Project and Learning Plan

I Intellectual Merit

The modern world is awash in data – from such an enormous variety of sources that it is hard to appropriately characterize the landscape without inadvertently under- or over-emphasizing some of its sections. In an interestingly similar vein, the essential challenge in the field of *deep learning*, which is at the core of this proposal, is appropriately characterizing a data set from a spectacularly sparse sampling of its underlying probability distribution. Rather than try to fully characterize the entire field of data science, we note the current interest in understanding data relating to the COVID-19 pandemic. Data are produced on positive test results, hospitalizations, intubations, deaths, etc., with relations to age, gender, socioeconomic status, race, location of residence, etc. All this data naturally raises the question – how to make sense of it? How to estimate underlying but hidden truths, and how to predict future events based on past experience. Although the field of statistical inference is hundreds of years old, its focus has largely been trained on the important, but limited, topic of Gaussian (or *normal*) distributions – i.e. data that obeys the bell-curve. While this, and other simple probability models, have proven useful, they are of limited value for many types of real-word data.

In contrast, the field of *machine learning* (which is what most modern commentators mean when using the term artificial intelligence) is focused on the design and training of models, rather than the analysis of simple parametric probability distributions. It is well known that there is no uniformly best model, or even best family of models, for all tasks and so the researcher's goal is to find and train a model that is useful for a given task. Ever since a watershed result from the University of Toronto in 2012, the preponderance of advances in machine learning have used *deep learning* models. Often the greatest impediment to applying deep learning is its need for large volumes of data that have been labeled with associated underlying ground truths, for use in model training. Recently, *generative adversarial networks* (GANs) were advanced, with the ability to produce unlimited quantities of realistic (though not labeled) data based on examples.

The proposed research will attempt to advance a promising new approach that we call PSG-GAN, for Parametrized Synthetic Generator GAN. Our work will target two very different applications: a robotic device for eliminating weeds from lawns, and non-destructive analysis of the health of rotating machines while in service. The first application area will involve extending our previous results in computer vision and the second involves a research collaboration with Dr. Mohammed Alabsi, a new member of the TCNJ Mechanical Engineering faculty. Our approach starts with an abstract physical model to produce artificially generated labeled data, and adds a GAN framework to "polish" the data. The synthetic data is polished in two ways – by tuning generator parameters, and by post-processing the data produced. The Generator/Discriminator framework allows us to produce artificial data that can best fool a discriminator trained to detect counterfeits. In this way we expect to produce labeled data for training that best leads to transfer of learning based on artificial sources, for ultimate use in analyzing natural, real-world, data.

Success in addressing our two target applications would produce direct benefits. The creation of machine learning algorithms for automated weeding of lawns would help to dramatically decrease the use of herbicides, thereby reducing pollution of groundwater, and protecting children and wildlife. The improvement of algorithms for predicting machinery failures can help prevent catastrophic disasters in equipment such as generators, motors, and airplane turbines, potentially reducing maintenance costs and saving lives. More generally, providing a viable alternative to expensive hand labeling of data would be a boon for applications of deep learning.

The faculty mentors have recently published nineteen works as journal articles and conference papers, many with undergraduate student researchers. We would expect the MUSE project described herein to result in a conference paper for Nithya at the IEEE Workshop on Advanced Imagery Pattern Recognition, in October 2021, and in a conference publication or journal article, under Dr. Alabsi, which would include Michael Franco-Garcia as a contributor.

II Role of Student and Mentor

Dr. Pearlstein and ECE students will work together in the Intelligent Media Processing Lab, which is housed in a recently renovated collaborative setting in Armstrong 124. As a new member of the Mechanical Engineering faculty Dr. Alabsi will be building out a machinery advanced analytics lab, which will be used for obtaining new real-world vibration measurements. Dr. Alabsi will focus on defining theoretical physics-based models of machine vibration, and will mentor students regarding their implementation, tuning, and verification. These models will be implemented via MATLAB scripts. Dr. Pearlstein will mentor students on the implementation and tuning of the PSG algorithm, which will "polish" the physics-based model data produced under Dr. Alabsi's tutelage, and also use the real-world data obtained by lab-based measurements.

The MUSE program will kick off with the faculty mentors introducing the students to the tools and components at their disposal. Mentors and students will work side-by-side. A proposed timeline for the summer experience is as follows:

Weeks	Tasks	Expected Actions
1	Orientation, literature review, lab administration	Students will come fully up to speed on the state-of-the-art in the technology areas to be pursued, and will have the ability to work with our lab equipment.
2	Experimental design	Students will define an initial set of experiments to carry out, and will organize the computer environment to record and back up all results in a methodical way.
3-4	Experimental implementation	Students will code scripts (using MATLAB, Python and TensorFlow) to carry out experiments, and will verify correct operation of their programs.
5-7	Carry out experiments	Students will run experiments and collect results.
8 and beyond	Data presentation and publication	Students will produce formal plots of results, and articles for presentation and/or publication.

III Broader Impacts

Nithya Nalluri, one of the proposed student researchers, is enrolled in the ECE Research Track, and participation in the MUSE program would provide her with a valuable opportunity to become immersed in her research work. Nithya is a female Computer Engineering student and is thus part of a group that is underrepresented in the field and in our program.

Michael Franco-Garcia, the other proposed student, has been working closely with Dr. Pearlstein to develop expertise in Linux system administration, Python programming and deep learning. He is highly motivated to pursue a career in deep learning, and this MUSE experience will help build his pool of skills and experience. In addition, Michael is a first-generation college student and is of Hispanic descent, thus belonging to a group that is both underrepresented in the field of Computer Engineering and within the TCNJ student body.

Funding of the proposed MUSE project would provide Dr. Pearlstein with an opportunity to re-engage fully with his research group after the interruptions caused by COVID. If this proposal is funded, Dr. Alabsi will hire an additional Mechanical Engineering student to work in his lab during the summer. This would boost Dr. Alabsi's lab, and help him to establish a thriving research program.

IV Collaborative Nature of the Project

The two faculty members have already started collaborative research, and are currently focusing on investigating the use of deep learning for the machinery diagnostics application, using public datasets. They are jointly mentoring two undergraduate researchers, who are testing the performance of improved deep learning models for machine fault detection and diagnosis, to be more robust to a variety of operating conditions.

The dearth of labeled data is a well-known problem in the machinery and health monitoring community. Hence, our research emphasis will focus on the generation of synthetic data using physics-based models, which be polished by GANs and ultimately used to train models for classifying real-world data. Dr. Alabsi will hire and advise an additional ME student to build a diagnostics simulator. Dr. Alabsi will evaluate different physics-based dynamic models to simulate different faults inside a ball bearing, and will supervise model implementation and parameter settings to generate labeled data. Dr. Pearlstein will mentor students to apply the PSG-GAN method to polish the generated data, and tune the system to enable effective transfer of learning based on artificial data, for the purpose of operating on real-world data. Dr. Alabsi will also hire and supervise an ME student, who will be responsible for the diagnostics simulator design and building. The simulator will consist of rotating machinery, planetary gear train, bearings and vibration and acoustics sensors. The design will allow TCNJ researchers to induce failures into different components to simulate real failure scenarios (e.g. cracked gear tooth or indent in bearing components). While the simulator is being assembled in the STEM workshop, Dr. Pearlstein and MUSE students will be working on model training and tuning. Once the simulator is fully operational, the generated deep learning models will be utilized for real-time classification and fault identification. Figure 1 shows a sample layout of the advanced analytics lab. Our ultimate vision is that Dr. Alabsi's simulator (Machine x) would be instrumented to stream vibration data into Dr. Pearlstein's NVIDIA Xavier embedded platform for deep learning, which would be configured with our jointly developed models for use in real-time classification.

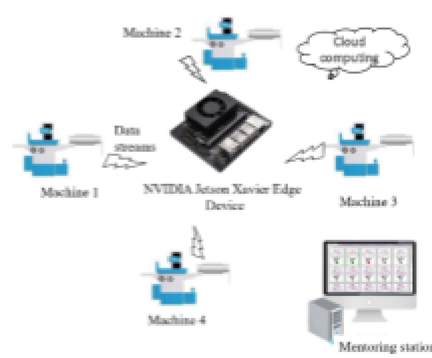

Figure 1: Advanced analytics lab layout