SUMMER RESEARCH 2024/25
PROJECT ABSTRACT

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

PROJECT # 24

SUPERVISOR/S:	Dr Han Gan, Jason Kurz & Tim Edwards
PROJECT TITLE:	Predicting lung cancer from breath samples
FIELD:	Data analytics
DIVISION/SCHOOL:	HECS - Au Reikura School of Computing and Mathematical Sciences
PROJECT LOCATION:	Hamilton

PROJECT ABSTRACT:

Previous research studies have shown that dogs can sometimes identify whether a patient has lung cancer based only upon smell. The pertinent question is therefore, what exactly are the dogs are identifying as markers for lung cancer? To further explore this, samples of breath were collected from patients, some with lung cancer, some without lung cancer, and a detailed chemical analysis of the breath samples was undertaken with Gas chromatography/Mass spectrometry (GCMS) instrument. We currently have two samples each from around 300 patients with approximately a third of these patients having a confirmed lung cancer diagnosis. There may be further samples collected by the start of this project. The primary goal of this cross-disciplinary project is to attempt to identify key features in the GCMS data that are associated with lung cancer. As taking breath samples is a non-invasive test compared to other approaches such as lung biopsies, it is hoped that this can be eventually used as a diagnostic tool for lung cancer.

STUDENT SKILLS:

- Data analysis knowledge
- Data wrangling skills
- Machine learning knowledge

PROJECT TASKS:

1. Familiarise themselves with the data set.
2. Clean and format the data set so it is ready to be analysed.
3. Build a variety of classification models, using both statistical and machine learning models.
4. Compare and contrast the classification models.
5. Map the key features of the spectrometry data identified by the classification models to known chemical compounds through GCMS software.
6. Create a final research poster summarising the findings.

EXPECTED OUTCOMES:

- Student's Research Poster (as per clause 6 of the Scholarship regulations)
- An organised and cleaned data set.