoustic and seismic modalities (Figure 21). The acoustic sensors consist of two arrays that are 2-3ft above the ground, with arrays placed orthogonal to each other. The acoustic sensors consist of 16 Bruel & Kjaer 4952 outdoor microphones, 2 sound devices 788T 8 channel recorders, pre-amplifiers, and a 24 bit A/D. The seismic data is recorded via a seismometer using a standard commercial off the shelf (COTS) sensor, a “Raspberry Shake”, which consists of a Raspberry Pi 3 controlled single channel electronically extended 4.5 Hz geophone with a 24 bit digitizer sampled at 50 HZ with the data presented in miniSEED format. The “Raspberry Shake” is traditionally used for earthquake recordings up to 300 miles from the epicenter, but for the purposes of the ESCAPE dataset only ten units were deployed around the testing area, with each sensor being placed in contact with the ground, located on a grass field, asphalt

driveway, and the concrete floor of the Butler building. It should be noted that the exact locations of these seismic sensors shift slightly between scenarios.

Once synchronization of sensors with respect to time is completed, each of the modalities come with their own unique challenges and caveats to consider. Starting with the EO data, as visual data is the most intuitive for human beings to understand, and coincidentally, the easiest way to confirm if the vehicle target is detected. This is advantageous, because obviously it is easier to track a vehicle if it can be seen, however it is also paradoxically *disadvantageous* to the overall goal of achieving sensor fusion, because if the neural network model can simply rely on just the EO data, the logical thing for the model to do is to just *not use the less reliable data*. Because the objective is *data fusion*, it therefore becomes important to reduce the amount of EO data available to the model in training, thereby making it impossible to perform well *solely* by using the EO data, and thereby *incentivizing* P-RF and other modality usage.

Besides this, there is also an amusing "prank" (as shown in Figure 22) played on researchers who use the EO modality in this dataset (after finding out that the dataset was not synchronized via GPS and will need to *manually* be synchronized) that depends on which EO sources are used. While the dataset is *theoretically* supposed to be focused on those aforementioned five types of vehicles (Figure 17), from *certain* camera angles from the EO data, an algorithm exposed to this dataset will also inevitably notice the *rows of parked vehicles* of the researchers and other staff involved in the creation of the dataset. This therefore necessitates a preprocessing method to adjust for this "prank", so that the model does not treat the parked vehicles as *legitimate targets*. Obviously, in a real-life

application, this is an obstacle that the algorithm should be able to handle on its own. But for the purposes of achieving sensor fusion and explainability, the EO data's stake in the fusion process is of decidedly less interest. From an algorithmic perspective, the focus on vehicle target detection is precipitated by the fact that these vehicle targets are physically moving, and so, the method chosen for preprocessing is to detect the moving targets.

Figure 22: DOF Preprocessing

While not optimal for real-time processing, the focus of this research was sensor fusion, and Dense Optical Flow (DOF) more importantly removed the parked vehicles from the data. Optical Flow looks for the "motion" of objects between consecutive frames of a sequence, caused by either the relative "movement" between the object and the camera. This displacement, and lack thereof, is used by the algorithm in order to identify what portions of the time series data are worth processing, with minimum movement not being highlighted by the algorithm at all. A Taylor Series Approximation of the system represented by the change in pixels in order to derive the optical flow equation and produce the image gradients along the horizontal, vertical, and time axis for the "motion estimation" of the object in question.

While Dense Optical Flow is a more computationally expensive method compared to Sparse Optical Flow, the use of the Farneback method [76] provided a considerably more visually effective transformation to the data. The Farneback method uses a two-frame motion estimation algorithm that uses polynomial expansion. The neighborhood of each image pixel is approximated by the polynomial, creating a quadratic polynomial that is used for producing a local signal model represented on a local coordinate system.

By using the computed neighborhood polynomials on the two subsequent images, it becomes possible to directly derive the displacement between the two, thereby measuring the "flow" of each image pixel. The changes in signal with regards to the weight and displacement are covered in the equations below respectively, where A is a symmetric 2x2 matrix of unknowns which are used to capture the signal (b), x is the image intensity, and w is weight function for the points in the neighborhood produced in the first equation.

$$\sum_{\Delta x \, \epsilon \, I} \quad w(\Delta x)|A(x+\Delta x)d(x) - \Delta b(x+\Delta x)|||^2 \ \#(2) \qquad d(x) =$$

$$(\Sigma \, w \, A^T A)^{-1} \Sigma \, w A^T \Delta b \ \#(3)$$

The second source of EO data used in this dataset is thresholding via Otsu's method. When previously tested [77] and compared with another form of preprocessing, thresholding via Otsu's method [78], it still came in second by all metrics (F1 score, Local and Global Impact) to Dense Optical Flow. The use of Otsu's method essentially calculates a threshold between which "class" pixel belongs to, either objects of interest or

not. It accomplishes this by minimizing intra-class variance, which is defined as a weighted sum of variances of the two aforementioned classes:

$$\sigma_{\omega}^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t) \#(4)$$

Where weights ω_0 and ω_1 are the probabilities of the two classes separated by a threshold t, and σ_0^2 and σ_1^2 are the variances of the two classes. The class probability $\omega_{\{0,1\}}(t)$ is calculated from the L bins of the histogram. By adding the variance of the background pixels to the variance of the foreground pixels for all possible thresholds, it becomes possible to select one for which the sum of the variances is the smallest. More specifically, minimizing the intra-class variance is equivalent to maximizing the inter-class variance:

$$\sigma_b^2(t) = \sigma^2 - \sigma_{\omega}^2(t) \#(5)$$
$$\sigma_b^2(t) = \omega_0(t)(\mu_0 - \mu_T)^2 + \omega_1(t)(\mu_0 - \mu_T)^2 \#(6)$$
$$\sigma_b^2(t) = \omega_0(t)\omega_1(t)[\mu_0 - \mu_1]^2 \#(7)$$

The variance is expressed in terms of class probabilities ω and class means μ. The class means $\mu_0(t), \mu_1(t), \mu_T(t)$ are calculated as:

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} ip(i)}{\omega_0(t)} \#(8)$$
$$\mu_1(t) = \frac{\sum_{i=t}^{L-1} ip(i)}{\omega_1(t)} \#(9)$$
$$\mu_T(t) = \sum_{i=0}^{L-1} ip(i) \#(10)$$

Where:

$$\omega_0\mu_0 + \omega_1\mu_1 = \mu_T \#(11)$$
$$\omega_0 + \omega_1 = 1 \#(12)$$

And the class probabilities and class means can be computed iteratively. Otsu's method generally performs well on data with a bimodal histogram, with a deep valley in between. While the modality helps highlight vehicles in a same way that DOF does, the problem is that it does not remove the inactive targets (the "prank") from the area, and thus on its own is only able to score a 0.73 F1 score compared to the maximum EO only input that DOF achieves as a 0.88 F1 score.

For the purposes of P-RF preprocessing in this paper, the Wigner-Ville Distribution (WVD), Continuous Wavelet Transform (CWT), Constant-Q Gabor Transforms (CQB), and Short-Time Fourier Transforms (STFT) were used. Our previous research [79] with these methods showed a considerable improvement over the previously used histograms. The Wigner-Ville distribution function was first proposed by Wigner [80] in 1932 in quantum mechanics and Ville [81] applied it to signal analysis in 1948. The Wigner-Ville distribution function takes time series x[t] and its non-stationary auto-covariance function is given by:

$$C_x(t_1, t_2) = \langle (x[t_1] - \mu\,[t_1])(x[t_2] - \mu\,[t_2])^* \rangle \ \#(13)$$

$$W_x(t, f) = \int_{-\infty}^{\infty} C_x\left(t + \frac{\tau}{2}, t - \frac{\tau}{2}\right) e^{-2\pi i \tau f} d\tau \ \#(14)$$

Where $\mu(t)$ is the mean (and any or may not be a function of time), taking the average over all possible realizations of the process. The Wigner function is then given by first expressing the autocorrelation function in terms of both the average time (t) and the time lag (τ), then applying the Fourier Transform to the lag. This is implemented in python using the Time-Frequency Toolbox with SciPy and matplotlib. For the Continuous Wavelet Transform, the function provides an overcomplete representation of

the signal by letting the translation and scale parameter of the wavelets continuously vary, as given by:

$$X_w(a,b) = \frac{1}{|a|^{\frac{1}{2}}} \int_{-\infty}^{\infty} x(t)\underline{\psi}\left(\frac{t-b}{a}\right) dt \ \#(15)$$

Where $\psi(t)$ is a continuous function in both the time and frequency domain, referred to as the "mother wavelet" and the bar on the ψ represents the operation of complex conjugate. The function was created in 1984 [82] by Alex Grossmann and Jean Morlet. The "mother wavelet" provides the source function to generate "daughter wavelets", aka the translated and scaled versions of the "mother wavelet". This is implemented using the PyWavelets library in python. The Constant-Q Gabor Transform is designed to transform a data series into the frequency domain, and essentially functions as a series of filters logarithmically spaced in frequency, with the kth filter having a spectral width equal to a multiple of the previous filter's width:

$$\delta f_k = 2^{\frac{1}{n}} * \ \delta f_{k-1} = \left(2^{\frac{1}{n}}\right)^k * \delta f_{min} \ \#(16)$$

Where δf_k is the bandwidth of the kth filter, f_{min} is the central frequency of the lowest filter, and n is the number of filters per octave. The transform was initially proposed in 1991 by Judith Brown [83] but has had multiple contributions for its current state. Similar to the acoustic data preprocessing, the Constant-Q Gabor transform is executed using python's Librosa. The last of the P-RF preprocessing methods used is Short-Time Fourier Transforms, proposed by Dennis Gabor [84] in 1946, a Fourier-related transform that determines the sinusoidal frequency and phase content of local sections of a signal as it changes over time. The function is multiplied by a window function (which is nonzero for a short period of time) that is then put through a Fourier

transform (one-dimensional function) of the resulting signal is taken. The window is then slid along the time axis τ until the end, resulting in a two-dimensional representation of the signal, represented by:

$$STFT\ \{x(t)\}(\tau, \omega)\ \equiv\ X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-iwt}\,dt\ \#(17)$$

Where $w(\tau)$ is the window function (typically Hann or Gaussian window) centered around zero, frequency ω, and x(t) is the signal to be transformed. $X(\tau, \omega)$ is the Fourier Transform of complex function x(t)w(t- τ), which represents the phase and magnitude of a signal over time and frequency. Typically phase unwrapping is employed along with either or both the time axis τ and the frequency axis ω. This suppresses any jump discontinuity of the phase result of the STFT. In practice, the procedure for computing STFTs divides the longer time signal into shorter segments of equal length, and then compute the Fourier transform separately on each segment.

For the purposes of preprocessing the Acoustic data, python's Librosa was utilized to implement Mel Spectrogram (MS). Originally, the data was going to also include Fast Fourier Transform Spectrograms and Power Spectrum Analysis via Librosa, but owing to poor results with the acoustic data, only the Mel Spectrograms were kept for the testing. The Mel Scale was originally proposed back in 1937 [85] as a unit of pitch such that equal distances in pitch sounded equally distant to the listener. Because of human evolution, humans do not perceive frequences on a linear scale, rather humans are better at detecting differences in lower frequencies as opposed to higher frequences. The Mel Spectrograms are relatively easy to use, requiring the following four basic steps to implement:

1. First breaking the audio signal into short frames.
2. The time signal must then be converted to the frequency domain using a Fourier Transform.
3. The frequencies are then converted to the Mel Scale, creating what is called the Mel Filter Bank.
4. After the Mel Filter Bank is completed, the Mel values are then plotted over time, generating the Mel Spectrogram.

By using the converted segments that are re-arranged on a time-frequency grid, the process provides rich data representing multiple dimensions from the original audio, turning the acoustic data into visual representation that is then read by the model for training. This provides an intuitively readable data source that both machines and human experts are able to understand. The Mel Spectrogram can essentially be thought of as a matrix of floating point numbers that depict frequency against time, such that when visualized the "shapes" of the audio signal are clear to see.

Lastly, for the purposes of preprocessing the Seismic data, Cepstrum analysis was used. Originally, methods such as Geometric Spreading function, DIP Filtering, and Deconvolution were going to be used. However, after initial testing, those three methods of preprocessing were similarly dropped owing to poor performance with the data, even after applying seismic denoising via Pyseistr [86]. Cepstrum analysis was first published in 1963 [87] and essentially is the result of computing the inverse Fourier transform of a logarithm of the estimated signal spectrum. This is represented as:

$$C_p = |F^{-1}\{log\ (|F\{f(t)\}|^2)\}\ |^2 \#(18)$$

Where f(t) is the signal as a function of time, C is the cepstrum, $\mathcal{F}$ is the Fourier Transform, F^{-1} is the inverse Fourier transform. In the context of its use in seismic data, it has been observed to have an additive periodic component of the logarithmic power spectrum of seismic signals caused by the echo signal, which has been useful for applications [88] related to exploring the differences in natural earthquakes and artificial explosions.

CHAPTER FOUR

P-RF FEATURE EXTRACTION

4.1 Early P-RF Preprocessing Attempts

Since 2019, our group has conducted and published several papers regarding EO/P-RF fusion with the ESCAPE dataset. Initially, P-RF data proved to be a difficult modality to work with [89] [18], but through machine learning based approaches, became more viable for vehicle target detection and differentiation. Back when histogram data was the only form of preprocessed P-RF data used, the vast majority of the distributions created by the histograms were indiscernible to the human eye, appearing to be noise (as seen in figure 7 at 0:01). It was clear from the improved scores that the model had learned something from the P-RF data, but it was initially unclear as to what features the fusion model was relying upon.

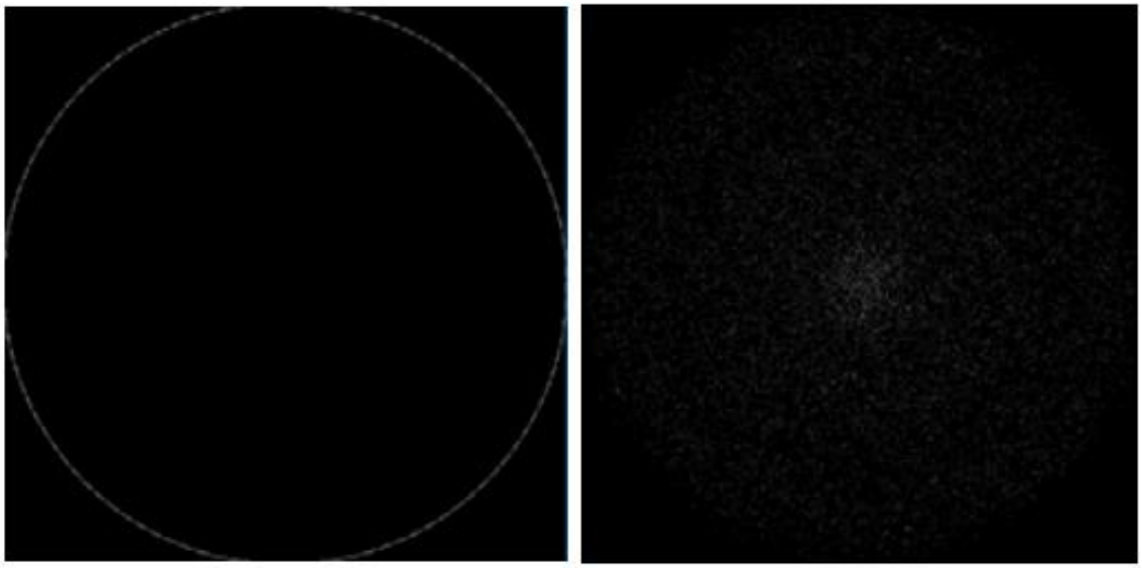

Figure 23. Comparison of I/Q Histogram collected for DIRSIG simulation 1 (left), simulation 9 (right)

This prompted further research, as up until that point our group had been using a DIRSIG (Digital Imaging and Remote Sensing Image Generation) model to experiment with vehicle detection using EO/P-RF fusion. The SIGINT provided by the P-RF data in the form of I/Q data histograms were quite clear (Figure 23) by comparison to the real data (Figure 25) collected by the ESCAPE dataset. There would certainly be some level of "noise" that appeared to correlate with multiple vehicle targets acting at the same time in the simulation, but nowhere near the degree of what the ESCAPE dataset's MTRI P-RF sensors were capturing.

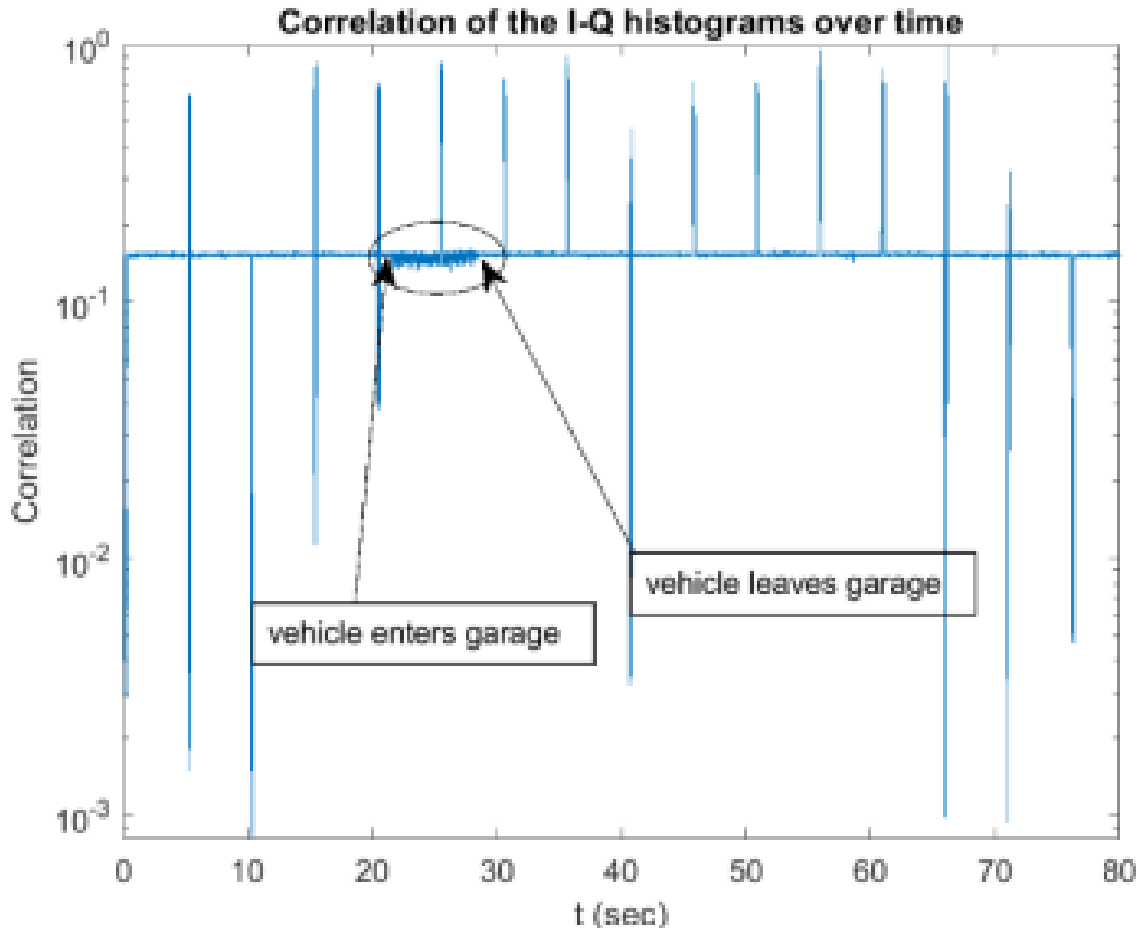

Figure 24. Correlation of P-RF histograms of d11 and d12 over time.

By taking a closer look at the histogram data from sensors d11 and d12 in scenario 1, it became noticeable at 0:21-0:28 that the histograms drastically change into a more discernable pattern (as seen in Figure 24). This occurs when the vehicle target enters and stays inside of the garage during that time period, as seen below (Figure 25).

d11 (left) and d12 (right) at 0:01 below:

d11 (left) and d12 (right) at 0:25 below:

Figure 25. P-RF Histograms (d11 and d12) of the ESCAPE dataset at 0:01 vs 0:25

As this indicated some level of correlation (as seen in Figure 24) between the histogram's changes and the model's improved performance, further research into the use of P-RF data was pursued [89]. These approaches included a comparison of single-target traditional classifiers [90]. These classifiers would include Naïve Bayes, Decision Tree,

K-Nearest Neighbors, and Nearest Centroid, and using XAI methods to further explore and compare the P-RF impact on the fusion model.

Initial attempts at using the raw I/Q data, for both the DIRSIG simulation and the ESCAPE dataset, were met with repeated failures. This was owing to a number of factors, as even after sufficient changes to the parameter inputs and model were made, and after the samples were synchronized with respect to time, the model had difficulty extracting useful features from the data, as seen by their dismally lower F1 scores. The usage of the histograms added two benefits, the first was easier fusion with respect to the EO data, and the second being the CNN's benefits from the histogram format being fed as a sensor modality.

4.2 XAI Research with P-RF Histogram Overlays

Once initial successes with P-RF fusion were made, the application of XAI methods, more specifically a Greedy Algorithm, were implemented in order to determine the P-RF impact. From the results of the Greedy Algorithm experiment, it was determined that the P-RF impact was relatively weak on a global scale, especially when compared to the EO data's impact. The noted improvements in F1 score and overall accuracy were promising when compared to non-fusion based EO models, which was obvious early on as with only one source of EO data, the model cannot determine when the "switch" is made in most scenarios, thereby driving down the overall score. But while improved scores are ideal to have, the question of how the model is utilizing the P-RF

data was clearly necessary to address. While it is difficult to determine how exactly a blackbox neural network utilizes P-RF data for differentiating between ten vehicles over three scenarios, it is possible to gain insights into the model's decision-making process using XAI methods.

To that end, the use of visualizations [91], GAN [92], Greedy Algorithm [93], and DiCE [77], were employed. Visualizations are a method of XAI which create a heatmap of neurons activated by an image input. DiCE is Diverse Counterfactual Explanations, an XAI method which generates feature attributions on the local and global level. The Greedy Algorithm used is ExplainX.ai, an advanced version of ProtoDash [94].

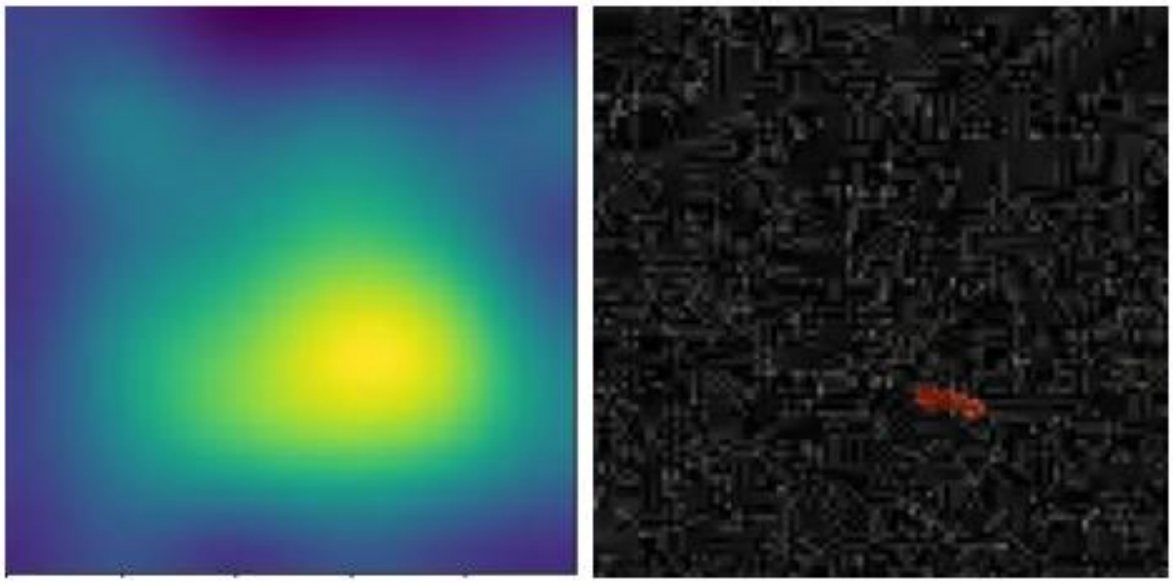

Figure 26: "EO Focused" Overlayed Heterogenous Input

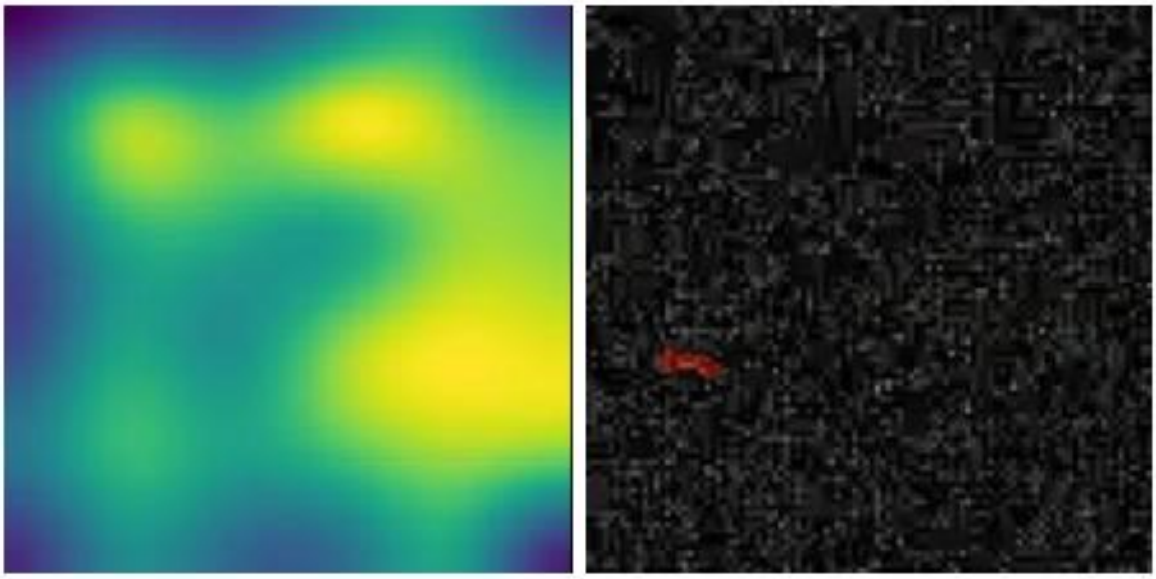

Figure 27: "P-RF Focused" Overlayed Heterogenous Input

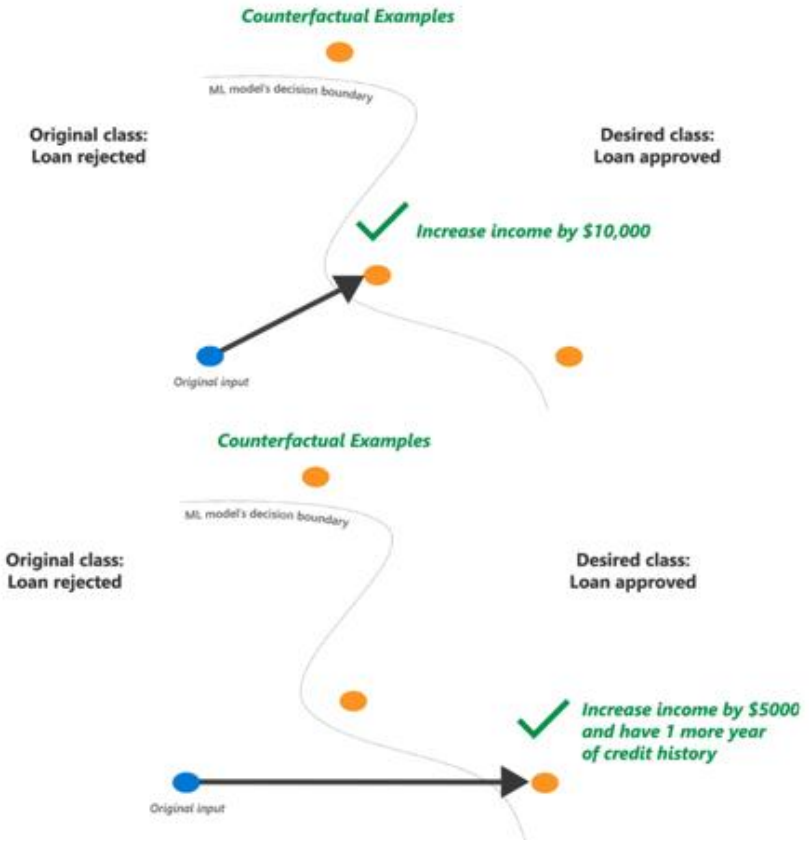

Figure 28. DiCE visual representation example. Source - github.com/interpretml/DiCE

During research using visualizations [91] on overlayed inputs that used both the EO and P-RF data, we had determined there to be three major trends in the visualizations produced for each scenario using the fused data. While visualizations and heatmaps are inherently more abstract than empirical, requiring human interpretation of the input (which can be argued to be objectively subjective), the quantity of similar categories in visualizations when plotted over time indicated that the three major visualization categories correlated to certain trends when taking into account the specific target.

These three categories were EO Focused (Figure 26), P-RF Focused (Figure 27), and Fusion Focused (Figure 29) visualizations. EO Focused visualizations were primarily focused on appearance of the vehicle target, P-RF Focused causing neuron activation focus on the overlayed P-RF data, and the Fusion Focused having both the vehicle target and P-RF data highlighted in the heatmaps. By plotting the frequency from which these different categories of visualizations occurred with respect to time in the scenario, we

were able to discern certain trends in the model's decision-making process based on which vehicle it was being tested on tracking.

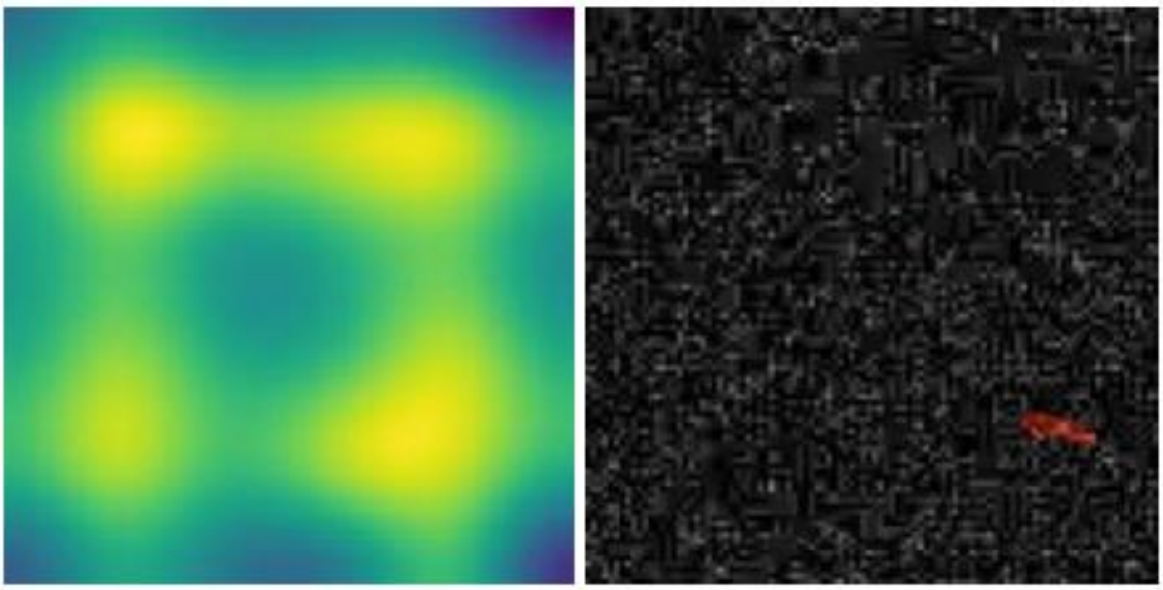

Figure 29: "Fusion Focused" Overlayed Heterogenous Input

The results indicated that, much like what the greedy algorithm research had confirmed previously, that the model primarily focused on the use of EO data whenever possible (EO Focused). This would often tend to occur with vehicle targets that were normally within view of the EO sensor most of the time. Unsurprisingly, because of the relatively unreliable P-RF histogram data, the P-RF data is only useful if the target isn't visible from the EO data (P-RF Focused), or if there are potential false positives (Fusion Focused). Such situations would occur when the vehicle target was behind the tree line or inside of the garage, and similarly when a similar vehicle target was visible but the actual vehicle target was not. These results were inferred from looking at the distribution of these categories with respect to time, and the knowledge of the events of those scenarios.

These distributions can be found in the source paper [91], and are discussed in further detail there. This research would steer us in the direction of the next natural step, which was to explore methods to potentially increase the impact of the P-RF modality impact on decision-making process of the fusion model.

4.3 Research with Incentivizing P-RF Data Usage

With both abstract (visualizations) and empirical (greedy algorithm, DiCE [95], LIME [96], SHAP, etc.) methods indicating that the P-RF data usage was relatively low, and even then, only if necessary, this raised the question of how to *better* incentivize model to use the P-RF data. The EO data was already reduced to a single sensor in order to incentivize the use of the P-RF data, but if non-DOF preprocessing was used the accuracy and performance of the model tended to drop even further. With that in mind, it becomes logical that if the P-RF data isn't performing well, the next logical step is to either forcibly incentivize the data's use in the neural network, or to see if it can be preprocessed in a better manner. This line of thinking lead to our research with Multi View Input – Deep Convolutional Generative Adversarial Network (MVI-DCGAN) [92], and seeing if the previously generated visualizations could be faithfully replicated.

It was our hope that the use of a DCGAN would provide realistic generations of the visualizations, better teaching the model to learn from the P-RF data based on what the previous model (having an F1 score of 0.9) had found to be useful features. During research with MVI-DCGAN however, we had mixed results attempting to train a more P-

RF focused model using adversarial training. The quality of the generated images was evaluated using Fréchet Inception Distance (FID).

$$d_F(\mu, \upsilon) \coloneqq \sqrt{inf_{\gamma \in \Gamma(\mu,\upsilon)} \int \quad ||x - y||^2 d\gamma(x, y)} \ \#(19)$$

Fréchet Inception Distance is calculated by using two probability distributions, μ, υ, over R^n having finite mean and variances, where $\Gamma(\mu, \upsilon)$ is the set of all measures on $R^n \times R^n$ with marginals μ and υ on the first and second factors respectively. The set $\Gamma(\mu, \upsilon)$ is also referred to as the set of all couplings of μ and υ. Rather than directly comparing the images pixel by pixel, the FID score instead compares the distribution of the generated images with the distribution of a set of real images (the ground truth).

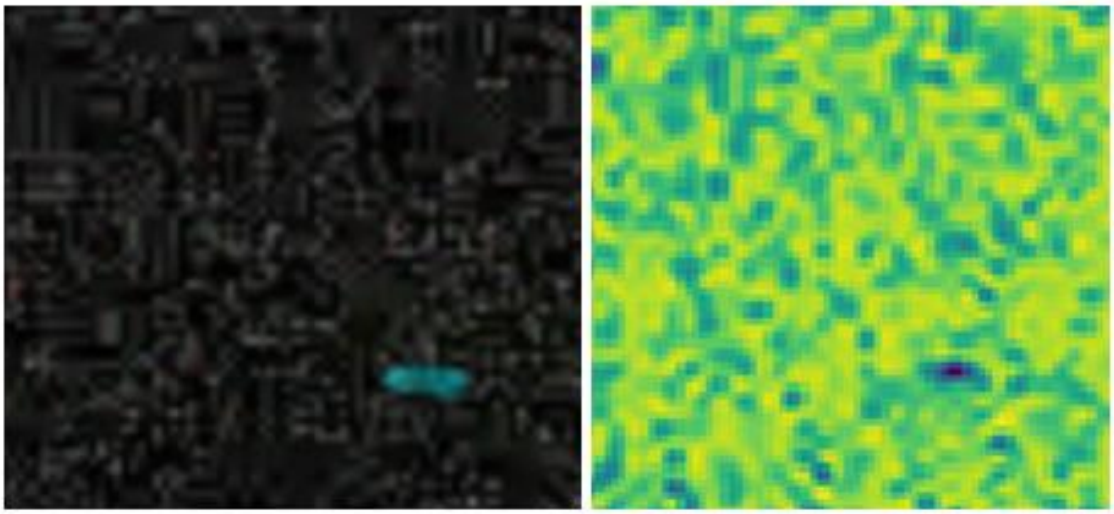

Figure 30: MVI-DC GAN comparison of overlay input and discriminator

While the training was successful in that the model was trained to focus exclusively on the P-RF data as opposed to the EO data (Figure 30), it was painfully clear that owing to what was clearly an insufficiently large enough amount of training data, the FID score was *decidedly lackluster*, to the point where it is visually obvious to even a

layman on how poor the generated images were. The best scoring generated FID visualizations that the model was able to produce looked closer to an AI generated prompt to recreate a grotesque version of Vincent van Goh's Starry Night, rather than any of the ground truth heatmap visualizations it was given as seen below in Figure 31.

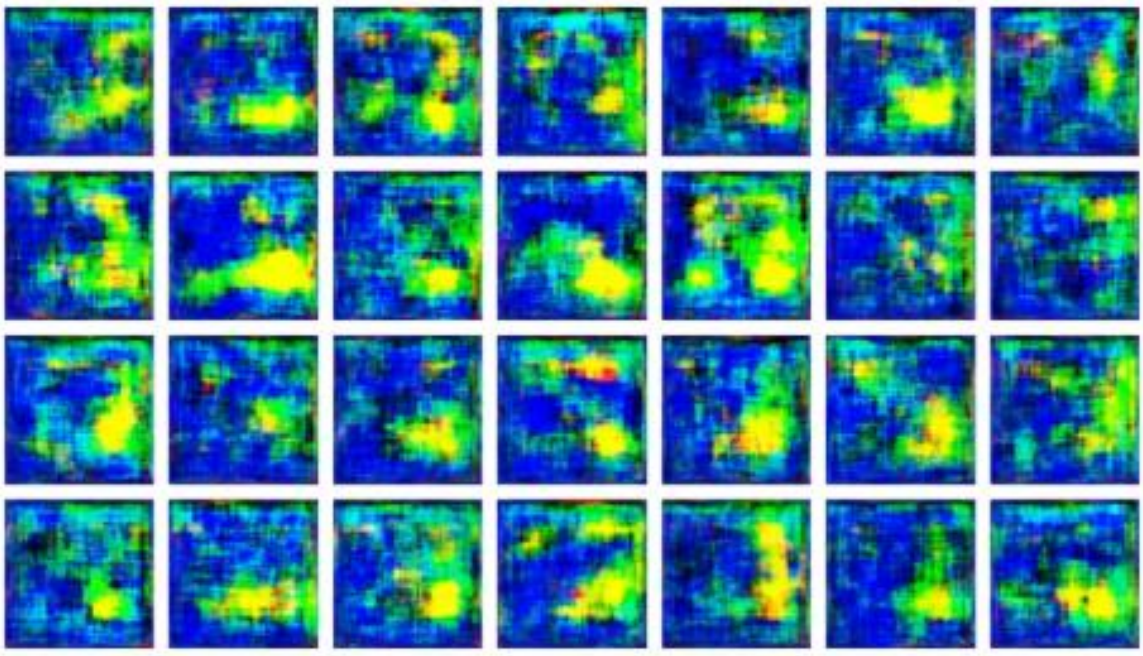

Figure 31: MVI-DCGAN generator (Scenario 2, Vehicle #2)

While we had previously attempted to utilize more traditional algorithms like canonical correlation analysis [90] [93] to incentivize P-RF usage, it was clear that the path forward with generative models was littered with roadblocks. From both a basic visual inspection and the more qualitative FID results of the model, it was clear that using generative models for further research would be difficult as we lacked both the sheer quantity of visualizations needed, and the quality with respect to the P-RF data's impact on the model's performance. With our attempts to synthesize high fidelity overlayed

visualizations clearly not being as successful as we would have liked, it was obvious that we had reached the limitations of what was possible using the P-RF histograms.

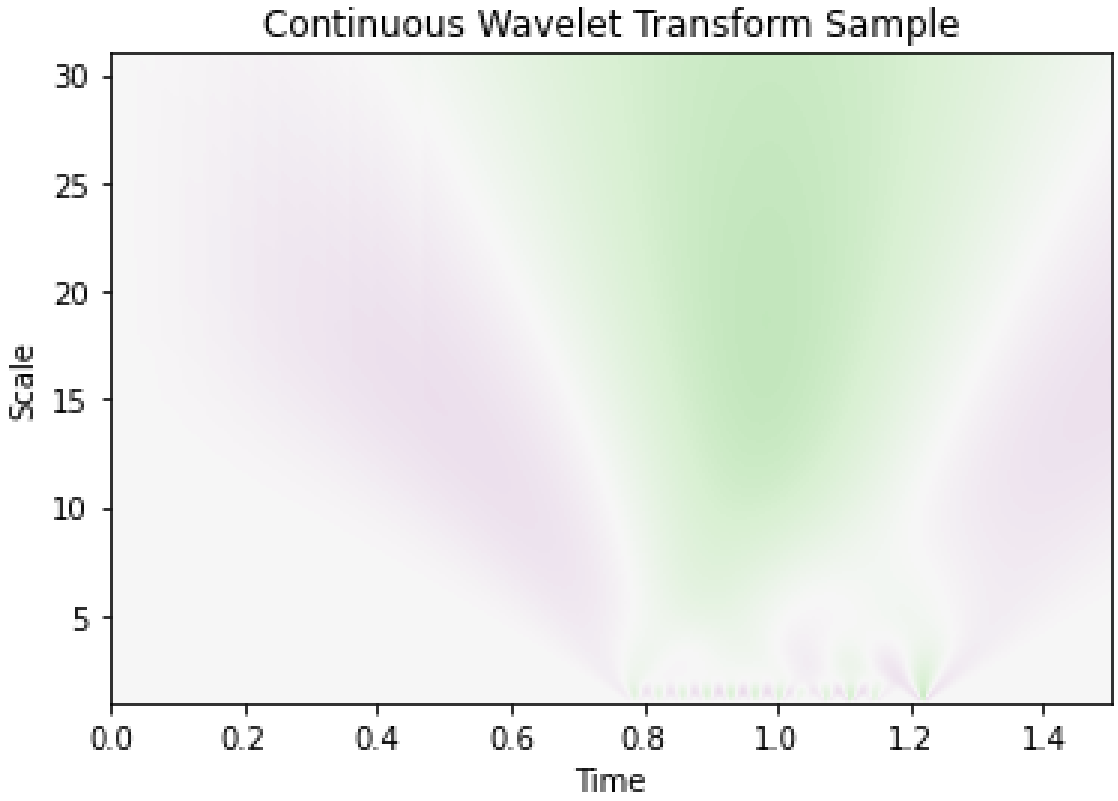

Figure 32. Continuous Wavelet Transform sample of Scenario 1

This therefore necessitated research into alternative methods of preprocessing the P-RF data. If research with the P-RF modality were to continue, histograms on their own would be drastically insufficient, as despite trying to use a myriad of different approaches, the quality of the data provided by the histograms, and the information the model could thereby exploit, was far too poor. Following the XAI research using generative models, we expanded on the methods of preprocessing the P-RF modality

[77], more recently showcasing our research with xLFER [79] regarding the use of Wigner-Ville Distribution (WVD), Continuous Wavelet Transform (CWT), Constant-Q Gabor Transforms (CQB), and Short-Time Fourier Transforms (STFT) as methods of preprocessing the P-RF data.

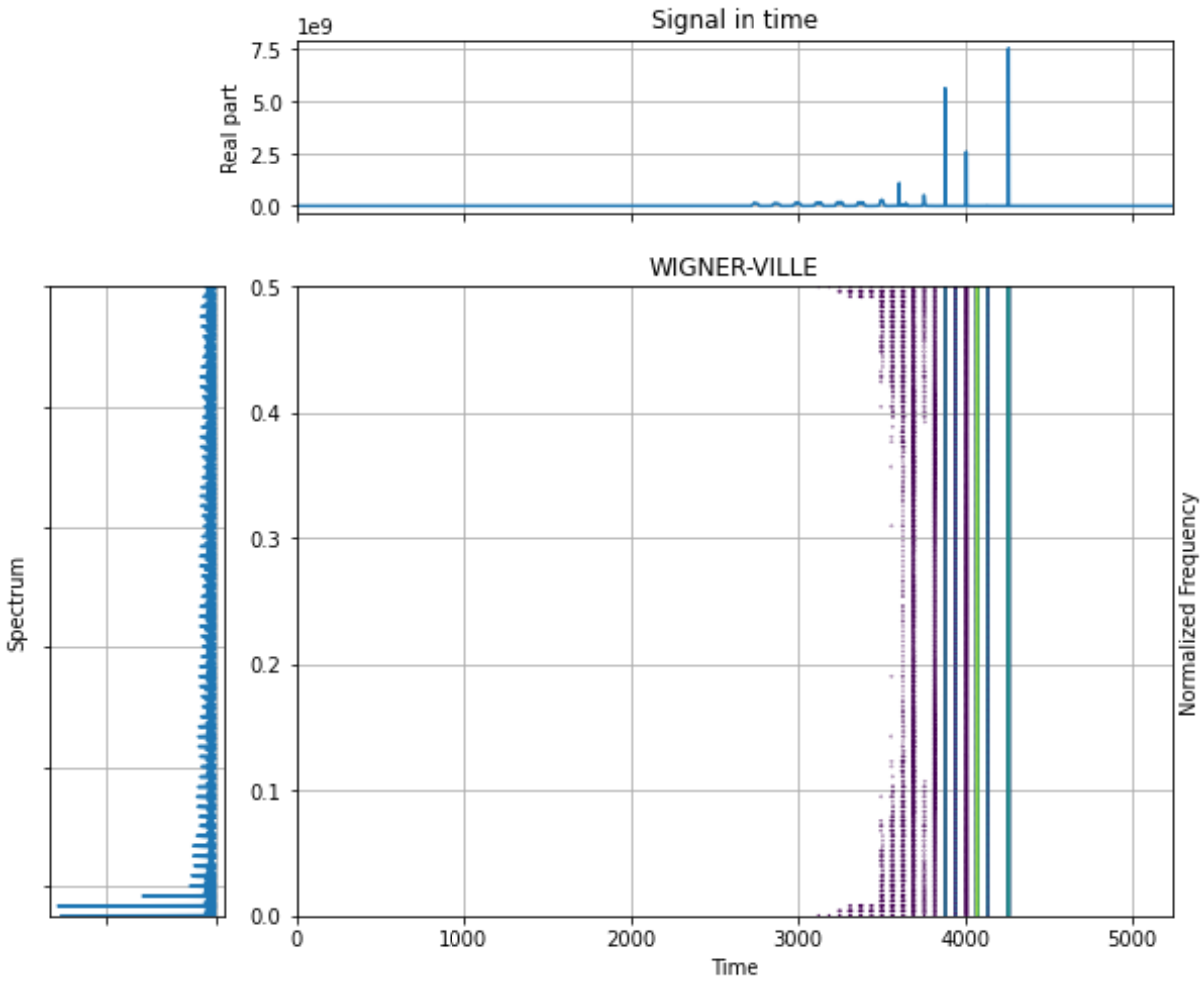

Figure 33. Wigner-Ville Distribution sample of Scenario 1

The derivation of these methods (CWT, WVD, CQB, STFT) were previously discussed in Chapter 3, Section 3, Sensors and Preprocessing. These sample were processed in python using the averaged bin files in python using Scipy and the Time-

Frequency analysis modules. These methods scored comparatively higher results in terms of both accuracy as well as local and global impact with the CWB and WVD modalities. It was based on those results using frequency domain and time-frequency analysis approaches, that warranted further investigation towards the use of CQB and STFT.

CHAPTER FIVE

HYBRID MODEL EVALUATION METRICS

5.1 Hybrid Model Approach

Chapter 5 extends the implementation of an explainable high level fusion model for tracking targets from multi-modal data, comparing the concept of hybrid models for counterfactual explanations. To the best of our knowledge, no research involving target detection and P-RF data has been conducted using counterfactual explanations. In this Chapter, the counterfactual explanations are generated using DiCE [95] (Diverse Counterfactual Explanations).

The hybrid model approach comes with the benefit of thereby challenging any model or algorithm meant to differentiate between potential targets when engaging in tracking, as similar targets being moved in a manner that even human users might have difficulty in differentiating between with only visual information. For the purposes of the scenarios and sensor data chosen, the three that were used in this hybrid model are designated as Scenarios 1, 2, and 3, which correspond to the dataset's Scenarios 1, 2C, and 2D respectively. The number of vehicle targets between the three scenarios totals 10, and each scenario deals with a different number of targets. The overall purpose of the dataset is that all the targets are designed to "evade" detection, by employing a number of different tactics that incentivize the fusion model to use different modality data input.

The three scenarios all involve multiple vehicles entering and exiting a garage, with multiple vehicles of similar make and build as well as more visually different vehicle targets. The model's evaluation metrics within these experiments are the accurate tracking and identification of potential targets (vehicle) using the different data sources provided, using the ESCAPE Dataset (Chapter 3) The data is preprocessed in a number of different ways, with EO being preprocessed with Dense Optical Flow (DOF), P-RF data being preprocessed into histograms. Within the scenarios provided, there are a number of vehicle targets which attempt to avoid detection, and therefore "escape", providing any model using the data an incentive to use more than just the EO sensors.

Hybrid Model Metrics for Counterfactual Explanations.

The usage of counterfactual explanations explores the hybrid model approach [77]. The hybrid model provides insights into how blackbox models can make use of P-RF data for the purposes of target detection. The hybrid model is able to combine the ability to fully utilize a black box model's learning capabilities with the explainability of a traditional model provides a decision tree for example, from an otherwise black box approach is extremely desirable, as it makes ensuring the comprehensibility and reliability of such a model easier. The hybrid model handles nonlinear data, while traditional methods might have issues with nonlinear data, if the decisions can be recreated using decision trees, it becomes extremely desirable for the purposes of explainability.

The hybrid use of black box systems can take many forms of implementing explainability, such as using decision tree pruning as backpropagation, generating decision trees from backpropagation, and rule extraction via decision tree induction. Being able to combine deep learning's ability to approximate complex relationships for problems with the explainability of decision trees allows for greater understanding of the hybrid model's decision-making process.

Early Level Fusion and Preprocessing

The two sources of data are preprocessed in a number of different ways, with EO being preprocessed with Dense Optical Flow (DOF) and thresholding, P-RF data being preprocessed into histograms and with FFT over the original I/Q data. While different sources of data enhancement has been tested in previous research, the usage of different interpretations of the same data is ideal for the purposes of gaining insights with the decision level fusion via decision trees. For the purposes of gaining explanations for the black box algorithms that will be used for the data, the usage of Counterfactual Explanations and saliency maps has been developed. The performance of the models will be kept separate from each other, with only the decisions they output being used to feed the decision level fusion model. The decision level fusion model from there will use the decisions made by the downstream models as training data.

Late Fusion Design

With the preprocessing completed and early models used, the next stage is of course the late-stage fusion. The Fusion Model takes four separate sources of decisions and inputs them into the decision level model. The late decision-level model will be a decision tree, in order to provide further insights using the generated decision tree's usage of the four inputs. The visualization of the tree after training provides different branches and nodes to explain the model's decision-making process and thereby provide greater explainability than relying on visualization methods.

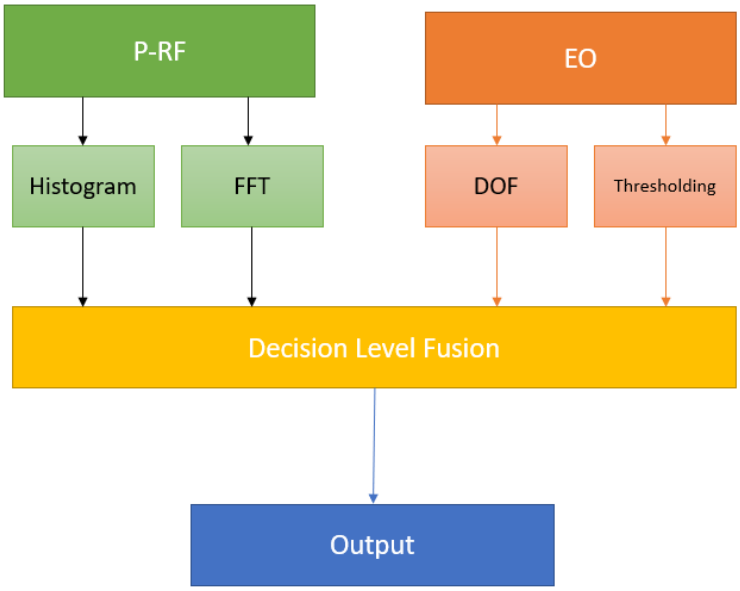

Figure 34. Late Fusion Model

Once the decision tree's results are completed, an analysis of the decisions made will be provided in the results section. In order to ensure the model's performance is on par with respect to other decision level fusion models that rely on more traditional means, comparison research will be conducted in parallel. Detailed in the next section below.

Comparison Research

With the usage of decision level fusion, naturally the comparison research for this data will not be limited to just the initial Decision Tree based fusion model. The comparison of the same input training data using Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and the Gaussian Naïve Bayes (GNB) will be used to compare the results. Evaluation will be conducted via F1 Score and Confusion Matrixes to gain a better insight into the performances of each model.

Results

Scenario 1 has two possible vehicle targets, both of which are of the same build and color as each other. For each of these scenarios, only one source of EO data was used, in order to maximize the need for the model to utilize the RF data rather than ignoring the P-RF input. The scenario starts as vehicle #2 travels into the garage in plain view of the EO sensors. As this happens, vehicle #2 travels into the garage from behind the tree line. While doing so, from the EO sensor's point of view, vehicle #1 is "hidden" due to visual obscuration that prevents the model from detecting its movements most of the time. Once vehicle #2 enters the garage, vehicle #1 then exits the garage, and the objective of the first scenario is to successfully determine when the "switch" is made. If the model incorrectly determines that the vehicle exiting as #2, then that means the model has failed and the vehicle has successfully "evaded" detection.

Figure 35. Scenario 1

Scenario 2 is nominally more complicated than scenario 1 by comparison. In this scenario, there are three total vehicles and essentially follows the same pattern as Scenario 1, but only two of the three look visually similar. The difference is that rather than vehicle #3, which is visible, or vehicle #1, which is not possible to obtain at the video angle chosen switching in the garage, is that vehicle #3 that was parked in the garage the entire time. This makes it appear that vehicle #1 enters and exits when in fact it is hidden inside of the garage, thus "escaping" detection successfully. The DOF EO input is not sufficient on its own to make that determination, as the difference between the two similar vehicles, as the source selected does not have access to all of the different EO sensors.

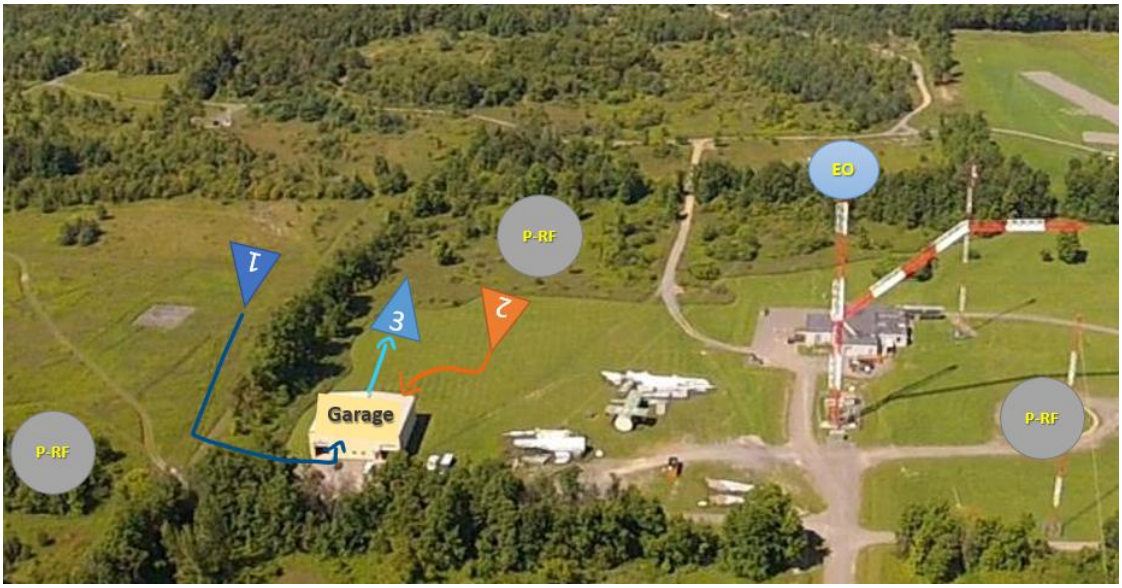

Figure 36. Scenario 2

Scenario 3 is the most complicated of the three and chosen due to the complexity of the five vehicle targets, all traveling at different speeds and with different makes. Four of these vehicles arrive out of the front of the garage, while the fifth vehicle arrives from out of view, thereby making the tracking at the end of the video input, linearly speaking, extremely difficult to conduct with only the EO input for that time frame. The variable speeds displayed by the five vehicle targets also presents an additional dimension of complexity with respect to tracking as the vehicles that are similar in design will overtake the other at different points within the scenario, making tracking a challenging process for Scenario 3.

Early Level Fusion and Preprocessing

The P-RF and EO sources of data are preprocessed in a number of different ways, with EO being preprocessed with Dense Optical Flow (DOF) and thresholding, P-RF data

being preprocessed into histograms and with FFT over the original I/Q data. While different sources of data enhancement have been tested in previous research [7] [19], the usage of different interpretations of the same data is ideal for the purposes of gaining insights with the decision level fusion via decision trees.

For the purposes of gaining explanations for the black box algorithms that will be used for the data, the usage of Counterfactual Explanations and saliency maps will be implemented. The performance of the models will be kept separate from each other, with only the decisions they output being used to feed the decision level fusion model. The decision level fusion model from there will use the decisions made by the downstream models as training data.

Late Fusion Design

With the preprocessing completed and early models used, the next stage is of course the late-stage fusion. The Fusion Model takes four separate sources of decisions and inputs them into the decision level model. The late decision-level model will be a decision tree, in order to provide further insights using the generated decision tree's usage of the four inputs. The visualization of the tree after training provides different branches and nodes to explain the model's decision-making process and thereby provide greater explainability than relying on visualization methods.

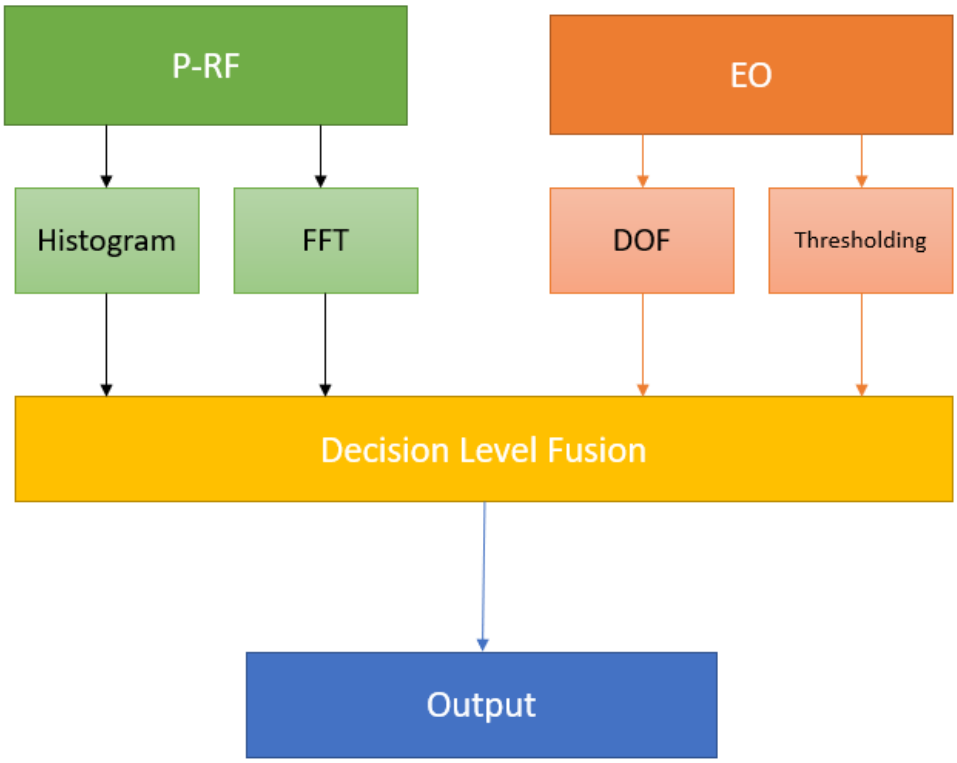

Figure 37. Late Fusion Model

Once the decision tree's results are completed, an analysis of the decisions made will be provided in the results section. In order to ensure the model's performance is on par with respect to other decision level fusion models that rely on more traditional means, comparison research will be conducted in parallel. Detailed in the next section below.

Comparison Research

With the usage of decision level fusion, naturally the comparison research for this data will not be limited to just the initial Decision Tree based fusion model. The comparison of the same input training data using Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and the Gaussian Naïve Bayes

(GNB) will be used to compare the results. Evaluation will be conducted via F1 Score and Confusion Matrixes to gain a better insight into the performances of each model.

Results

In this section, an overview of the experimentation results is provided below.

Early Level Fusion

While the standalone P-RF modalities are still being tested, the EO based modalities, as seen below in Table 1, are able to achieve relative success [77]. It is impossible for the EO modality on its own to achieve a perfect score, owing to how the ESCAPE dataset is designed. However, having a higher score is promising with respect to the information and its level of accuracy that it will provide to the decision level model.

TABLE I. EARLY LEVEL FUSION COMPARISON

Base Modality	Preprocessing	F1 Score
EO	Thresholding	0.73
EO	Dense Optical Flow	0.88
P-RF	Histogram	0.27
P-RF	CWT	0.41
P-RF	WVD	0.56

Comparison with Similar Models

The comparison of Decision tree algorithm with Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes (GNB) is used to showcase its performance with respect to other traditional methods, using the information from the EO (DOF), EO (Thresholding), PRF (Histograms), PRF (CWT), and PRF (WVD) models. While previous research has shown better results with earlier level fusion, the results overall were favorable for the decision level fusion approaches that were tested. The Decision tree's (DT) performance aside, the improvement that the two other PRF models provide can be seen in the increased F1 scores.

TABLE II. COMPARISON OF LATE DECISION LEVEL FUSION MODELS (INCOMPLETE)

Model:	F1 Score:
DT	0.92
LR	0.93
SVM	0.91
KNN	0.89
GNB	0.94

Explainable AI

The KNN model had the least impact, with the SVM model showing a decent performance, while the LR and GNB were able to outperform the Decision tree's performance entirely, as seen above in Table II. The Decision tree's decision making was unsurprisingly focused on the EO sources, with Thresholding predictably having the

lower impact of the two. The P-RF sources for CWT and WVD also predictably had a bigger impact than the decision-making process than the Histograms did, but otherwise is more or less in line with the weights the counterfactuals. While the DT model didn't fare as well as the LR and GNB models, the insights the tree weights provided were useful comparisons for the counterfactuals. While the baseline traditional models did not perform as well as the decision level fusion CNN, their performance was considerably improved from prior research with only P-RF histogram input.

In order to gain explanations from the models tested, counterfactual explanations are generated using DiCE (Diverse Counterfactual Explanations). Counterfactual explanations provide this information by showing feature-perturbed versions of the same sample based on different features, and thereby gaining insights into the model. This makes it possible to generate feature importance scores using a summary of the counterfactuals generated, determining its impact on both a local and global level [77].

TABLE III. COMPARISON OF EARLY LEVEL FUSION IMPACT

Modality	Local	Global
EO (Thresholding)	0.76	0.61
EO (DOF)	0.83	0.68
P-RF (Histogram)	0.16	0.32
P-RF (CWT)	0.37	0.41
P-RF (WVD)	0.53	0.57

The global importance scores per feature are estimated by aggregating the scores over individual inputs, while the local feature importance scores are computed for a given instance by summarizing a set of counterfactual examples around the point. As seen above in Table 3, the early level fusion feature importance scores largely favor the EO

modality. The results are largely in line with previous research with a greedy algorithm that assigned local and global feature impact scores, with the general trend being stronger EO impact than P-RF but with P-RF having a much larger impact on the global scale. Unlike the aforementioned research with ExplainX.ai, the use of counterfactuals relies on the model's decisions as opposed to approximating the original model via Greedy Algorithm. The use of counterfactuals also provides a more empirical explanation than methods such as saliency maps.

5.2 Model Performance

As the aim of this research is *not purely* just the optimization of the model's performance, there are a few *caveats* that need to be elaborated on, and a need to explain what metrics will be used to measure the results. A lot of the Sensor Fusion research our team has done has been aimed at understanding what *impact* P-RF data has on the model's capabilities, and as such *overall performance* isn't the main focus of this research. For the purposes of this research, the main focus is on Model Performance, Model Fusion, and Model Efficiency. Model Performance will be the model's actual ability to identify and differentiate vehicle targets based on the data provided. Model Fusion will relate to the model's ability to utilize the other non-EO modalities for sensor fusion. Lastly, Model Efficiency will relate to the processing costs associated with the training of the model.

Because prior research has already been done regarding sensor fusion using the EO/P-RF data with high F1 scores for the purposes of target detection and differentiation, the main focus is on improving the model's ability to utilize other modalities. Because the EO (DOF) data has been dominant in prior research, for just the EO data there is a counterintuitive focus on *reducing* rather than increasing its impact on the model, from both a local and global perspective. This only applies to the EO data because of the model's tendency towards using it over the other sources of data provided. For the P-RF, Acoustic, and Seismic data, the secondary objective remains to improve their impact in training while the model's general performance is measured and compared with different combinations of input.

With regards to the model's evaluation in terms of actual performance for the purposes of target detection and differentiation, and not efficiency or modality impact via XAI, the use of the F1 score metric is utilized. The use of F1 score gives a more comprehensive evaluation, especially in imbalanced datasets, by taking into account both false positives and false negatives. This is especially preferable in scenarios in which false positives and false negatives have different implications, such as in the medical field. Accuracy is a straightforward measure of correctness, whereas F1 score provides a more nuanced evaluation, especially in skewed datasets.

$$F_1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad \#(20)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \#(21)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \#(22)$$

The F1 score is calculated as the harmonic mean of the precision and recall, measuring positive predictive value and sensitivity for machine learning. It assesses the predictive ability of a model by examining its performance on each class individually. Precision is the measurement of type I error, false positives, while recall is the measurement of type II error, false negatives, respectively. The highest possible value of an F-score is 1.0, which indicates a perfect precision and recall, while the lowest possible value is a 0.0, if the precision and/or recall are 0.

5.3 Model Fusion

Besides the model's performance with respect to its task, it is important to measure or better understand how the input data is being utilized. For that purpose, by using SHapeley Additive exPlanations (SHAP) [97], it becomes possible to infer insights on how each modality is used on a global and local level. Local SHAP values can be calculated for each individual prediction to understand factors that contribute to that specific prediction. Globally, these SHAP values can be aggregated across multiple instances to understand the model's overall behavior and identify the most important features that positively or negatively influence its predictions. The theoretical SHAP values generated are defined by iterating through all possible coalitions, scaling exponentially with the number of players. The resulting SHAP values that are generated provide a strong estimation of relative usage and importance for the multiple sensors being provided to the neural network.

Our previous research using such methods had indicated a general trend in reducing the impact of the EO data, which is why while performance, both in terms of computational efficiency and F1 Score, will be an important, a soft objective for the fusion models is to ideally reduce overall reliance on the EO data. Ideally, with less reliance on EO data, the more the acoustic, seismic, and P-RF data will be used by the fusion model. This is not always the case, but is a general trend that has been observed in prior research exploring the local and global impact of modalities for the fusion model. This is largely dependent on scenario and target, but other factors the data after pre-processing may yet provide more opportunities for the other modalities to shine.

The dominance of the EO data for this dataset is that despite being restricted to only one EO source instead of all three, even the obscured EO sensor data provides the more reliable input for vehicle detection from the model's point of view. The value of the P-RF, acoustic, and seismic data increases when the target is either obscured or has a "double" in view of the EO source. While the P-RF, acoustic, and seismic data have been through preprocessing, it's painfully clear that further work will be needed to make either of those three modalities capable of detecting and differentiating between multiple vehicle targets. Some proposed ideas for improving their performance are covered in Section 7.2.

With regards to the use of P-RF data, the removal of the Histogram pre-processing method was chosen owing to its consistently poor performance [77] [79] across many experiments. After all, it was this relatively poorer performance in the incentivizing P-RF fusion research that prompted the research into other preprocessing

methods in the first place. Owing to having multiple sources of P-RF data that score a higher performance in F1 score and for local and global impact, keeping the Histogram data as a source of data would not contribute enough to justify using it. Therefore, for this reason, the P-RF histogram data is not used in the experimentation.

With regards to the use of Acoustic data, the exclusion of the Fast Fourier Transform (FFT) Spectrograms and Power Spectrum Analysis (PSA) were chosen owing to primarily the modality local and global impact for the model. When initially training what would have been fusion Model #1-E through Model #1-H, the F1 score would vary drastically, warranting further investigation into what was going on in the training for both modalities. When comparing the local and global impact after completing different training epochs, the impact scores would *actually decrease* as more training was conducted.

50 Epochs:

Model #1-E:		Local:	Global:	F1 Score:
EO	DOF	0.59	0.63	0.64
P-RF	CWT	0.27	0.41	
Acoustic	FFT	0.03	0.05	
Seismic	GS	0.04	0.03	

100 Epochs:

Model #1-E:		Local:	Global:	F1 Score:
EO	DOF	0.71	0.74	0.83
P-RF	CWT	0.39	0.43	
Acoustic	FFT	0.01	0.03	
Seismic	GS	0.02	0.01	

50 Epochs:

Model #1-F:		Local:	Global:	F1 Score:
EO	DOF	0.59	0.63	0.71
P-RF	CQB	0.31	0.39	
Acoustic	PSA	0.02	0.04	
Seismic	DIP	0.03	0.05	

100 Epochs:

Model #1-F:		Local:	Global:	F1 Score:
EO	DOF	0.72	0.77	0.86
P-RF	CQB	0.34	0.41	
Acoustic	PSA	0.01	0.03	
Seismic	DIP	0.02	0.04	

50 Epochs:

Model #1-G:		Local:	Global:	F1 Score:
EO	DOF	0.61	0.64	0.67
P-RF	WVD	0.33	0.4	
Acoustic	FFT	0.05	0.04	
Seismic	Deconvolution	0.04	0.02	

100 Epochs:

Model #1-G:		Local:	Global:	F1 Score:
EO	DOF	0.69	0.73	0.82
P-RF	WVD	0.34	0.41	
Acoustic	FFT	0.03	0.02	
Seismic	Deconvolution	0.02	0.01	

Figure 38. Reduced Modality Impact Over Training Epochs

Even ignoring the initially dismal values that the acoustic and seismic data started with, this is clearly *not* a desirable performance by any of the proposed metrics. The reduced local and global impact over the course of *more* training epochs indicates that the model was learning to *not use* the data (Figure 38). Owing to the model's local and global impact at four, three, and two modality inputs wouldn't exceed 0.06 after the requisite amount of training epochs, it was painfully clear that when compared to the Acoustic (MS) data's performance, the Mel Spectrogram should be chosen over the Fast Fourier Transform and Power Spectrum Analysis.

With regards to the use of Seismic data, Geometric Spreading (GS) function, DIP filtering, and Deconvolution, similar issues ensued with these modalities. The repeating issues from the acoustic modalities tested caused similar checks to be made at 50 and 100 epochs. The results indicated similar to the Acoustic (FFT) and Acoustic (PSA), the Seismic GS, DIP, and Deconvolution were also being trained to have next to no local and global impact on the decision-making process. Because the local and global impact results for these fusion models aren't incentivized to use this data for vehicle detection, it is difficult to justify their inclusion in the experimentation section.

5.4 Model Efficiency

Finally, for the purposes of measuring efficiency, and because of SHAP's PyTorch support, the efficiency of each of the models is measured using PyTorch's flopth function. While the measurement of FLOPs within neural network models with the same

number of layers, and batch sizes should theoretically have similar results, measuring the actual memory and power efficiency of different models will differ based on factors like hardware and input dimensions that the data takes. Traditionally, measuring FLOPs on a high-performance computing system are calculated as follows:

$$FLOPS = racks \times \frac{nodes}{rack} \times \frac{sockets}{node} \times \frac{core}{socket} \times \frac{cycles}{second} \times \frac{FLOPs}{cycle} \quad \#(23)$$

This calculation can be simplified based on the total number of CPUs, however it doesn't help illustrate the nature of calculating neural network performance when taking into account its architecture. In order to calculate operations for convolutional kernels, it's crucial to remember the number of channels the kerel should match the number of channels in the input. Breaking the calculations down to the main components of the CNN, the following equations are used to estimate FLOPs for different parts of a CNN:

$$Convolutions: FLOPs = 2 \times \#\ of\ Kernels \times Kernel\ Shape \times Output\ Shape \quad (24)$$

$$Fully\ Connected\ Layers: FLOPs = 2 \times Input\ Size \times Output\ Size \quad \#(25)$$

$$Pooling\ Layers: FLOPs = Height \times Depth \times Width\ of\ an\ image \quad \#(26)$$

$$With\ a\ Stride: FLOPs = \frac{Height}{Stride} \times Depth \times \frac{Width}{Stride}\ of\ an\ image \quad \#(27)$$

While the actual quantity of FLOPs for any given neural network is dependent on the hardware and software implementations, it is possible to derive a typical quantity from expanding the layer-by-layer operations for each parameter and weight, by making reasonable implementation assumptions for each activation function. By taking this approach, it becomes more clearly obvious that input dimensions proportionately affect the computations for the first layer of the convolutional model.

Owing to the lack of optimization in these models, which will be discussed more in Chapter 7.2, the constants of having both the same device handling training and the models not being optimized for efficiency will put the calculations for efficiency on even

footing for comparison. While, as discussed in the literature review, FLOPs on its own is insufficient metric, because the model architecture and hardware are shared, the use of the metric is justified for this research, albeit again not being optimized for efficiency on the software end. Sensor fusion models will be compared to see if the use of additional sensor modalities will impact their F1 Score and computational efficiency in FLOPs. Doing so will aid in determining if a modality's impact is worth the added computational complexity of the model.

CHAPTER SIX

EXPERIMENTATION

6.1 Phase I

For the purposes of experimentation, the original plan was to compare the use of four-source, three-source, and two-source multimodal fusion models for EO, P-RF, Acoustic, and Seismic data. The impact of each of the modalities (Local and Global), the model's performance (F1 Score), and the computational costs (FLOPs) would be compared for each of these models, as covered in chapter 5. Phase I would be the preliminary experimental results with the smallest number of fusion models as different combinations of sensor data were fed into the models, with Phase II (three-source multimodal fusion models) and Phase III (two-source multimodal fusion models) having more models to compare owing to a larger set of permutations.

With the performance of Mel Spectrogram and Cepstrum being the only viable preprocessing methods tested for Acoustic and Seismic data from the ESCAPE dataset, the experimentation for Phase I and incidentally Phase II and III is reduced somewhat from what was originally planned during the proposal. Like the removal of the P-RF histogram data, the decision to remove the modalities was owing to performance and the fusion model's actual usage of the data. The justifications for the removal of those modalities are covered in Section 5.2 (Figure 38) and in Section 7.2 under the caveats for the research conducted.

Because of the singular viable sources of acoustic and seismic data, the number of four-source models is limited only by the number of P-RF sources (CWT, CWB, WVD, and STFT). These four models are referred to as Model #1-A, Model #1-B, Model #1-C, Model #1-D respectively. As expected from the use of four or more sources of data, the results in terms of overall performance are the highest of all the phases, with Model #1-A having the lowest F1 score of 0.95, and with Model #1-B and Model #1-C having an F1 score of 0.98. However, owing to the large amount of data, larger parameters, and the greater complexity of the models, the computational costs of the models are also considerably larger by comparison, as they can be measured in Teraflops, as seen below in Figure 39.

Model #1-A:		Local:	Global:	F1 Score:	FLOPs:
EO	DOF	0.74	0.67	0.95	206.73 Gflops
P-RF	CWT	0.39	0.42		
Acoustic	MS	0.09	0.18		
Seismic	Cepstrum	0.13	0.21		

Model #1-C:		Local:	Global:	F1 Score:	FLOPs:
EO	DOF	0.71	0.64	0.98	211.69 Gflops
P-RF	CQB	0.43	0.49		
Acoustic	MS	0.19	0.23		
Seismic	Cepstrum	0.26	0.21		

Model #1-B:		Local:	Global:	F1 Score:	FLOPs:
EO	DOF	0.79	0.72	0.98	182.42 Gflops
P-RF	WVD	0.38	0.43		
Acoustic	MS	0.11	0.14		
Seismic	Cepstrum	0.09	0.17		

Model #1-D:		Local:	Global:	F1 Score:	FLOPs:
EO	DOF	0.77	0.71	0.96	198.01 Gflops
P-RF	STFT	0.36	0.39		
Acoustic	MS	0.17	0.28		
Seismic	Cepstrum	0.21	0.25		

Figure 39: Phase I Model #1-A through Model #1-D comparison

The performance of Model #1-B and Model #1-C in particular scored the highest in terms of F1 score. Model #1-C is notably less reliant on the EO data (lower local and global impact) but also simultaneously is computationally more expensive by comparison with 211.69 Gigaflops as opposed to Model #1-B's 182.42 Gigaflops. The overall impact of the seismic data is greater in Model #1-B's model however, though the acoustic data is less impactful compared to Model #1-C. The greatest impact for both the acoustic and

seismic data is seen in Model #1-D, where their global impact is 0.28 and 0.25 respectively.

Conversely, the reliance on EO data is especially high for Model #1-B and Model #1-D when compared to Model #1-A and Model #1-C. The higher impact of the other non-EO modalities is certainly seen in Model #1-C, but appears to have come at a cost of almost 30 Gigaflops.

Model #1-E:		Local:	Global:	F1 Score:	FLOPs:
EO	Thresholding	0.78	0.74	0.85	199.31 Gflops
P-RF	CWT	0.35	0.33		
Acoustic	MS	0.11	0.15		
Seismic	Cepstrum	0.14	0.19		

Model #1-F:		Local:	Global:	F1 Score:	FLOPs:
EO	Thresholding	0.73	0.76	0.82	216.44 Gflops
P-RF	CQB	0.38	0.42		
Acoustic	MS	0.17	0.25		
Seismic	Cepstrum	0.18	0.24		

Model #1-G:		Local:	Global:	F1 Score:	FLOPs:
EO	Thresholding	0.79	0.81	0.89	219.41 Gflops
P-RF	WVD	0.35	0.39		
Acoustic	MS	0.14	0.19		
Seismic	Cepstrum	0.06	0.14		

Model #1-H:		Local:	Global:	F1 Score:	FLOPs:
EO	Thresholding	0.79	0.75	0.87	206.92 Gflops
P-RF	STFT	0.33	0.38		
Acoustic	MS	0.19	0.09		
Seismic	Cepstrum	0.16	0.23		

Figure 40. Model #1-E through Model #1-H comparison

When compared to Models #1-A through #1-D, the models that used Thresholding (. Model #1-E through Model #1-H comparison) as a source of EO data suffered in terms of both F1 score performance and computational efficiency, and to some degree even in terms of sensor fusion. The impact of Acoustic and Seismic Data was rather disappointing overall, as the Thresholding based models appeared to be somehow even more reliant on the EO data compared to the DOF based models. Naturally, the same can be said regarding the P-RF data as well.

While the local and global impact of the acoustic and seismic data wasn't all that different from the DOF based models, there are a few notable trends. With the exception of Model #1-F, all of the Thresholding based models had a higher local impact of Acoustic data when compared with their DOF counterparts, with Model #1-G and Model

#1-F both having a higher Global impact as well. The global impact of the Seismic data for Models #1-F was notably higher than its DOF counterpart Model #1-C, and similarly the Seismic local impact of Model #1-E was higher than that of its DOF counterpart Model #1-A.

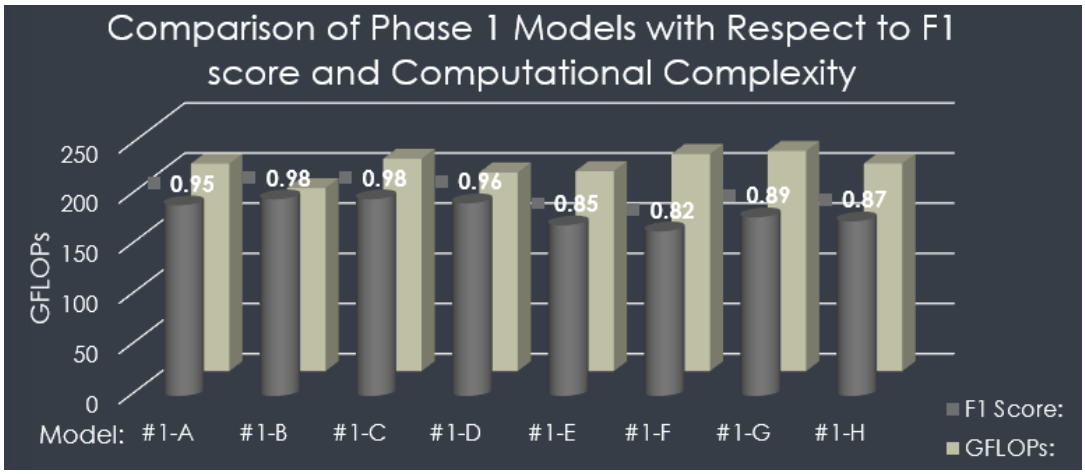

Figure 41. Comparison of Phase I Models

Model #1-C's usage is dwarfed by Model #1-G's FLOPs, and the lowest of the four, Model #1-E, has more FLOPs than Model #1-B and Model #1-D respectively. The F1 Scores were relatively lower, as Model #1-F scores the lowest of all eight fusion models for Phase I. There is not much else to say about the four Phase I models otherwise, as having access to four sources of data resulting in higher Gigaflops and higher F1 scores overall was always within expectations.

6.2 Phase II

The second phase of testing hosts a considerably larger number of models than Phase I, owing to testing for all available combinations of three modality sources. Models

#2-A through #2-H (Figure 42) use different combinations of EO, P-RF, and either the Seismic or Acoustic sensor data, as seen in Figure 42. Models #2-I through #2-L (Figure 45) use different combinations of P-RF, Seismic, and Acoustic data, while Model #2-M (Figure 46) uses only EO, Seismic, and Acoustic data. The majority of these models still perform relatively well, as seen below in Figure 42 and Figure 46, but models without access to the EO sensor data suffer in terms of performance, as seen in Figure 43.

Model #2-A:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.72	0.79	0.93	131.74 Gflops
P-RF	CWT	0.44	0.46		
Acoustic	MS	0.21	0.23		

Model #2-B:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.77	0.81	0.96	128.81 Gflops
P-RF	CQB	0.43	0.48		
Acoustic	MS	0.26	0.29		

Model #2-C:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.83	0.76	0.89	103.89 Gflops
P-RF	WVD	0.47	0.51		
Acoustic	MS	0.2	0.24		

Model #2-D:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.74	0.69	0.93	119.02 Gflops
P-RF	STFT	0.41	0.49		
Acoustic	MS	0.27	0.31		

Model #2-E:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.78	0.71	0.94	148.36 Gflops
P-RF	CWT	0.43	0.49		
Seismic	Cepstrum	0.19	0.25		

Model #2-F:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.79	0.67	0.91	127.54 Gflops
P-RF	CQB	0.39	0.41		
Seismic	Cepstrum	0.22	0.28		

Model #2-G:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.73	0.76	0.97	120.33 Gflops
P-RF	WVD	0.4	0.42		
Seismic	Cepstrum	0.23	0.27		

Model #2-H:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.76	0.66	0.92	114.97 Gflops
P-RF	STFT	0.39	0.43		
Seismic	Cepstrum	0.2	0.24		

Figure 42: Results for Models #2-A through #2-H

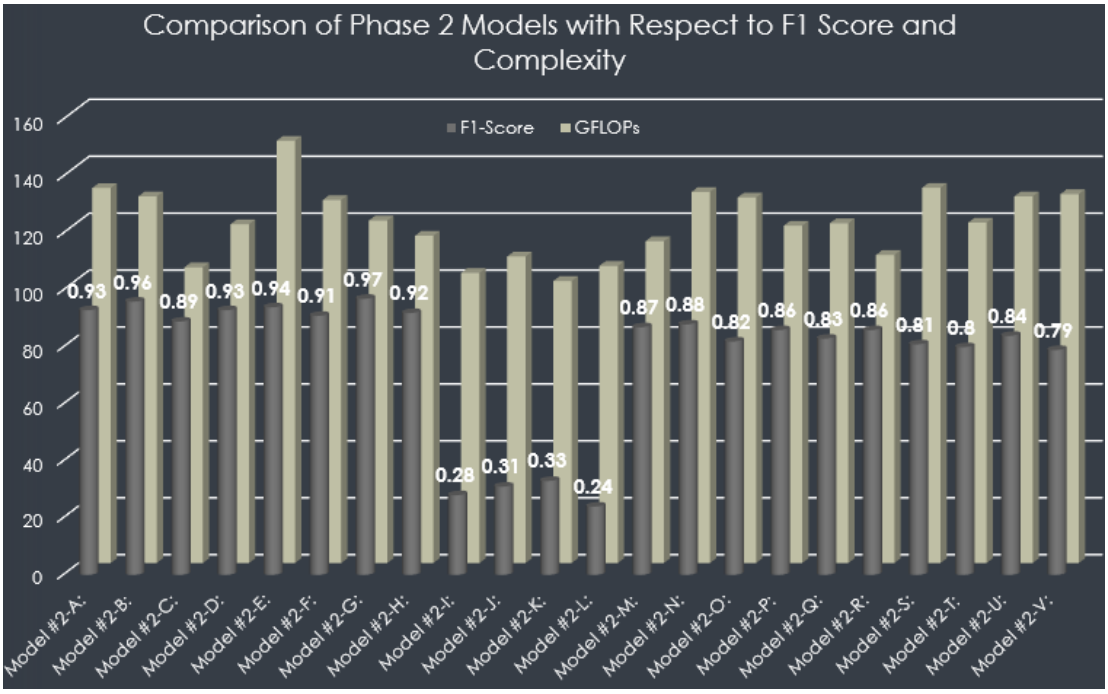

Figure 43. Comparison of Phase 2 Models

Overall, the reduced level of complexity for supporting three sources of data instead of two can be seen in the results for the Phase II models. While the measurements are still relatively high, even when compared to Model #1-B, the highest computational cost of the Phase II models is only Model #2-E's 148.36 Gigaflops vs Model #1-B's 182.42 Gigaflops. Even with only three sources of sensor data, as opposed to four, the majority of the models still perform relatively well, with Model #2-G being barely below Model #1-B and Model #1-C's F1 score of 0.98. Model #2-G also has considerably lower computational costs and lower EO local impact when compared to #1-B.

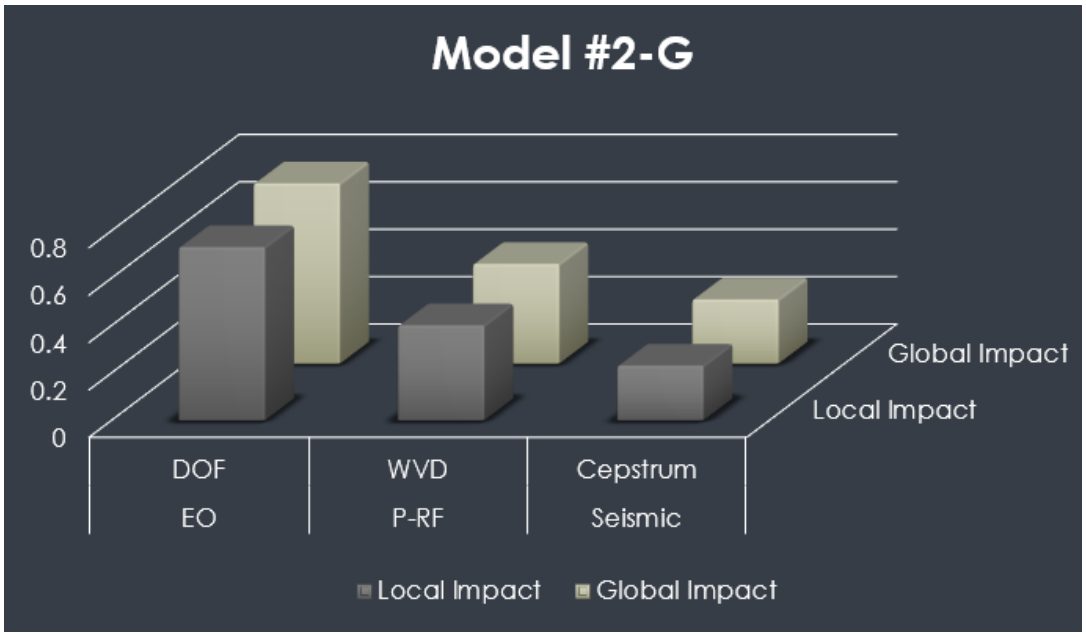

Figure 44. Model #2-G

Model #2-I:		Local:	Global:	F1-Score:	FLOPs:
P-RF	CWT	0.25	0.23	0.28	101.92 Gflops
Acoustic	MS	0.11	0.19		
Seismic	Cepstrum	0.14	0.21		

Model #2-J:		Local:	Global:	F1-Score:	FLOPs:
P-RF	CQB	0.17	0.24	0.31	107.68 Gflops
Acoustic	MS	0.09	0.13		
Seismic	Cepstrum	0.1	0.16		

Model #2-K:		Local:	Global:	F1-Score:	FLOPs:
P-RF	WVD	0.26	0.22	0.33	99.08 Gflops
Acoustic	MS	0.19	0.16		
Seismic	Cepstrum	0.12	0.15		

Model #2-L:		Local:	Global:	F1-Score:	FLOPs:
P-RF	STFT	0.21	0.14	0.24	104.39 Gflops
Acoustic	MS	0.12	0.18		
Seismic	Cepstrum	0.17	0.21		

Figure 45: Results for Models #2-I through #2-L

For Model #2-I through #2-L, the performance for a model that is supposed to differentiate between ten targets is rather poor. The overall impact of each of the modalities is rather low, which coupled with the low F1 Scores indicates a large amount of guessing on the part of these four models. There is also a notable drop in computational costs for these when compared to #2-M and models #2-A through #2-H, with the highest of these four being Model #2-J, with Model #2-K's computational costs notably dropping below 100 Gigaflops.

Model #2-M:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.77	0.73	0.87	113.06 Gflops
Acoustic	MS	0.13	0.16		
Seismic	Cepstrum	0.09	0.14		

Model #2-N:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.74	0.71	0.88	130.37 Gflops
Acoustic	MS	0.13	0.21		
Seismic	Cepstrum	0.11	0.15		

Figure 46: Results for Model #2-M and #2-N

Model #2-M performs relatively poorly compared to its EO/P-RF counterparts, underperforming even Model #2-C's F1 score. The relatively lower impact of the acoustic and seismic data indicates Model #2-M's difficulties in successfully utilizing the modalities, as its Thresholding counterpart, Model #2-N, outperforms it in every metric, having greater computational cost, marginally higher F1 score, and also having relatively higher impact on local and global level for Acoustic and Seismic data. It should also be noted that this F1 score is not much higher than previous single modality EO models have previously achieved with this dataset.

Model #2-O:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.74	0.81	0.82	128.44 Gflops
P-RF	CWT	0.39	0.42		
Acoustic	MS	0.19	0.21		

Model #2-P:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.79	0.83	0.86	118.53 Gflops
P-RF	CQB	0.36	0.42		
Acoustic	MS	0.21	0.24		

Model #2-Q:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.83	0.76	0.83	119.37 Gflops
P-RF	WVD	0.45	0.48		
Acoustic	MS	0.17	0.23		

Model #2-R:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.78	0.72	0.79	108.23 Gflops
P-RF	STFT	0.35	0.39		
Acoustic	MS	0.23	0.29		

Model #2-S:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.82	0.73	0.81	131.83 Gflops
P-RF	CWT	0.34	0.42		
Seismic	Cepstrum	0.16	0.21		

Model #2-T:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.81	0.76	0.8	119.6 Gflops
P-RF	CQB	0.36	0.39		
Seismic	Cepstrum	0.19	0.26		

Model #2-U:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.73	0.76	0.84	123.83 Gflops
P-RF	WVD	0.4	0.42		
Seismic	Cepstrum	0.23	0.27		

Model #2-V:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.79	0.72	0.82	129.61 Gflops
P-RF	STFT	0.39	0.43		
Seismic	Cepstrum	0.16	0.2		

Figure 47. Model #2-O through #2-V

The performance of the Thresholding based models are otherwise rather lackluster, with the highest performing model being Model #2-P at an F1 score of 0.86. In terms of computational cost, none of the Thresholding based models exceed the FLOPs of Model #2-G nor do any of them use less than their DOF equivalent Model #2-C or

even Model #2-K. The observed trends overall indicate a reduced reliance on the P-RF, Acoustic, and Seismic data for the Thresholding based models, consistent with their equivalents in Phase I.

6.3 Phase III

The last phase of experimentation, Phase III, showcases the use of multiple two-source sensor fusion models. As seen in Figure 48 and Figure 50, as was previously the case with the Phase II models, any models with EO sensor data performed relatively better than models without access to EO sensor data. The reliability of a sensor fusion model designed for ten possible classification performs considerably worse when the sensor with the most consistent and strongest impact on decision-making is missing. When compared to Phase II's models, the computational costs are relatively lower, but not by as much of a drop compared to the difference between Phase I and Phase II's sensor fusion models. The performance of Phase III's EO/P-RF models outperform the other sensor combinations, similar to in Phase II, with Model #3-C (Figure 48) having comparable F1 score to Model #2-G. However, unlike Model #2-G, Model #3-C requires only the EO and P-RF data, and its computational costs are nearly half as much as Model #2-G's. Conversely, the EO local impact is much higher for Model #3-C than it is for Model #2-G.

Model #3-A:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.79	0.66	0.94	73.95 Gflops
P-RF	CWT	0.38	0.46		

Model #3-B:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.76	0.68	0.96	69.07 Gflops
P-RF	CQB	0.42	0.45		

Model #3-C:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.8	0.67	0.97	68.33 Gflops
P-RF	WVD	0.49	0.56		

Model #3-D:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.81	0.67	0.91	67.38 Gflops
P-RF	STFT	0.38	0.41		

Figure 48: Results for Models #3-A through #3-D

Similar to Phase II, the Phase III models that were trained only using P-RF data, and either seismic or acoustic data performed relatively poorly. The computational costs were not notably smaller, and given the model's low local and global impacts and low F1 score would indicate that like Models #2-I through #2-L, Phase III's Models #3-E through #3-L failed to learn relevant features without the context the EO data provides, as seen below in Figure 49.

Model #3-E:		Local:	Global:	F1-Score:	FLOPs:
Acoustic	MS	0.13	0.18	0.31	69.05 Gflops
P-RF	CWT	0.22	0.24		

Model #3-F:		Local:	Global:	F1-Score:	FLOPs:
Acoustic	MS	0.11	0.19	0.19	74.01 Gflops
P-RF	CQB	0.16	0.2		

Model #3-G:		Local:	Global:	F1-Score:	FLOPs:
Acoustic	MS	0.11	0.17	0.27	79.28 Gflops
P-RF	WVD	0.14	0.22		

Model #3-H:		Local:	Global:	F1-Score:	FLOPs:
Acoustic	MS	0.14	0.22	0.23	63.49 Gflops
P-RF	STFT	0.19	0.21		

Model #3-I:		Local:	Global:	F1-Score:	FLOPs:
Seismic	Cepstrum	0.08	0.14	0.23	68.73 Gflops
P-RF	CWT	0.13	0.19		

Model #3-J:		Local:	Global:	F1-Score:	FLOPs:
Seismic	Cepstrum	0.06	0.11	0.29	65.83 Gflops
P-RF	CQB	0.15	0.23		

Model #3-K:		Local:	Global:	F1-Score:	FLOPs:
Seismic	Cepstrum	0.09	0.17	0.19	76.04 Gflops
P-RF	WVD	0.12	0.2		

Model #3-L:		Local:	Global:	F1-Score:	FLOPs:
Seismic	Cepstrum	0.05	0.11	0.22	71.56 Gflops
P-RF	STFT	0.13	0.19		

Figure 49: Results for Models #3-E through #3-L

Model #3-M:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.76	0.82	0.83	61.96 Gflops
Acoustic	MS	0.21	0.17		

Model #3-N:		Local:	Global:	F1-Score:	FLOPs:
EO	DOF	0.79	0.85	0.77	71.70 Gflops
Seismic	Cepstrum	0.17	0.14		

Figure 50: Results for Models #3-M and #3-N

For the EO-Acoustic and EO-Seismic models (#3-M and #3-N respectively), the fusion of the data performed worse when compared to Phase II's Model #2-M, as seen below in Figure 50. The fusion with the EO-Acoustic data had a lower F1 score, having

both lower computational costs and performance. Model #3-N's EO-Seismic sensor fusion model performed notably even worse, while also still having a higher computational cost compared to Model #3-M.

Model #3-O:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.79	0.85	0.81	72.49 Gflops
Acoustic	MS	0.19	0.14		

Model #3-P:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.73	0.69	0.79	78.16 Gflops
Seismic	Cepstrum	0.21	0.19		

Model #3-Q:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.72	0.67	0.83	77.31 Gflops
P-RF	CWT	0.34	0.47		

Model #3-R:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.73	0.65	0.85	74.27 Gflops
P-RF	CQB	0.39	0.43		

Model #3-S:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.77	0.74	0.8	69.19 Gflops
P-RF	WVD	0.45	0.51		

Model #3-T:		Local:	Global:	F1-Score:	FLOPs:
EO	Thresholding	0.74	0.67	0.78	81.63 Gflops
P-RF	STFT	0.33	0.39		

Figure 51. Results for Model #3-O through Model #3-T

Notably, for the Thresholding based models, Model #3-P outperforms its DOF equivalent Model #3-N, having a marginally higher F1 score, and with higher local and global impact for the Seismic data. It also is computationally more expensive, a trait that it shares with Model #3-O when compared to its DOF equivalent Model #3-M, which conversely has a higher EO impact but lower F1 score. Model #3-T achieves the highest computational cost of the Phase III models, at 81.63 Gflops.

For the most part, the Thresholding based models do not outperform their DOF counterparts (Figure 51). The Thresholding based models follow a trend of consistently having an overall larger number of FLOPs for consistently lower F1 scores. Besides the

previously mentioned exceptions, the most promising model results were from Models #3-C and Models #3-B for Phase III models tested.

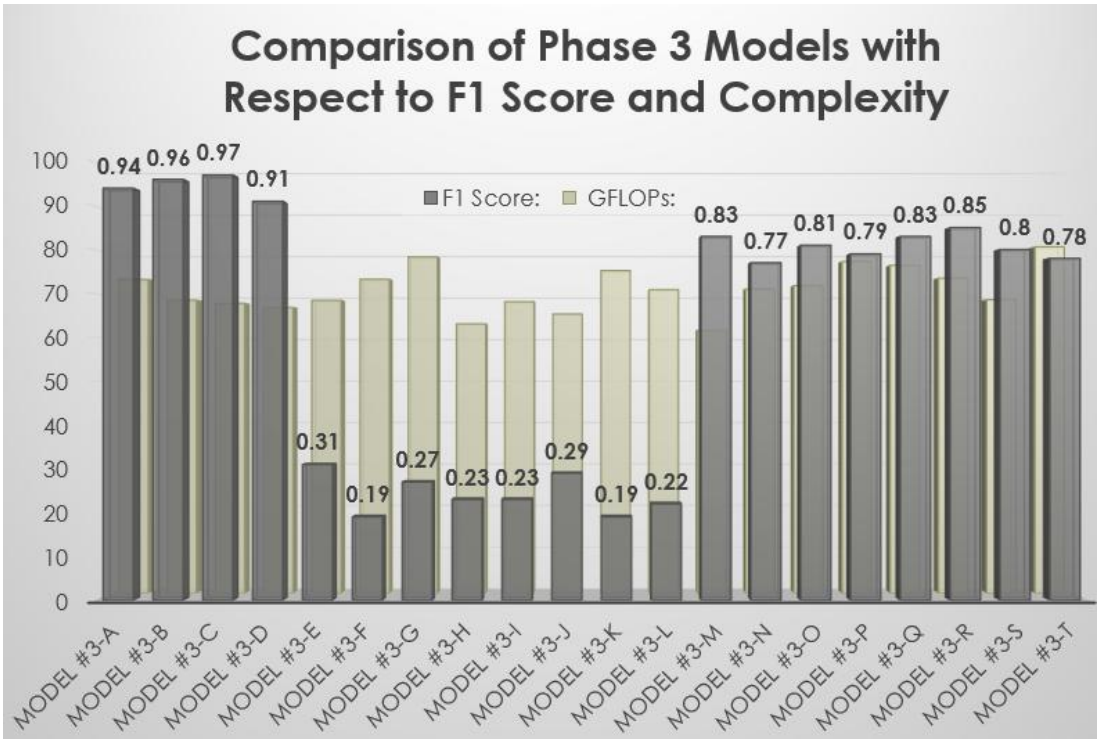

Figure 52. Comparison of Phase 3 Models

Based on the completed results of the three phases of models, the top performing models of the 31 fusion models tested were in Phase I, as both Model #1-B and Model #1-C managed to achieve an F1 score of 0.98. Models #2-G and #3-C managed to achieve an F1 score of 0.97, while Models #1-D and #3-B managed to achieve an F1 score of 0.96 respectively. Of the models that achieved an F1 score of above a 0.95, Model #-C required the lowest computational complexity cost, measured in FLOPs (at 68.33 Gigaflops) using only EO (DOF) and P-RF (WVD), with Model #3-B having a lower F1 score but almost as few FLOPs needed for training (at 69.07 Gigaflops).

With respect to performance and modality impact, Model #1-C, while requiring 16% more FLOPs to complete training compared to Model #1-B, had a more desirable distribution of local and global impact for the modalities tested. #1-C had the largest P-RF, Acoustic, and Seismic impact of Phase I, while having the lowest EO impact. In Phase II, the two highest performing models, #2-B and #2-G showcase relatively higher use of P-RF data, Acoustic, and Seismic data, with #2-G having the lower amount of EO impact on the local and global level. In Phase III, Model #3-C had the lowest global impact of the two highest performing models for that phase (#3-B and #3-C) while also having the highest WVD impact.

With respect to the P-RF modalities, WVD and CQB generally performed above average. Models #1-B (0.98), #2-G (0.97), #3-C (0.97) all used WVD for P-RF, while models #2-B (0.96), #3-B (0.96) used CQB. In particular, CQB performed better with acoustic data in Phase II, as opposed to CWT and WVD performing better with seismic data by comparison, with Model #2-C underperforming in particular. Fusion models that utilized STFT generally underperformed. In Phase I Model #1-D only barely outperforming #1-A in Phase I, but this underperformance is seen particularly in Phase II, with #2-D, #2-H, #2-L, and #3-D underperforming for the EO/P-RF fusion models in Phase III. In Phase III, the STFT models placed second lowest for their respective categories, with #3-L (STFT) narrowly outperforming #3-K (WVD), Model #3-F (CQB) narrowly outperforming #3-H (WVD). These results were generally in line with prior research [79] on P-RF usage we've previously conducted.

Lastly, the impact of the Acoustic and Seismic data was relatively low in the models they were used in, as seen particularly painfully in Models #2-M, #3-M, #3-N, in which the consistently highest performing sensor modality, EO, was not enough to improve to their performance to even above an F1 score of 0.9. Prior research has indicated the fusion results, specifically #3-N, were lower than previously tested standalone EO only trained models. The relatively poorer performance of these modalities is addressed in section 7.2, along with other caveats for the metrics and algorithms used in this research.

CHAPTER SEVEN

CONCLUSION

7.1 Overview of Thesis

The results of the research presented in this thesis indicate a few interesting trends towards the usage of the ESCAPE dataset's sensors for heterogenous multimodal fusion, with respect to performance and computational efficiency. In this thesis, we present the relevant research related to the multimodal fusion for vehicle detection, and in particular the research conducted to better understand the P-RF modality, which inherently necessitated further research into explainability. By implementing XAI methods and comparing computational cost and model performance, we can better understand how to more efficiently implement multimodal fusion from the available sources of data in order to facilitate uncertainty reduction while also avoiding feeding unnecessary sources of sensor data.

The incredible potential of deep learning has many potential avenues and applications which can enable advances in the progress of human scientific understanding. The progress of human knowledge is a marathon across time. Even less than a hundred years ago the idea that carrots somehow enabled people to see in the dark was somehow more believable than the technology that enabled the broadcasting of music and news could broadly and inexplicably somehow be used to reliably detect planes, instead of using the tried-and-true method of listening to a concrete acoustic mirror by the ocean. A tidbit of nutritional misinformation that somehow still manages to

persist in a warped form to the modern day. As the AI field continues to develop, it is important to integrate good practices for the training of reliable and transparent deep learning models.

7.2 Caveats

As the work presented in this thesis covers a lot of topics, modalities, and metrics, it seemed prudent to address certain points and certain choices that were made in the process of the research. To that end, covering a few caveats about the research within this thesis is necessary, in particular with regards to the metric of computational complexity cost and the underwhelming performance of two modalities that theoretically should have been more viable for these experiments. As well as addressing why certain metrics were chosen, and also with respect to the performance of certain modalities.

With regards to the usage of F1 score for comparing multiple models for the same classification task, while F1 score is typically used for classification tasks, other metrics could have been selected. When it comes to statistical hypothesis testing for two matched samples, the Wilcoxon signed-rank test is largely considered to be a good option, though in this particular case an approach such as the Friedman's Test would likely be a better approach. As a non-parametric alternative to ANOVA, the rankings it would provide would likely be a more ideal statistical measurement of model performance. However, in the interests of time, and because the evaluation of computational complexity costs and how well the model fused the data were of *greater interest* in the context of this research,

the usage of F1 score was not addressed any further. Especially as F1 score is normally used for classification tasks, which the sensor fusion deep learning models tested in this research were trained to complete.

With regards to the usage of the FLOPs as a measurement of computational costs, it should be stated that the work presented in this thesis has not been optimized for reducing computational costs. Deep learning requires a considerable amount of calculations, power, and runtime to complete training, and considerations between the types of processors [98] used and algorithmic [63] changes are important for the purposes of calculating or reducing computational costs incurred during training.

With that in mind, the total number of FLOPs calculated for the fusion models experimented with are considerably higher than the theoretical minimum number of FLOPs required for training. Owing to inefficiencies within the training process, a lack of algorithmic optimization, and older hardware, the reason for including the metric is that all models from Phase I through III were tested and trained on the same device, and therefore the total number of FLOPs provides a means of determining computational costs relative to each of the generated fusion models.

Besides the usage of FLOPs for measuring computational efficiency of the fusion model, the relatively poorer performance of the Acoustic and Seismic data should also be addressed. While other researchers [99] [100] have made better use of these sources of data, the use of Seismic and Acoustic sensor modalities was relatively limited in the research presented in this thesis. Previously published work with the ESCAPE dataset did not use either of those modalities and owing to the nature of Seismic and Acoustic data,

the use of a CNN for the fusion models, rather than a model capable of taking into account the temporal aspect of the data (such as an LSTM or RNN) also did not help with the sensor modalities used.

Lastly, it should be noted that Diverse Counterfactual Explanations are but *one* potential tool in attempting to solve deep learning's explainability problem, particularly when it comes to how the approach generates "blind" perturbations [101]. This can lead to issues wherein the resulting counterfactuals may not use valid datapoints (not creating "native" counterfactuals, i.e. using features in ways that do not naturally occur), may lack the sparsity of good counterfactuals (if they modify too many features), or may even lack diversity (if the generated counterfactuals minimally vary from each other). That being said, among causal-based methods, DiCE provides several benefits that are ideal for human users [102], that make it a better choice over attributable methods. Ideally the DiCE can be beneficial in providing explanations to those affected by its decisions, situationally provide actionable feedback, potentially aid developers in identifying, detecting, and fixing bugs and other performance issues. While there are general concerns in the field over the actionability, sparsity, and data manifold closeness, particularly with regards to *how* counterfactuals are generated, the results presented were *consistent* with the results provided by other Post-Hoc explainability methods, including both attributable [93] and explanation form based [91] methods.

With regards to how useful the the local and global scores provided by DiCE regarding each modality are, there are several ways in which this information can be used to enhance decision-making and optimize fusion strategies. Adaptive sensor weighting

[103], for example, can assign higher weights to the inputs to sensors with higher global scores. If local scores are varying, providing a context-aware weighting mechanism to adjust the weights based on real-time conditions can provide less variance. More traditional algorithms, such as Bayesian inference or Dempster-Shafer theory [104] can integrate sensor trust scores into the fusion process, and even introduce confidence thresholds into highly varying local scores. Depending on if sensors are consistently performing better in specific scenarios, adaptive fusion methods that switch between decision-level, feature-level, or score-level fusion may be worth considering for more dynamic fusion strategy selection. If a sensor has low local scores, this may also indicate noisy data or malfunction, which anomaly detection techniques can aid in fault detection and finding outliers. In Reinforcement Learning applications, rewarding optimal weighting strategies by rewarding correct decisions and penalizing incorrect ones can be done by leveraging the local and global scores as feedback. Generally speaking, reducing or pruning redundant sensors can optimize computational efficiency and reduce system complexity, which local and global scores can help in making those decisions, or be used to indicate if additional algorithms and techniques, such as probabilistic fusion (Kalman filters, particle filters, or probabilistic graphical models) can aid in uncertainty reduction.

7.3 Contributions and Future Direction

In this paper, we present a deep learning framework evaluation approach (Figure 53) that provides methods of evaluating a multimodal sensor fusion model, and its results

when applied to the ESCAPE dataset. This evaluation framework provides three different metrics for comparison. The first is through predictive performance, measuring each fusion model's predictive performance in each different iteration of sensor fusion via F1 score. The second is providing insights with post-hoc causal XAI method, DiCE. Through these insights, the data provided by the counterfactuals can provide some understanding of how the model's decisions are made with respect to the sensor data it is fed on both a local and global scale. Lastly, through GFLOPs, it becomes possible to measure comparative performance with respect to the computational complexity of each of the sensor fusion models tested. This framework and its three metrics can be used to evaluate alternatives other than the naive use of all available data for a multimodal sensor fusion model, which can then be used to make decisions regarding resource usage.

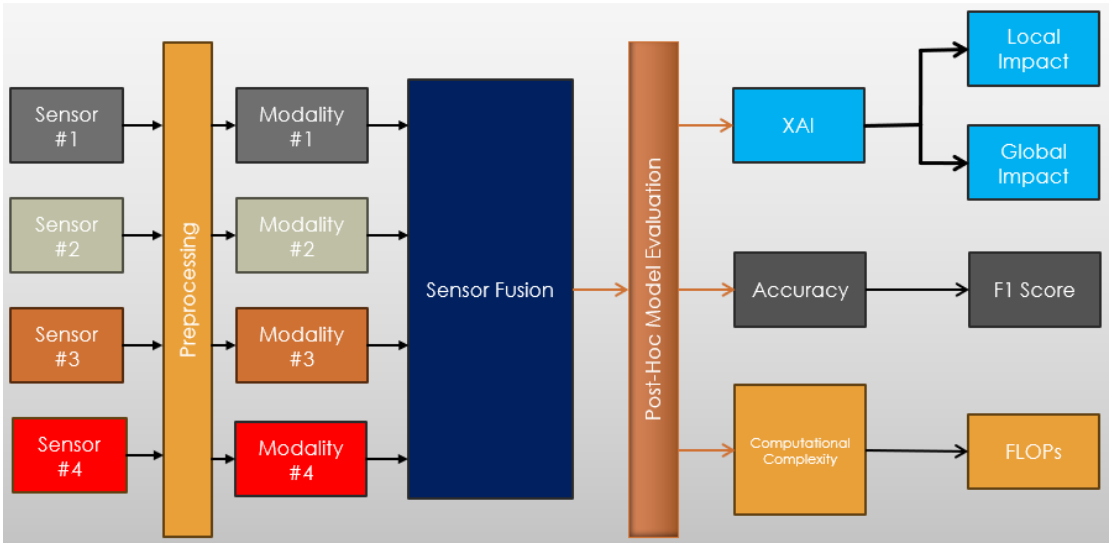

Figure 53. DL Framework Evaluation Approach

As shown with the results of its testing with the aforementioned dataset, it becomes clearer what "optimal" options may be worth considering with respect to the available resources. While the naive use of all sensor modalities was able to achieve an

F1 score of 0.98 (as seen with both Models #1-B and #1-C), the highest score, it was also at the highest costs in computational complexity relative to F1 score. However, through this testing framework, it is also brought to our attention that comparatively similar performance (F1 score of 0.97) is still achievable with *fewer modalities* required, as well as at a *lower computational complexity* (Models #2-G and #3-C). If one is not concerned with cutting costs for the sensor fusion model, this approach can still provide insights into how much more reliant one sensor fusion scheme is compared to another sensor fusion scheme on certain modalities, as seen with #1-C's tradeoffs of having a higher computational complexity cost but lower reliance on EO data compared to #1-B (Figure 54).

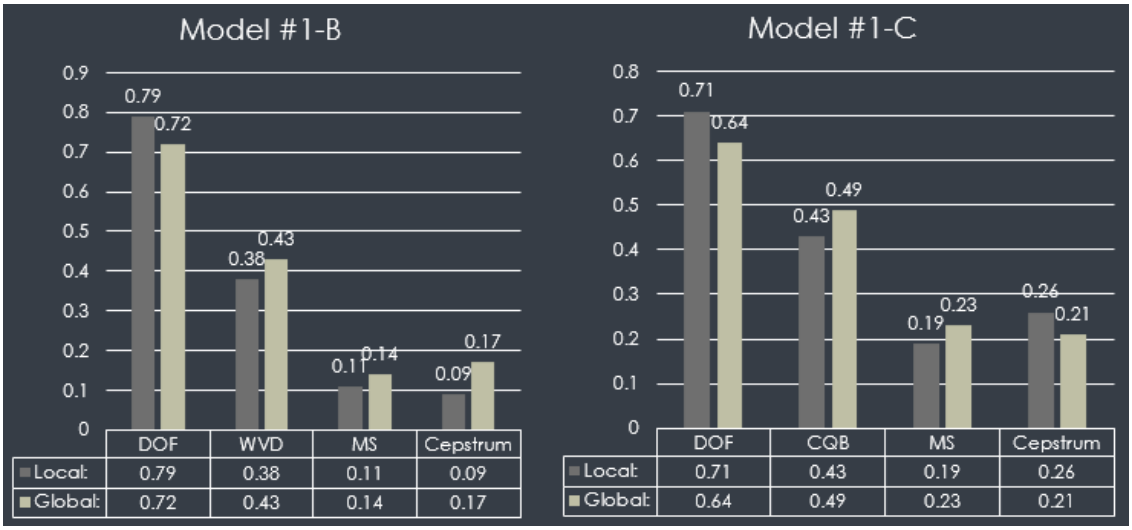

Figure 54. Model #1-B vs #1-C

While my more recent [77] research has moved in the direction that I initially had envisioned when I was originally working on my Master's Thesis, I would ideally prefer to also have more research devoted to the use of acoustic and seismic data modalities, as both are relatively less invasive than EO data. It was originally my hope that the research

for this thesis might manifest results and a fusion model that was less dependent on the EO data for this dataset, but as it currently stands that objective is difficult to accomplish. Should more comparative research be done, a change in direction towards the topic of metacognitive learning would be highly desirable. It would also be ideal to further explore more research on the efficiency of machine learning for these experiments, as the field continues to evolve further, especially with respect to implementing the fusion model for use in the field.

References

[1] H. J. GOP, "https://x.com/," [Online]. Available: https://x.com/JudiciaryGOP/status/1833154509222129884. [Accessed 28 1 2025].

[2] J. Heier, "Design Intelligence - Taking Further Steps Towards New Methods and Tools for Designing in the Age of AI," in *Artificial Intelligence in HCI*, Springer, Cham, 2021.

[3] M. Kawai, "Newly-acquired pre-cultural behavior of the natural troop of Japanese monkeys on Koshima islet," *Primates,* vol. 6, pp. 1-30, 1965.

[4] Z. Tufekci, "Facebook Said Its Algorithms Do Help Form Echo Chambers, and the Tech Press Missed It," *New Perspectives Quarterly,* vol. 32, pp. 9-12, 2015.

[5] J. Fayyad, M. Jaradat, D. Gruyer and H. Najjaran, "Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review," *Sensors,* vol. 15, p. 4420, 2020.

[6] J.-W. Hong and D. Williams, "Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent," *Computers in Human Behavior,* vol. 100, pp. 79-84, 2019.

[7] C. Barabas, M. Virza, K. Dinakar, J. Ito and J. Zittrain, "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

[8] J. E. Fountain, "The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms," *Government Information Quarterly,* vol. 39, no. 2, p. 101645, 2022.

[9] J. T. Mtetwa, K. Ogudo and S. Pudaruth, "Explainable Algorithmic Trading: Unlocking the Black Box with GAN-Based Visualizations," in *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Port Louis, Mauritius, 2024.

[10] Y. Zhang, Y. Weng and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics,* vol. 12, no. 2, p. 237, 2022.

[11] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li and S. Labi, "Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems," *Transportation Research Part C: Emerging Technologies,* vol. 156, p. 104358, 2023.

[12] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim and M. Hobbhahn, "Will we run out of data? Limits of LLM scaling based on human-generated data," in *arXiv*, arXiv, 2024.

[13] B. Cottier, R. Rahman, L. Fattorini, N. Maslej and D. Owen, "The rising costs of training frontier AI models," in *arXiv*, arXiv, 2024.

[14] N. Boyko, "Data Processing and Optimization in the Development of Machine Learning Systems: Detailed Requirements Analysis, Model Architecture, and Anti-Data Drift Strategies," *Journal of Applied Data Sciences,* vol. 5, no. 3, 2024.

[15] A. Lbath and I. Labriji, "Energy Efficiency in AI for 5G and Beyond: A DeepRx Case Study," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, Antwerp, Belgium, 2024.

[16] B. Kahler and E. Blasch, "Sensor Management Fusion Using Operating Conditions," in *2008 IEEE National Aerospace and Electronics Conference*, Dayton, OH, USA, 2008.

[17] E. Blasch and D. Lambert, High-Level Information Fusion Management and Systems Design, Norwood, MA: Artech House, 2012.

[18] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing and J. Li, "A Survey of Multimodal Sensor Fusion for Passive RF and EO Information Integration," *IEEE Aerospace and Electronic Systems Magazine,* vol. 36, no. 7, pp. 44-61, 2021.

[19] Y. Zheng, E. Blasch and Z. Liu, Multispectral image fusion and colorization, Bellingham, Washington: SPIE Press, 2018.

[20] E. Blasch, T. Pham, C.-Y. Chong, W. Koch, H. Leung and D. Braines, "Machine Learning/Artificial Intelligence for Sensor Data Fusion–Opportunities and Challenges," *IEEE Aerospace and Electronic Systems Magazine,* vol. 36, no. 7, pp. 80-93, 2021.

[21] D. Shen, E. Blasch, P. Zulch, M. Distasio, R. Niu, J. Lu, Z. Wang and G. Chen, "A joint manifold leaning-based framework for heterogeneous upstream data fusion," *Journal of Algorithms & Computational Technology,* vol. 12, no. 4, pp. 311-332, 2018.

[22] D. K. Seo, Y. H. Kim, Y. D. Eo, M. H. Lee and W. Y. Park, "Fusion of SAR and Multispectral Images Using Random Forest Regression for Change Detection," *ISPRS Int. J. Geo-Information,* vol. 7, no. 10, p. 401, 2018.

[23] S. Kim, W.-J. Song and S.-H. Kim, "Double Weight-Based SAR and Infrared Sensor Fusion for Automatic Ground Target Recognition with Deep Learning," *Remote Sensing,* vol. 10, no. 1, p. 72, 2018.

[24] N. T. B. Bui, D. C. Pham, B. Q. Nguyen and S. T. Le, "Tracking a 3D target with fusion of 2D radar and bearing-only sensor," in *IEEE International Conference on Industrial Technology (ICIT)*, Lyon, France, 2018.

[25] Q. Zhang, Y. Liu, R. S. Blum, J. Han and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Information Fusion,* vol. 40, no. 1, pp. 57-75, 2018.

[26] J. J. Zhang, A. Papandreou-Suppappola and M. Rangaswamy, "Multi-target tracking using multi-modal sensing withwaveform configuration," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, 2010.

[27] M. Robinson, J. Henrich, C. Capraro and P. Zulch, "Dynamic sensor fusion using local topology," in *IEEE Aerospace Conference*, Big Sky, MT, USA, 2018.

[28] D. Hall and J. Llinas, "An introduction to multisensor data fusion," in *Proceedings of the IEEE*, 1997.

[29] N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," in *arXiv:1811.00908*, 2019.

[30] M. Kulin, T. Kazaz, I. Moerman and E. D. Poorter, "End-to-End Learning From Spectrum Data: A Deep Learning Approach for Wireless Signal Identification in Spectrum Monitoring Applications," *IEEE Access,* vol. 6, pp. 18484-18501, 2018.

[31] R. Saley, R. Queiroz and K. Czarneck, An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software, Safety of the Intended Functionality: SAE, 2020, pp. 13-25.

[32] Z. Yang, A. Zhang and A. Sudjianto, "Enhancing Explainability of Neural Networks Through Architecture Constraints," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 32, no. 6, pp. 2610-2621, 2021.

[33] R. R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar and K. P. Sycara, "Transparency and Explanation in Deep Reinforcement Learning Neural Networks," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.*

[34] W. C. Barott, E. Coyle, T. Dabrowski, C. Hockley and R. S. Stansbury, "Passive multispectral sensor architecture for radar-EOIR sensor fusion for low SWAP UAS sense and avoid," in *IEEE/ION Position, Location and Navigation Symposium - PLANS 2014*, Monterey, CA, USA, 2014.

[35] S. Kemkemian and M. Nouvel, "Sense-and-Avoid System Based on Radar and Cooperative," *Encyclopedia of Aerospace Engineering,* pp. 5-14, 2-15.

[36] G. Fasano, D. Accardo, A. E. Tirri, A. Moccia and E. D. Lellis, "Radar/electro-optical data fusion for non-cooperative UAS sense and avoid," *Aerospace Science and Technology,* vol. 46, no. 1, pp. 436-450, 2015.

[37] M. A. Davenport, C. Hegde, M. F. Duarte and R. G. Baraniuk, "Joint Manifolds for Data Fusion," *IEEE Transactions on Image Processing,* vol. 19, no. 10, pp. 2580-2594, 2010.

[38] H. Martín, A. M. Bernardos and J. Iglesias, "Activity logging using lightweight classification techniques in mobile devices," *Personal Ubiquitous Computing,* vol. 17, no. 4, pp. 675-695, 2013.

[39] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics,* vol. 4, no. 37, 2019.

[40] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. Díaz-Rodríguez and F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion,* vol. 99, no. 1566-2535, p. 101805, 2023.

[41] D. Minh, H. X. Wang, Y. F. Li and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review,* vol. 55, no. 5, pp. 3503-3568, 2021.

[42] G. Menghani, "Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better," *ACM Computing Surveys,* vol. 55, no. 12, pp. 1-37, 2023.

[43] A. Stringer, G. Dolinger, D. Hogue, L. Schley and J. G. Metcalf, "A Metacognitive Approach to Adaptive Radar Detection," *IEEE Transactions on Aerospace and Electronic Systems,* vol. 60, no. 1, pp. 168-185, 2024.

[44] A. Stringer, T. Sharp, G. Dolinger, S. Howell, J. Karch and J. G. Metcalf, "Application of Generative Machine Learning for Adaptive Detection with Limited Sample Support," in *2024 IEEE Radar Conference (RadarConf24)*, Denver, CO, USA, 2024.

[45] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller and A. Holzinger, "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists," *Cognitive Systems Research,* vol. 86, p. 101243, 2024.

[46] S. Hariharan, R. R. R. Robinson, R. R. Prasad, C. Thomas and N. Balakrishnan, "XAI for intrusion detection system: comparing explanations based on global and local scope," *Journal of Computer Virology and Hacking Techniques,* vol. 19, pp. 217-239, 2022.

[47] Y. Wang, "A Comparative Analysis of Model Agnostic Techniques for Explainable Artificial Intelligence," *Research Reports on Computer Science,* vol. 3, no. 2, pp. 25-33, 2024.

[48] H. Qin, Y. Gu, W. Deng, D. Xu and G. Wang, "Self-interpretable Bayesian Rule Lists Model with Deep Prototype Layer," in *2021 7th International Conference on Big Data and Information Analytics (BigDIA)*, Chongqing, China, 2021.

[49] D. Song, J. Yao, Y. Jiang, S. Shi, C. Cui, L. Wang, L. Wang, H. Wu, H. Tian, X. Ye, D. Ou, W. L. b, N. Feng, W. Pan, M. Song, J. Xu, D. Xu, L. Wu and F. Dong, "A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM," *Computer Methods and Programs in Biomedicine,* vol. 235, p. 107527, 2023.

[50] D. Bhati, F. Neha and M. Amiruzzaman, "A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging," *Journal of Imaging,* vol. 10, no. 10, p. 239, 2024.

[51] T. B. Alexey Dosovitskiy, "Proceedings of the IEEE Conference on Computer Vision and Pattern," 2016.

[52] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. A. Riedmiller, "Striving for Simplicity: The All Convolutional Net," in *International Conference on Learning Representations*, 2015.

[53] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, 2014.

[54] G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition,* vol. 65, no. 1, pp. 211-222, 2017.

[55] M. Du, N. Liu, Q. Song and X. Hu, "Towards Explanation of DNN-based Prediction with Guided Feature Inversion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 2018.

[56] A. Henelius, K. Puolamäki and A. Ukkonen, "Interpreting Classifiers through Attribute Interactions in Datasets," in *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, 2017.

[57] D. Hogue, T. Sharp, J. Karch, G. Dolinger, A. Stringer and L. Schley, "Using Informative AI to Understand Camouflaged Object Detection and Segmentation," in *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, Barcelona, Spain, 2023.

[58] B. Lavender and S. Sen, "Model-Agnostic Policy Explanations: Biased Sampling for Surrogate Models," in *Explainable and Transparent AI and Multi-Agent Systems. EXTRAAMAS 2024*, Springer, Cham, 2024.

[59] Q. Zhang, X. Li, X. Che, X. Ma, A. Zhou, M. Xu, S. Wang, Y. Ma and X. Liu, "A Comprehensive Benchmark of Deep Learning Libraries on Mobile Devices," in *WWW '22: Proceedings of the ACM Web Conference 2022*, New York, NY, USA, 2022.

[60] Y. L. T. W. N. D. W. L. J. C. Haochen Hua, "Edge Computing with Artificial Intelligence: A Machine Learning Perspective," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1-35, 2023.

[61] Y. Shi, K. Yang, T. Jiang, J. Zhang and K. B. Letaief, "Communication-Efficient Edge AI: Algorithms and Systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167-2191, 2020.

[62] Z. Chang, S. Liu, X. Xiong, Z. Cai and G. Tu, "A Survey of Recent Advances in Edge-Computing-Powered Artificial Intelligence of Things," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 13849 - 13875, 2021.

[63] R. Desislavov, F. Martínez-Plumed and J. Hernández-Orallo, "Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning," *Sustainable Computing: Informatics and Systems*, vol. 38, p. 100857, 2023.

[64] J. Park, M. Naumov, P. Basu, S. Deng, A. Kalaiah, D. S. Khudia, J. Law, P. Malani, A. Malevich, N. Satish, J. M. Pino, M. Schatz, A. Sidorov, V. Sivakumar, A. Tulloch and X. Wang, "Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications," *CoRR*, vol. abs/1811.09886, 2018.

[65] M. t. A. E. o. N. Networks, "Danny Hernandez; Tom B. Brown," arXiv:2005.04305, 2020.

[66] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *North American Chapter of the Association for Computational Linguistics*, 2019.

[67] I. S. a. G. E. H. Alex Krizhevsky, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, Lake Tahoe, Nevada, United States, 2012.

[68] M. M. H. Shuvo, S. K. Islam, J. Cheng and B. I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," in *Proceedings of the IEEE*, 2023.

[69] K. Z. Ibrahim, T. Nguyen, H. A. Nam, W. Bhimji, S. Farrell and L. Oliker, "Architectural Requirements for Deep Learning Workloads in HPC

Environments," in *2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, St. Louis, MO, USA, 2021.

[70] Langerman, A. Johnson, K. Buettner and A. D. George, "Beyond Floating-Point Ops: CNN Performance Prediction with Critical Datapath Length," in *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, 2020.

[71] A. Asperti, D. Evangelista and M. Marzolla, "Dissecting FLOPs Along Input Dimensions for GreenAI Cost Estimations," in *Machine Learning, Optimization, and Data Science*, 2022.

[72] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, 2019.

[73] M. E. Paoletti, J. M. Haut, X. Tao, J. Plaza and A. Plaza, "FLOP-Reduction Through Memory Allocations Within CNN for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5938-5952, 2021.

[74] M. Zawish, S. Davy and L. Abraham, "Complexity-Driven Model Compression for Resource-Constrained Deep Learning on Edge," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3886-3901, 2024.

[75] P. Zulch, M. Distasio, T. Cushman, B. Wilson, B. Hart and E. Blasch, "ESCAPE Data Collection for Multi-Modal Data Fusion Research," in *2019 IEEE Aerospace Conference*, Big Sky, MT, USA, 2019.

[76] G. Farneback, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *SCIA'03: Proceedings of the 13th Scandinavian conference on Image analysis*, Berlin, Heidelberg, 2003.

[77] A. Vakil, E. Blasch, R. Ewing and J. Li, "Explainable Hybrid Decision Level Fusion for Heterogenous EO and Passive RF Fusion via xLFER," in *Prof. of 2022 9th International Conference on Soft Computing and Machine Intelligence*, Dayton, Ohio, 2022.

[78] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 9, no. 1, pp. 62-66, 1979.

[79] A. Vakil, R. Ewing, E. Blasch and J. Li, "Insights into Heterogenous Sensor Fusion and Passive RF Feature Extraction for xAI," in *Naecon*, Dayton, OH, USA, 2024.

[80] E. Wigner, "On the Quantum Correction for Thermodynamic Equilibrium," *Physical Review*, vol. 40, no. 1, pp. 749-759, 1932.

[81] J.-A. Ville, "Theory et Application de la Notion de Signal Analytique," *Cables et Transmissions*, vol. 20A, 1948.

[82] A. Grossmann and J. Morlet, "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 1, pp. 723-736, 1984.

[83] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425-434, 1991.

[84] D. Gabor, "Theory of communication. Part 1: The analysis of information," *Journal of the Institution of Electrical Engineers*, vol. 93, no. 26, 1946.

[85] S. S. Stevens, J. Volkmann and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 1, p. 185–190, 1937.

[86] Y. Chen, A. Savvaidis, S. Fomel, Y. Chen, O. M. Saad, Y. A. S. I. Oboue, Q. Zhang and W. Chen, "Pyseistr: a python package for structural denoising and interpolation of multi-channel seismic data," *Seismological Research Letters*, vol. 94, no. 3, pp. 1703-1714, 2023.

[87] B. P. Bogert, J. R. Healy and J. W. Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," *Proceedings of the Symposium on Time Series Analysis*, vol. 1, no. 1, pp. 209-243, 1963.

[88] W. Fu-sheng and L. Ming, "Cepstrum analysis of seismic source characteristics," *Acta Seismologica Sinica*, vol. 16, no. 1, pp. 50-58, 2003.

[89] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing and J. Li, "Feature Level Sensor Fusion for Passive RF and EO Information Integration," in *Proc. of 2020 IEEE Aerospace Conference*, Dayton, Ohio, 2020.

[90] E. Blasch, A. Vakil, J. Li and R. Ewing, "Multimodal Data Fusion Using Canonical Variates," in *Proc. of 2021 IEEE Aerospace Conference*, Dayton, Ohio, 2021.

[91] A. Vakil, E. Blasch, R. Ewing and J. Li, "Visualizations of Fusion of Electro Optical (EO) and Passive Radio-Frequency (PRF) Data," in *Proc. of IEEE National Aerospace and Electronics Conference*, Dayton, Ohio, 2021.

[92] A. Vakil, E. Blasch, R. Ewing and J. Li, "MVI-DCGAN Insights into Heterogenous EO and Passive RF Fusion," in *Prof. of 2022 9th International Conference on Soft Computing and Machine Intelligence*, Toronto, Canada, 2022.

[93] A. Vakil, E. Blasch, R. Ewing and J. Li, "Finding Explanations in AI Fusion of Electro-Optical/Passive Radio-Frequency Data," *Sensors*, vol. 23, no. 3, p. 1489, 2023.

[94] K. Gurumoorthy, A. Dhurandhar, G. Cecchi and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights," in *IEEE International Conference on Data Mining (ICDM)*, 2019.

[95] R. K. Mothilal, A. Sharma and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, NY, United States, 2020.

[96] M. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, NY, United States, 2016.

[97] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir and G. Menegaz, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems,* vol. Early View 2400304, 2024.

[98] Y. Wang, Q. Wang, S. Shi, X. He, Z. Tang and K. Zhao, "Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training," in *20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, Melbourne, VIC, Australia, 2020.

[99] D. Roy, Y. Li, T. Jian, P. Tian, K. Chowdhury and S. Ioannidis, "Multi-Modality Sensing and Data Fusion for Multi-Vehicle Detection," *IEEE Transactions on Multimedia,* vol. 25, no. 1, pp. 2280-2295, 2023.

[100] D. Garagić, D. Pelgrift, J. Peskoe, R. D. Hagan, P. Zulch and B. J. Rhodes, "Machine Learning Multi-Modality Fusion Approaches Outperform Single-Modality & Traditional Approaches," in *2021 IEEE Aerospace Conference,* Big Sky, MT, USA, 2021.

[101] B. Smyth and M. Keane, "A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations," in *International Conference on Case-Based Reasoning*, Nancy, France, 2022.

[102] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson and C. Shah, "Counterfactual Explanations for Machine Learning: A Review," *ACM Computing Surveys,* vol. 56, no. 12, pp. 1-42, 2020.

[103] E. Blasch, "Value-based sensor and information fusion," in *Proceedings Volume 13057, Signal Processing, Sensor/Information Fusion, and Target Recognition XXXIII*, National Harbor, Maryland, United States, 2024.

[104] E. Blasch, J. Dezert and B. Pannetier, "Overview of Dempster-Shafer and belief function tracking methods," in *Proceedings of Signal Processing, Sensor Fusion, and Target Recognition XXII*, 2013.

[105] T. Molom-Ochir and R. Shenoy, "Energy and cost considerations for GPU accelerated AI inference workloads," in *2021 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA, 2021.