**Project Title**: Deep Regression Techniques for Decoding Dark Matter with Strong Gravitational Lensing

**Name:** Mehul Jain

**Email:** mehul.jain.codedev@gmail.com

**University**: Northeastern University

**Major**: MS in Data Science

**GitHub**: https://github.com/mehulfollytobevice

**Location**: Boston, Massachusetts

**Time Zone**: UTC/GMT -4 hours

**Mentors:**

- Michael Toomey (Brown University)
- Stephon Alexander (Brown University)
- Brandon Ames (University of Alabama)
- Sanaz Kiyadeh (University of Alabama)
- Yurii Halychanskyi(University of Washington)
- Saranga Mahanta (Institut Polytechnique de Paris)
- Karthik Sachdev (RWTH Aachen)

## Background: Why choose me?

Being someone who is profoundly interested in deep learning, I think this is a wonderful opportunity for me to get hands on experience in the field. I am also very eager to learn more about the field of computer vision over the summer and contribute to the open-source community of ML4SCI.

Currently, I am a graduate student pursuing my master's degree in Data Science from Northeastern University, Boston. I completed my undergraduate degree in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. Having a strong academic background in computer science and data science, I believe will help me in contributing to the success of this project.

In the past, I have worked on multiple research papers and open-source projects tackling problems in the domains of ML, DL, CV and NLP. Here are some of my relevant open-source projects:

1. **Galaxy-Morphology-Classification** – Classifying galaxy morphology using CNNs.
2. **image_captioning_flickr** – Image captioning system
3. **Deep-learning** – repository with a collection of micro-projects based on deep learning.
4. **MachineLearning**- repository with a collection of micro-projects based on machine learning.
5. **genre_classification**- A multi-label genre classification system.

One of my earliest passions has been Astronomy. I have always been fascinated by the wonders of the universe and the mysteries it holds. During my undergraduate years, I stumbled upon Machine Learning, and it was love at first sight. Since then, I have been eager to combine these two passions and explore the possibilities of using machine learning in the field of astronomy. In my previous semester, I developed a CNN based system to classify morphologies of distant galaxies using their images. This project was based on the Galaxy Zoo Challenge on Kaggle. The opportunity to contribute to ML4SCI aligns perfectly with my goal of exploring the intersection of these two domains. I am confident that I possess the necessary skills and experience to succeed in this program, and I am excited to work alongside experienced mentors to further develop my abilities.

## Proposal Abstract:

Strong gravitational lensing is a promising probe of the substructure of dark matter to better understand its underlying nature. Deep learning methods have the potential to accurately identify images containing substructure and differentiate WIMP particle dark matter from other well motivated models, including vortex substructure of dark matter condensates and superfluids.

This project will focus on further development of the DeepLense pipeline that combines state-of-the art of deep learning models with strong lensing simulations based on lenstronomy. The focus of this project is on using deep regression techniques for estimating dark matter properties, including population-level quantities and properties of dark matter particle candidates (e.g. CDM, WDM, axions, SIDM).

## Features and Deliverable Specifications:

Deep regression tasks involve predicting continuous output values based on input data. To find and tune the optimal model for deep regression tasks for DeepLense, we need to follow a systematic approach that involves the following steps:

1. **Data preprocessing:** Collect and preprocess the data to ensure that it is clean, properly formatted, and ready for analysis. This includes loading the dataset, applying transforms on the images, and creating batches which can be used to train the model.
2. **Model selection:** Choose a suitable model architecture that can handle the complexity of the data and generate accurate predictions.
3. **Hyperparameter tuning:** Fine-tune the model by adjusting its hyperparameters to optimize its performance on the data.
4. **Evaluation:** Test the model's performance on a validation set and select the best model based on its performance metrics.

Following are the features and deliverable specifications for this project.

## Features for the DeepLense pipeline and optimal regression model:

- Ability to handle large datasets with high-dimensional input features and continuous output values.
- Robustness to noise, missing data, and outliers in the input data.
- Ability to capture complex patterns and relationships between the input and output variables.
- Ability to generate accurate predictions with low error rates.
- The ability to provide insights into the model's decision-making process, such as feature importance and contribution to the output.

## Deliverables:

- An optimized deep regression model for estimating dark matter properties, including population-level quantities and properties of dark matter particle candidates (e.g., CDM, WDM, axions, SIDM).
- A DeepLense pipeline that integrates the optimized deep regression model with strong lensing simulations based on lenstronomy.
- Performance evaluation of the optimized deep regression model, including metrics such as mean absolute error, root mean squared error, and R-squared score.
- Error analysis of the optimized deep regression model to identify sources of error and provide insights for model improvement.
- A user-friendly interface that allows users to input data and generate predictions using the trained models.
- A technical report detailing the methodology, results, and findings of the project, along with recommendations for future work.
- A user-friendly documentation of the DeepLense pipeline and the optimized deep regression model, including instructions for installation, usage, and customization.
- A presentation summarizing the project's goals, methodology, results, and impact, suitable for technical and non-technical audiences.

## Project Timeline:

This is a tentative outline, and I have aimed to be as flexible as possible in allocating weekly tasks. Typically, the final 1-2 days of a week will be set aside for code review and documentation, or as buffer time for any incomplete tasks. My plan is to be regularly in touch with the mentors throughout the week to get some valuable feedback and to ensure that the progress is made in the right direction. I have designed the schedule to allow for regular feedback on code written thus far and plan code reviews at appropriate intervals.

Here's the timeline:

**May 29 - June 28 (4 weeks) (Initial setup, data preprocessing and research)**

- Week 1: Setup and familiarization with the DeepLense pipeline and dataset.
- Week 2: Data preprocessing, data exploration and data augmentation.
- Week 3-4: Research and experimentation with various deep regression models and techniques.

**June 28 - July 14 (2.5 weeks) (Model selection and optimization)**

- Week 5: Initial model selection and optimization.
- Week 6-7: Refinement and optimization of the selected model.

**July 14 - August 21 (5.5 weeks) (Integration, UI development and model evaluation)**

- Week 8-10: Integration of the optimized deep regression model with the DeepLense pipeline.
- Week 11-12: Performance evaluation and error analysis.
- Week 13: Technical report writing and documentation preparation.

**August 21 - August 28 (1 week) (overall review)**

- Week 14: Final week for submission of the final work product and mentor evaluation.

Note that this timeline is subject to adjustment based on the specific details of the project and the progress made during the work period.

## Summer Availability:

- Semester End Date: 01/05/2023
- Work Hours per week: Approximately 42 hours per week or the amount of time required to finish the tasks allocated for the week (whichever is higher).