Statement of Academic Purpose

Shreyas Kalvankar *M.S. Applicant*

My academic and professional journey in Machine Learning and Software Development has been deeply influenced by a strong passion for creating knowledge-based systems to tackle diverse challenges, frequently employing deep learning algorithms. I am interested in exploring the fundamental mathematics driving Machine Learning and deep learning.

I'm driven by questions regarding the theoretical foundations that explain deep networks' success in approximating complex functions and the potential inherent limits to their expressiveness. I am interested in delving into the dynamics of training deep networks and understanding the insights gained from this process. Moreover, I'm fascinated by the potential applications of learned representations across domains like computer vision, natural language processing, and speech recognition. I'm equally intrigued by the challenges surrounding the interpretability and explainability of learned representations.

My primary goal is to explore these theoretical quandaries and utilize this knowledge to develop advanced learning algorithms to contribute to research in physics and mathematics to solve problems such as approximating numerical computations using neural networks, astronomical simulations, assisted theorem proving, etc. and design systems capable of incremental learning and autonomous improvement. My unwavering passion for engineering and research is evident in my academic achievements, reflected in my transcripts and my active involvement in extracurricular activities during my undergraduate years. My journey has underscored the importance of further specialization and enrichment, which I believe a master's program can provide, laying the foundation for a future Ph.D. pursuit. In this regard, I am convinced that Carnegie Mellon University is the ideal place to nurture my potential.

During my bachelor's I was extremely fascinated by astrophysics and its intersection with computer science. In my junior year, I led a team of three to apply deep learning techniques to classify galaxy morphologies. Identifying galaxy morphologies has important implications in many astronomical tasks, e.g., studying galaxy evolution. The recent data influx in astronomy necessitates a robust and automated system for processing large amounts of images. We aimed to set up a 7-class classification system that classified galaxy images using CNNs, which can surpass existing benchmarks. I was responsible for establishing a robust training pipeline and optimizing various neural network architectures from the ground up. Our collective efforts outperformed the second-best submission on the Kaggle Public Leaderboard. This immersive journey provided a deep understanding of intricate, deep-learning methodologies rooted in a strong theoretical foundation. Dr. Yuanzhi Li's work in deep learning theory, particularly her paper on *'Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data,'* has been fascinating and insightful. Her research aligns well with my interests, and having her as my master's advisor would be immensely valuable for me.

The Galaxy Morphology Classification project proved to be a stepping stone in connecting astrophysics to computer science. Soon after that project, I began contributing to *EinsteinPy*, an open-source Python package designed to address issues in General Relativity and gravitational physics. My work specifically involved incorporating various symbolic computations, such as adding the Reissner–Nordström metric and the calculations for the event horizon and ergosphere of a Kerr-Newmann black hole. While symbolic computations are valuable in controlled environments, real-world problems often necessitate approximate rather than exact solutions. The potential for deep learning in simulating complex system behaviors is vast. However, the challenge remains—these problems often lack labeled data, making it difficult to apply modern ML algorithms. Dr. Maria Balcan's research, focusing on algorithms that learn effectively from limited data in collaboration with domain experts is extremely relevant to address these challenges and her work resonates with my interests and aspirations. Being mentored by her would be an incredible opportunity.

Following my graduation, I started working as an ML engineer, later transitioning into an ML consultant role at Relfor Labs Pvt. Ltd., an AI startup. For the past two years, I have been responsible for building novel neural architectures to classify audio data, represented as mel spectrograms, and setting up the ML training pipeline. I also built a data processing pipeline to build a proper dataset from the large database of raw audio. During this time, I encountered challenges related to data management including data storage and data pre-processing for training and inference, as well as problems in deploying and maintaining ML models, which were constantly plagued by issues like data drift and synchronization. For this very reason, The prospect of courses like *'ML with Large Datasets'* offered by Dr. Talwalkar and Dr. Gordon excites me as it systematically covers most of the practical problems that occur when dealing with large datasets and promises a comprehensive exploration of various techniques for scalable deep learning and low latency inference that extend beyond my previous academic and professional experiences.

The scientific landscape has been consistently evolving, but many longstanding problems have eluded complete solutions for many years. Looking into the future, I wish to make a difference in this significant era. My determination to pursue a research career is unwavering, and I aspire to work in an academic setting and eventually pursue a Ph.D. The vibrant community of students and researchers at CMU is the perfect place for me to nurture myself and further my commitment to help expedite this scientific advancement. Thank you for considering me as a prospective student at your university.