

# Statement of Purpose

Moseli Mots'oeahi

February 10, 2023

## 1 Background

I was born and raised in Lesotho. I studied Actuarial Science at the University of Pretoria. I then obtained a Master's degree with distinction in data science with a research dissertation titled "Comparison of Adversarial and Non-adversarial LSTM Music Generative Models", while at the same time working at Vodafone South Africa as a data scientist in the credit and risk department. For my dissertation at the university of Pretoria, I compared adversarial and non-adversarial music-generated music samples. During this period, I got to learn how to use deep learning libraries such as Tensorflow and Pytorch, and learned how to prepare data for different modalities such as audio, text, and images. It was through the work I did in natural language processing and summarization with Professor Jan Platos of the Technical University of Ostrava that I got a poster accepted at NeurIPS Black in AI 2018, where I met Professor Peter Sadowski, who would go on to be my advisor during my Masters's degree in Hawaii.

During my Master's at UH (2019-2021), I took all the AI and ML classes offered. I also took at least one class from the core computer science competency areas. I appreciated the creative freedom in designing and implementing our final projects in most classes. During my masters, a large part of my research with Professor Peter Sadowski was on building machine learning models for large-scale fish stock estimation using a dataset of 1.3 Million fish images belonging to 163 species. In this, all the software engineering skills from ICS314 came in handy: I was writing better code. I found that code repositories that used to intimidate me such as the Detectron2 code base from Facebook, didn't intimidate me anymore. Having worked on: audio, text, and image data, it was during this period I discovered, I want to spend the rest of my PhD working on computer vision, especially image classification.

## 2 Current and Future Research

In terms of research output since 2019, I have 3 first-author conference papers: (1 published in proceedings, 1 accepted for publication, 1 to be submitted), and 1 coauthor paper in NLP for low resource languages that is a potential journal submission in 2023. I have also been an active reviewer for the following conference venues: (1) Computing Conference 2023, and (2) Future Technologies Conference 2023 and (3) Intelligent Systems Conference 2023. My current research under the supervision of Professor Kyungim Baek, is centered around active learning algorithms for image data, in the presence of label noise. My inclination towards this area was inspired by the work I did with Professor Peter Sadowski on fish taxonomic classification and size estimation. Active learning is concerned with the development of learning heuristics that

allow the learning to select training samples it is trained on. This is done to ensure only the samples that maximize the model's performance are selected and labeled, thus we get the best model within a predetermined data labeling budget. However, data labeling is prone to human error for a number of reasons, hence the need for noise-robust active learning methods.

Existing literature in this space relies heavily on image datasets and the use of convolutional neural networks. I am particularly interested in seeing how the current advances in state-of-the-art visual transformers affect active learning algorithms on image classification with noisy labels. Although transformers have surpassed all convolutional neural network-based models on image classification benchmarks, the majority of the active learning literature is still anchored on convolutional neural networks. Having written a literature review in this area, it has become evident there is a need for more standardized evaluation methods for active learning algorithms on noisy labels. The reasons I chose this area are two-fold, firstly, Our work with Professor Sadowski granted us access to a 1.3 Million fish dataset, with noisy labels that has not been explored, and building tools around this specialized dataset would be highly impactful. Secondly, as the world becomes ever more so digital, datasets are likely to grow bigger and noisier, more and more car makers will enter autonomous driving, more and more companies will seek to build in-house image-based document readers, and many more applications in fields like precision health, manufacturing and mining safety. I would like to contribute towards understanding and building the best active learning algorithms for datasets with noisy labels.

### **3 Career Goals**

I see this research output as a viable B2B startup capable of raising venture capital and making an impact, given it is led by experts in the field. I want a PhD in this field because it is intense, rigorous and will likely mean I am an expert in my area once all requirements are met. I hope to eventually launch and grow a startup based on the technical and research skills I gain during my PhD. The other reason I want this degree is to someday much much later, be able advise students, and teach a course or two at a university in southern Africa or my home country Lesotho. There is a huge shortage of professors in the STEM fields, and all the capable students would rather take high-paying industry jobs after their undergraduate studies than consider graduate studies. I like being around students, they are full of exciting ideas and still have hope in their eyes. It is fulfilling to observe a student's progress, even as a TA.