Graduate Research Plan Statement

Introduction: Machine Learning (ML) is a pivotal technology for space applications, particularly for real-time geospatial analysis aboard satellites. ML is being used to study environmental and socio-economic destruction in conflicts[1], or assessing damages caused by natural disasters[2]. These tasks are time critical and require quick action to provide needed aid to affected areas. To conduct these assessments, researchers use multispectral Earth observation data. This data is collected by remote sensing satellites such as the Sentinel-2 satellites, but the data from these satellites are large and can take time to downlink observation prior to analysis on the ground[3]. This wasted time could have been used to analyze the imagery collected onboard the satellite and downlink analysis results to ground stations. This will allow for faster image processing which provides quicker decision making.

There are two focuses to attain reliable onboard inferencing: **(1)** radiation resiliency (fault tolerability), and **(2)** improvement of computation capabilities under low power systems. This is a challenging task as increasing radiation tolerability comes at the cost of decreasing processing power. Radiation in orbit is higher than the radiation on Earth, leading to faster component deterioration. Because of this, engineers design systems and ML models that are fault tolerable. For example, models are deployed with fixed-point instead of floating-point quantization because it is more resilient to radiation effects[4]. This comes at the cost of decreasing computation power, conflicting with the second focus for computation improvements. To test the radiation resiliency of satellite hardware, researchers have to irradiate the hardware while inferencing ML models in several testing campaigns. This type of experimentation permanently damages the system which renders it inoperable for further use, a time and cost consuming testing method. An alternative approach is to utilize fault injectors (FI). These systems simulate the effects of radiation by perturbing the model's parameters prior to inferencing, a much more cost effective and reliable form of radiation testing.

However, researchers in the fault injection field have not explored a wide range of different model architectures. **To address these challenges, I propose a novel framework that compresses and tests the fault tolerability of geospatial models for satellite deployments.** Existing literature has mostly conducted work on convolutional neural networks (CNNs) rather than transformer-based models[5, 6]. This is because CNNs have lower parameter counts that makes it possible to deploy onto satellites. Despite this feasibility, transformers attain better performance than their counterparts. With recent works building geospatial models on transformer-based architectures[7], it is crucial to compress and explore their fault tolerance to attain dependable satellite deployment and performance.

Research Plan: I propose **(Aim #1)** to create a pipeline that measures the fault tolerance of geospatial models over a range of simulated radiation dosages and across the entire computation pipeline. Subsequently, I will **(Aim #2)** evaluate several compression techniques on geospatial models to observe what methods are capable of maintaining performance and fault tolerance. These contributions will advance the state of space ML by **(1)** enabling models to be deployable onto satellites for onboard inference and **(2)** accelerate the utilization of improving hardware being integrated into newer satellites.

Aim #1, Fault-Tolerability Measurement and Inferencing Pipeline: The first step of this work is to construct a pipeline that simulates radiation effects holistically. The goal of this aim is to simulate radiation effects in the sensor, memory, and processor (instruction level) spaces. The construction of this pipeline will result in a library that can be used to simulate perturbation across the entire computation process. Firstly, to emulate radiation effects on satellite sensors, perturbations are applied to images, such as gaussian and salt-and-pepper filters[8]. For simulated effects on memory, there are two fault injector frameworks that this proposal can leverage: Ares FI[9] and PyTorch FI[10]. Ares FI focuses more on static faults where perturbations of parameters are permanent during inference. PyTorch FI is more focused on transient faults where perturbations are temporary. Lastly, to emulate the effects in the instruction level, I propose using Low-Level Tensor FI[11]. This FI perturbates instructions at the low-level virtual machine intermediate representation level, applying perturbations during instruction execution.

With these methods combined, this work will apply perturbations to these targeted areas during or before inferencing. This will give a holistic understanding of what areas of the entire inferencing process fail to maintain fault tolerability. This can be done in whole or in part of the entire processing pipeline. For instance, experiments can run with only one type of perturbation to help isolate areas where model design or hardware architecture is mostly vulnerable to radiation effects. The performance results from the perturbation campaigns are compared to the unperturbed models' performance to see what the fault tolerance is across a range of simulated radiation dosages.

Aim #2, Compression Effectiveness Study: Significant work has been established to compress models for high performance edge computing systems or smaller devices. In these areas, there are five main compression techniques: model pruning, parameter quantization, low-rank decomposition, knowledge distillation, and lightweight model design[12]. These methods provide several advantages, such as the ability to deploy into space vehicles with limited hardware and quicker inferencing time. Although parameter quantization and lightweight model design have been used in prior literature for deploying CNNs[13, 14], these techniques largely have been less commonly applied to transformer-based geospatial models. Thus, this aim applies these techniques to transformer-based geospatial models and utilizes the pipeline built in Aim #1 to test its fault tolerance. The goal is to provide further insights on the fault capabilities of compressed models under different perturbation rates. Furthermore, this work will be an extension to Aim #1's work, making a framework that compresses models and tests their fault tolerability.

Intellectual Merit: My proposed framework will **(1)** help shed light on how compressed geospatial models perform with perturbations in the input, memory, and processing space. Literature on fault tolerance have only looked at these three different areas one at a time, making it difficult to emulate space deployment without considering all other areas. My framework tackles this by simulating the entire process and provides insights on compressing models if they are too large for deployment criteria. Furthermore, my framework **(2)** enables researchers to iteratively build model architectures geared towards deployment conditions. Unlike radiation campaigns where iterative development is difficult due to the hardware being permanently damaged, this framework can allow researchers to iteratively develop models on the same hardware, undamaged. This can direct researchers to develop suitable models for mission specific deployments that are fault tolerable and small enough for onboard inferencing. Lastly, this framework is **(3)** applicable to areas outside of the geospatial domain. Models trained for space navigation and geospatial tasks are largely developed in PyTorch. Because the fault injectors in Aim #1 are applicable to PyTorch models, my framework can accommodate for various different space applications. This framework can accommodate space missions from earth observation to autonomous systems on extraterrestrial surfaces. Altogether, my work not only provides a system that tests and verifies whether ML models are deployable for onboard inferencing on satellites, but also enables more types of architectures of geospatial models to be deployed to these systems. With institutions integrating better computation hardware into satellites, my work will accelerate the effectiveness of space ML.

Broader Impacts: Being able to compress and inference fault tolerable geospatial models onboard satellites has major implications for public policy and rapid response to disasters. Because the data collected from the satellites is directly processed onboard, the results can be transferred to the ground, giving administrations faster time to react. This is especially true for countries undergoing conflicts. Governments can quickly direct populations to evacuate cities if there is a deployment of hostile troops directed to the area of interest. In times of natural disaster where important infrastructure (e.g., electricity and communication towers) is destroyed, regional governments can pull analyzed results from satellites. Government officials can make crucial decisions to direct emergency resources to affected areas where people are stranded, or are in desperate need of aid. The time difference between downlinking data and then processing it on Earth versus downlinking already processed results is the difference between life and death for people in these situations.

References: [1]Aung et al. *Environmental Research Communications.* 2021. [2]Shafian & Hu. *Buildings.* 2024. [3]Drusch et al. *Remote Sensing Environment.* 2012. [4]Hanif et al. *IEEE/ACM DAC.* 2019. [5]Mateo-Garcia et al. *Scientific Reports.* 2021. [6]Kain et al. *IEEE HPEC.* 2020. [7]Rolf et al. *ICML.* 2024. [8]Zhu et al. *Nuclear Engineering and Technology.* 2024. [9]Reagan et al. *55th DAC.* 2018. [10]Mahmoud et al. *50th DSN-W.* 2020. [11]Agarwal et al. *ISSRE.* 2022. [12]Li et al. *Computers.* 2023. [13]Traiola et al. *ETS.* 2023. [14]Chintalapati. *Advances in Space Research.* 2024.