DEEP LEARNING OF INFRARED SPECTRA FOR BOILING POINT PREDICTION

1. Introduction

Knowledge of boiling point is of importance in tasks such as finding out a chemical's vapor pressure and hence toxicity. However, boiling points are not always available in literature, and are expensive and time consuming to measure. Hence, computational methods are of interest.

Three main issues with current boiling point estimation are:

1	It requires knowledge of the structure of a compound.
2	It requires a pure sample of the compound, which is not always available.
3	To derive the structure of a compound, specialized apparatus is required for analysis, and this can take up to half an hour.

Current computational methods include the OPERA (OPEn (quantitative) structure-activity Relationship Application) model for boiling point proposed by Mansouri et al, which utilizes molecular descriptors generated using the PaDEL software. This model faces the first issue - molecular descriptors can only be generated for a known structure. The model has an RMSE of 22.08.

Contributions:

1	Created a model which predicts boiling points from IR spectra (which can be quickly and efficiently taken)
2	Showed than CNNs perform significantly better than conventional machine learning models and MLPs

2. Methodology

Project Pipeline:

Curating Datasets → Boiling points, molecular descriptors

Baseline model → Takes in molecular structure, generates a molecular descriptor, and predict the boiling point, based on machine learning principles.

Models using IR spectra will be compared to the baseline model to find out if using IR spectra is feasible for boiling point prediction.

Curating Datasets → Boiling points, IR spectra

Random Forest → Hypothesize worse performance than baseline

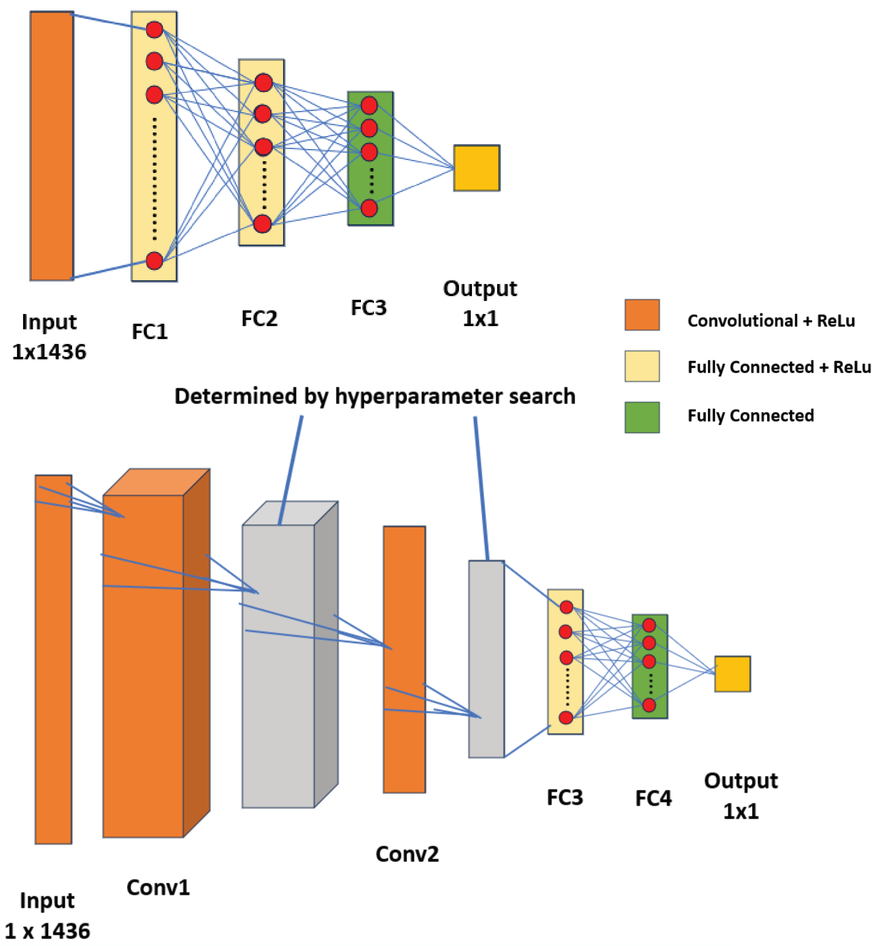
XGBoost → Hypothesize XGBoost will do better as it was shown online (e.g. Kaggle competitions) to do well in various tasks

FFNN → Hypothesize FFNNs better capture links between different peaks and substructures

CNN → Hypothesize that with differently sized filters convolved over the spectral input, CNNs identify differently sized peaks in the data and outperform FFNNs

3. Model Architecture

Feed Forward Neural Network (FFNN) architecture:



Convolutional Neural Network (CNN) architecture:

4. Results and Discussion

Data	Molecular descriptors		IR spectra			
Model	OPERA	Baseline	Random Forest	XGBoost	FFNN	CNN
RMSE	22.08	20.79	58.99	56.02	47.60	42.41

1. Is it feasible to explore the use of IR spectra for prediction of boiling points?

- CNN has a higher RMSE than baseline and opera models but shows promise for boiling point prediction in scenarios when the chemical is unknown and hence molecular structure and descriptors are unknown, rendering models like baseline and OPERA inapplicable.

- Lower accuracy could be due to IR spectra data being more prone to noise from experimental conditions or contamination.

2. Do CNNs outperform FFNNs in handling spectral input?

- **C**NNs outperform FFNNs in handling spectral input as it can leverage differently sized filters to learn patterns during the convolution process.

- Hence, CNNs can be explored for models with spectral inputs.

Further work

- Contrastive Learning to encode implicit molecular information in the model
- More data and data augmentation for other physical states of chemicals than just gas.

Member:
Ching Yuhui Natalie, River Valley High School
Mentors:
Mr Alvin Liew, DSO National Laboratories
Mr Chong Yihui, DSO National Laboratories
Dr Chiew Hai Leong, DSO National Laboratories

YDSP Young Defence Scientists Programme
DSTA Defence Science & Technology Agency
DSO NATIONAL LABORATORIES