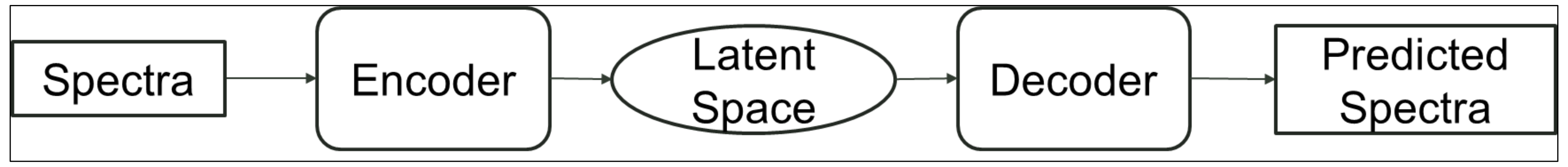
UNSUPERVISED SPECTRAL COMPRESSION

Patrick C. Riley[1], Samir V. Deshpande[1], Kyle P. O'Donnell[2], Ruth Dereje[2], Brian S. Ince[1], Charles S. Harden[2], Brian C. Hauck[1], and Mary M. Wade[1]
[1]U.S. Army Combat Capabilities Development Command Chemical Biological Center, Aberdeen Proving Ground, MD, [2]Science & Technology Corporation, 21 Enterprise Pkwy, Suite 150, Hampton, VA 23666

Introduction

Dimensionality reduction and machine learning (ML) based feature selection improves the accuracy of classification tasks by helping models generalize to relevant features for prediction. This down-selection of features has an added advantage of reducing the size of data needed to make complex ML based predictions. In our recent work, random forest (RF) feature selection was used to build a model to generate ion mobility spectrometry (IMS) spectral features to improve classification of a long short-term memory (LSTM) based neural network (NN). This work investigates an unsupervised deep learning (DL) algorithm called a variational autoencoder (VAE). VAE's use two nearly identical neural networks to encode or compress data into a latent representation of features, then decode this data into the original value as shown in figure 1.

VAE BLOCK DIAGRAM



Spectra → Encoder → Latent Space → Decoder → Predicted Spectra

Figure 1. Block diagram of simple encoder/decoder model, showing spectra as input, latent space as the encoder output, and predicted spectra as the decoder output.

Comparing the loss of the encoder and decoder allows the VAE model to learn without a need to label the data. VAE's have shown promise in a number of fields to include synthetic data generation, drug design, and image generation.

Problem

Compact IMS based chemical detectors are used to detect the presence of gas phase chemical hazards. These detectors operate in complex environments often leading to overlapping chemical signatures and thus false alarms. Variance in IMS detectors increases the alarm window size and complicates the development of detection algorithms. Traditional approaches rely on acquiring huge amounts of data acquired from multiple instruments at varied conditions. This drastically increases the time and cost of fielding chemical detectors. Figure 2 depicts the frequency of occurrence for an IMS based detector, showing how peaks for methyl salicylate (MeS) vary. This work hypothesized that a VAE trained on IMS spectral images will learn spectral features and could be used to generate spectra. Here we demonstrate a VAE trained to compress IMS spectra to a two-dimensional latent space. The resulting latent features are examined by are comparing to features generated by principal component analysis (PCA).Comparison of the accuracies predicted by several classification models demonstrates the power of a using a VAE latent features over PC's as inputs for ML models.

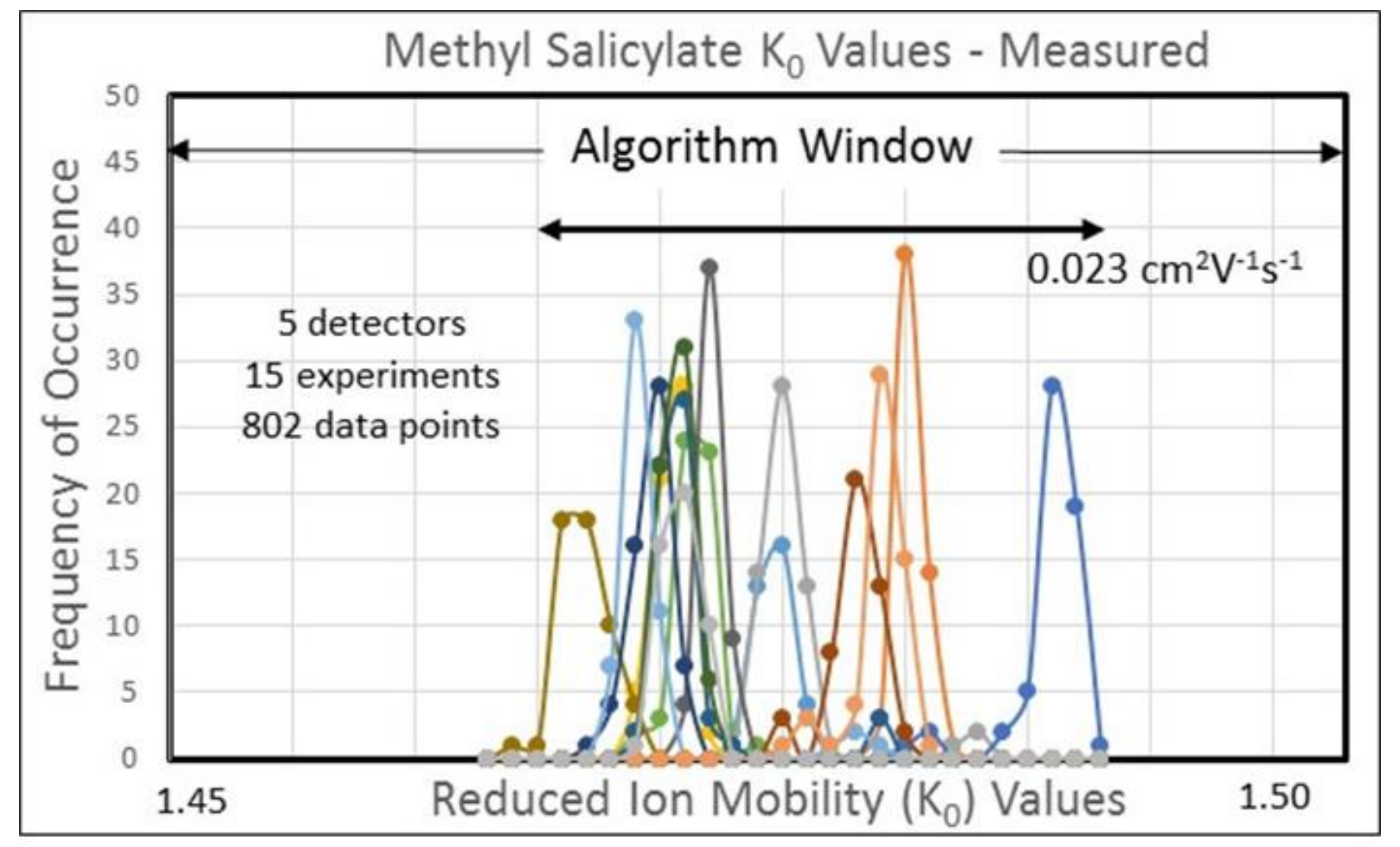


Figure 2. Frequency of occurrence of MeS peak position across five IMS detectors.

DEVCOM CBC @ DTRA CBD S&T Conference
Scan the QR Code to view all of CBC's
2022 DTRA CBD S&T Conference materials
https://cbc.devcom.army.mil/cbdst-conference/

Methods

The IMS training data set consisted of 1000 positive (+ve) and negative (-ve) detection mode spectra from 9 chemical classes and 1 background (BKG) class totaling 10000 spectra. Figure 3 shows a typical +ve mode spectrum for di(propylene glycol) monomethyl ether (DPM), where the three Gaussian shaped peaks represent: the reactant ion peak (RIP), the monomer peak, and the dimer. The features are represented by a measurement of current expressed as amplitude (y-axis) at intersecting coordinates of reduced mobility (K0) (x-axis). This allowed common autoencoder techniques, regularly used to learn important features and representations from images, to work with this dataset.

+VE MODE DPM SPECTRUM

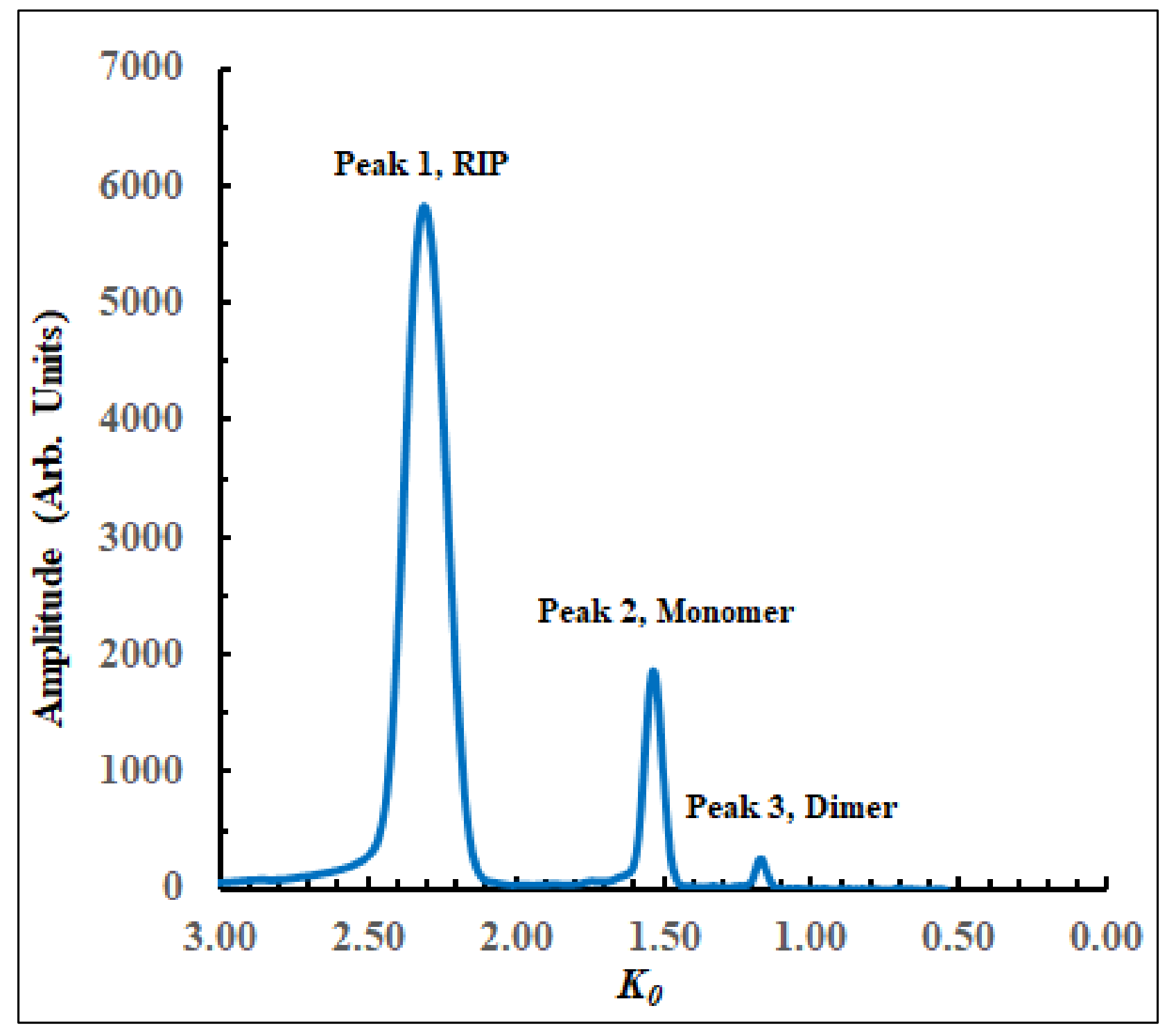


Figure 3.Typical DPM +ve mode spectrum

HEAT MAP OF DPM COMBINED +VE/-VE SPECTRUM

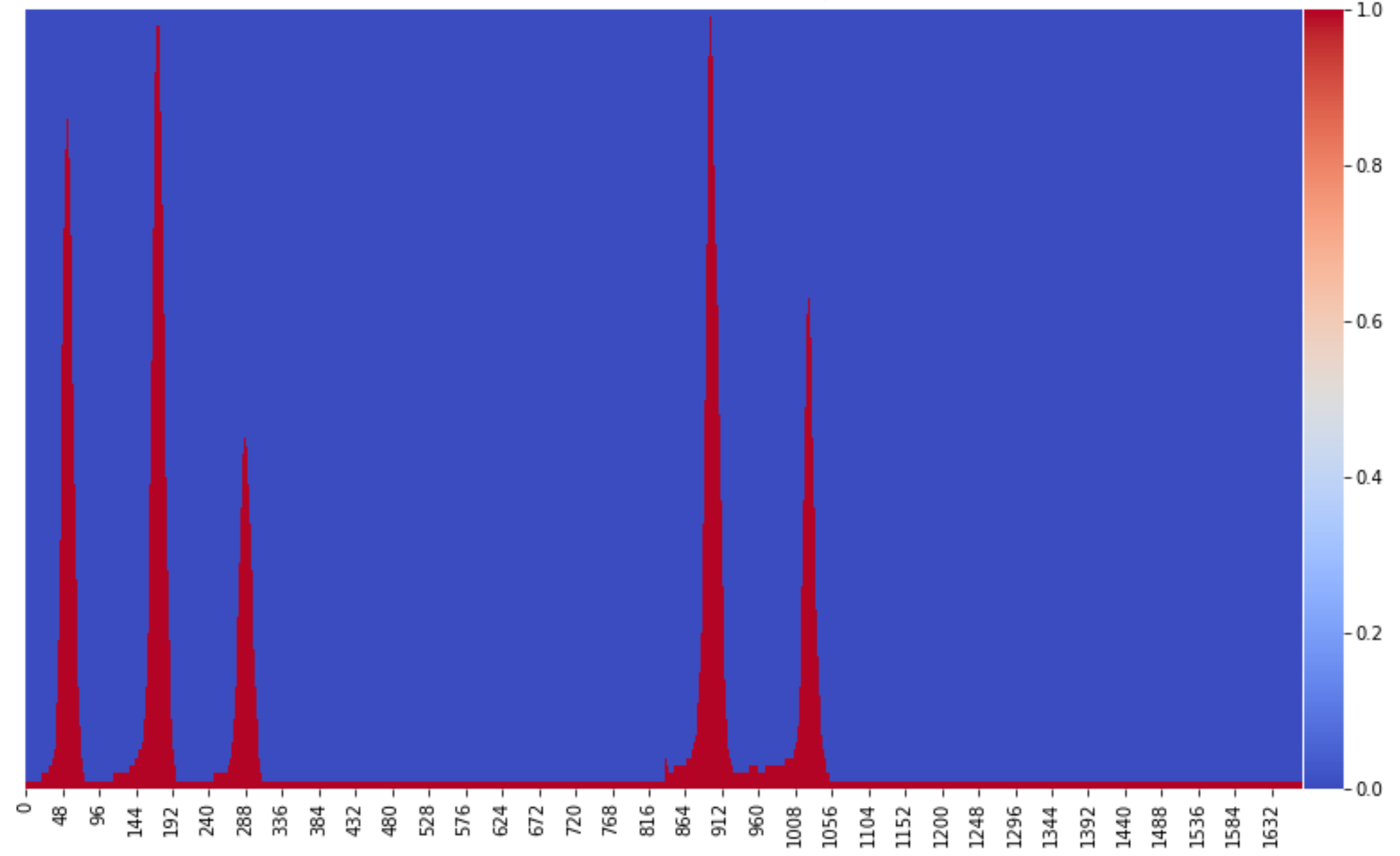


Figure 4. Heat map of DPM +ve and -ve spectrum converted to 2D vector

These features are converted to spectral bins representing an ascending count of K0 bins. The features for each spectra are normalized individually into a range of [0,1], then expanded to a 2D vector of binary, where the area under the curve of each spectra is represented as a 1, and area above the curve of the spectra represented as a 0. Figure 4 demonstrates the final result, depicting a heat map of 0's, 1's representing the combined +ve and -ve spectra for DPM.

While training the VAE was used to generate a latent feature space of 2 dimensions. The decoder used binary cross entropy (BCE) to calculate the reconstruction loss and it minimized at ~150 epochs. Figure 5 shows the log-loss over all epochs for the train and test data. Figure 6 depicts a predicted DPM spectrum, showing a heat map of probable locations for peaks. The reconstruction loss combined with the predicted spectrum demonstrate a VAE that successfully converged.

BCE RECONSTRUCTION LOSS OF DECODER

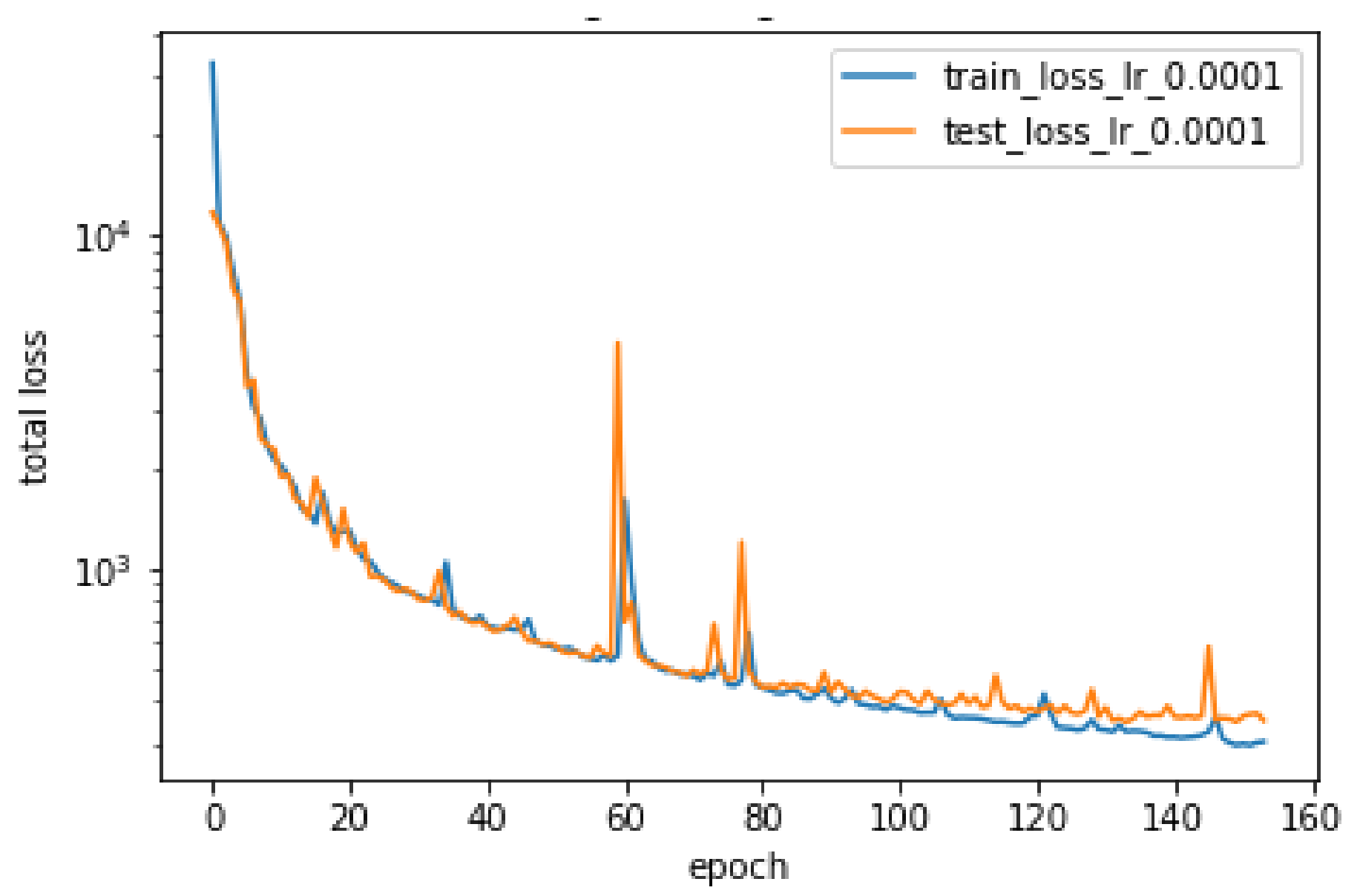


Figure 5. Log-loss of decoder calculated using BCE, with a learning rate of 0.0001.

HEAT MAP OF PREDICTED DPM +VE/-VE SPECTRUM

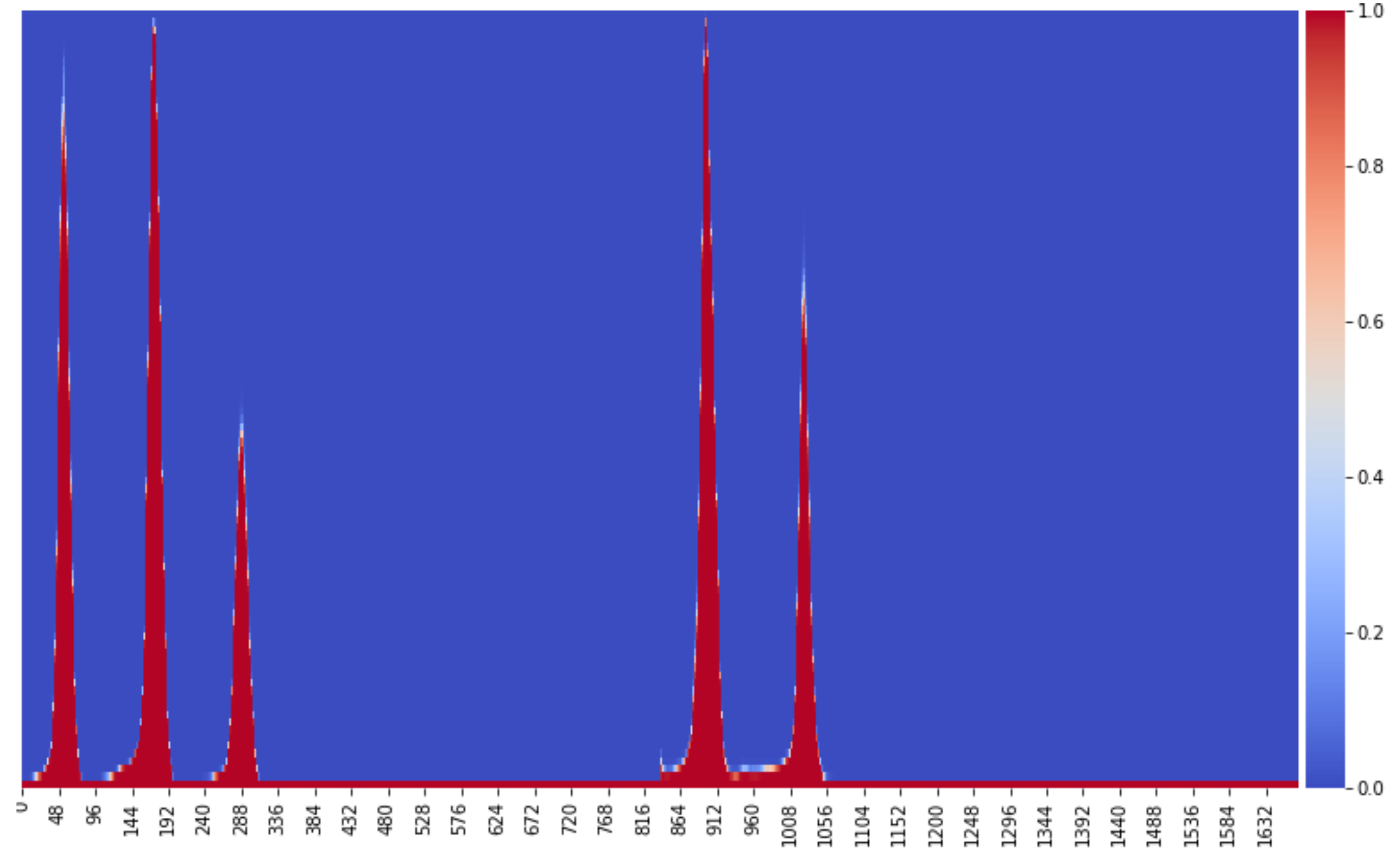


Figure 6. Heat map of predicted DPM +ve and -ve spectrum showing most probably peak locations.

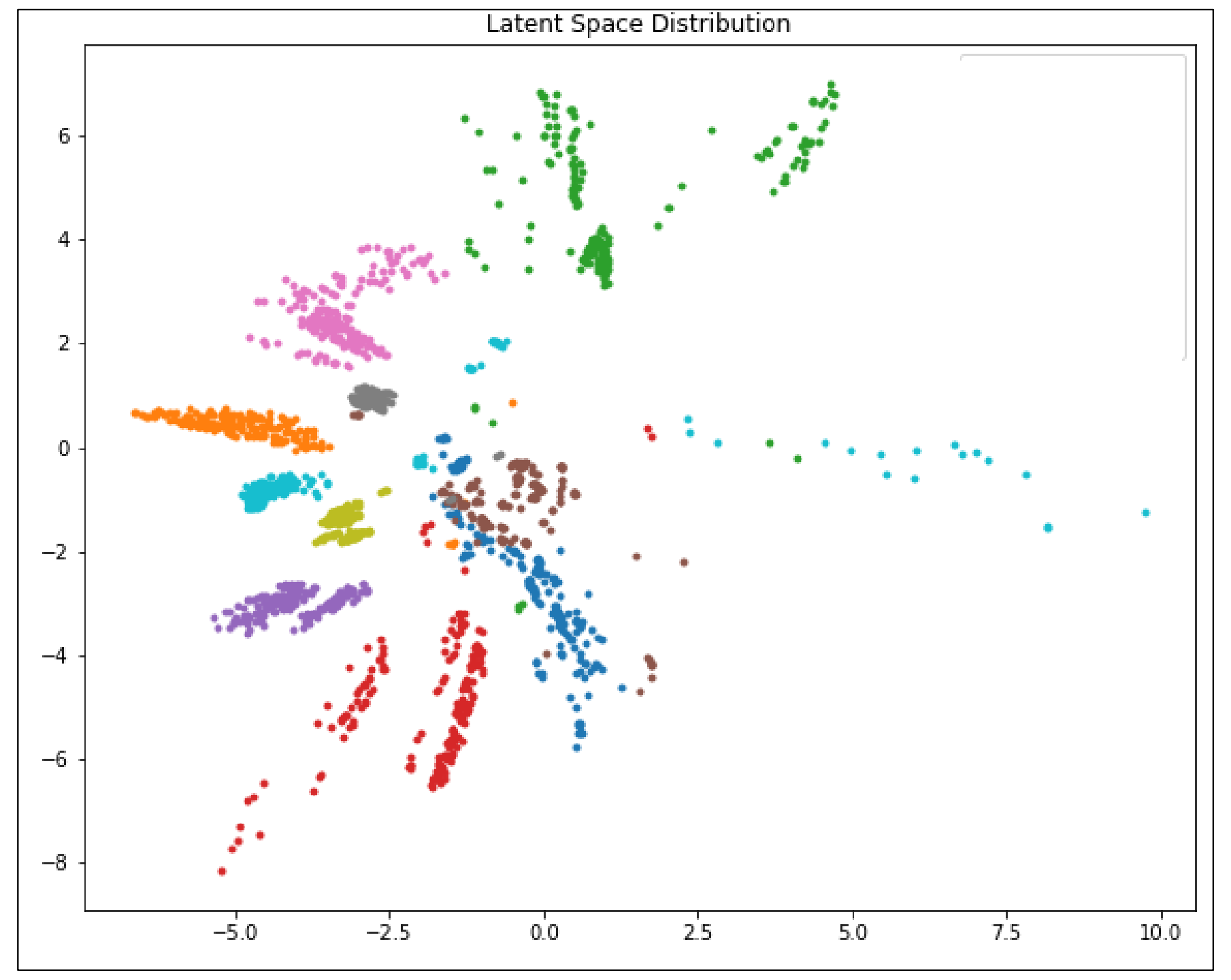


Figure 7. VAE latent feature space for 10 class problem.

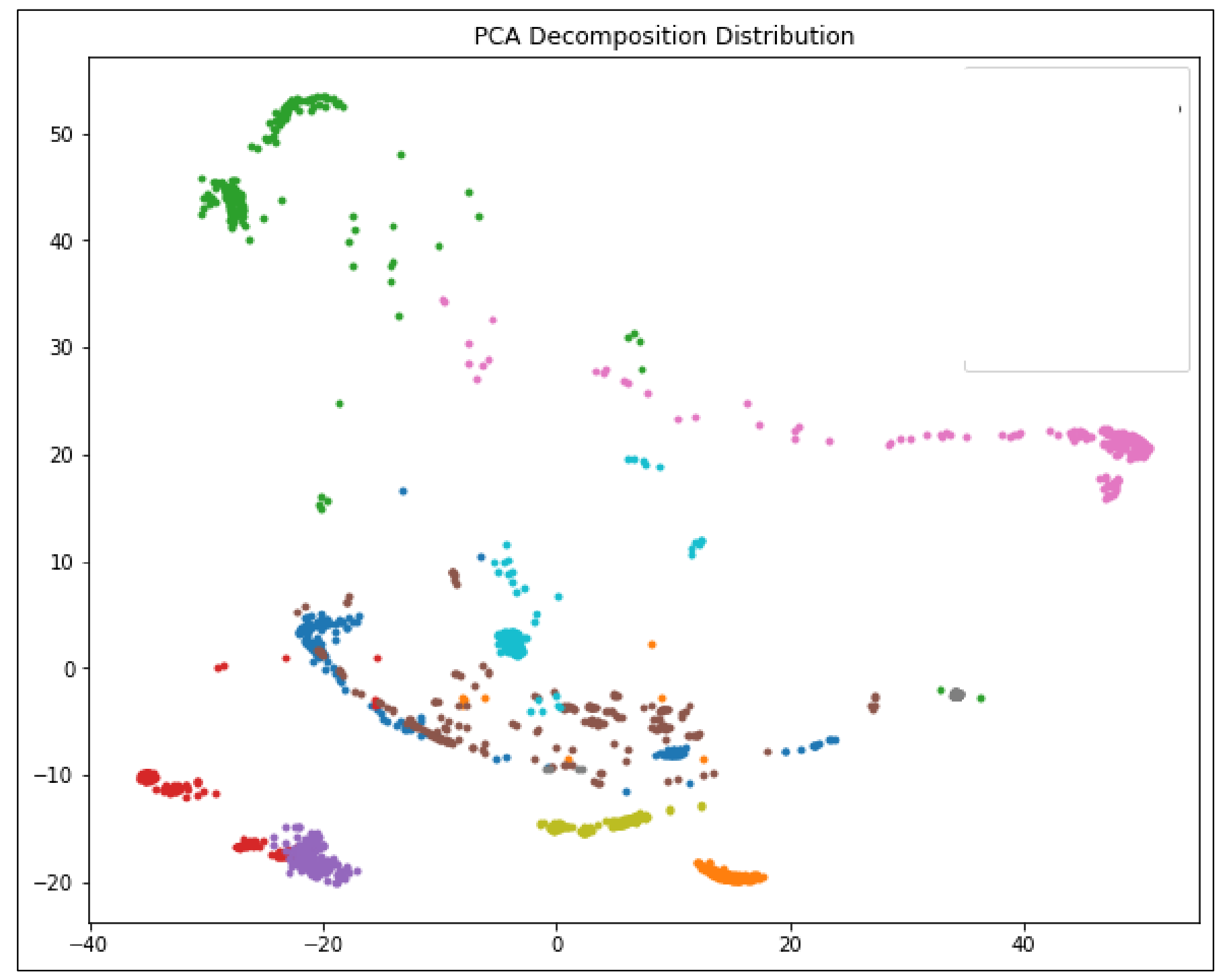


Figure 8. PCA distribution space for 10 class problem.

Results

The resulting 2 dimensional latent feature space is shown in figure 7. The encoder was trained using unlabeled data, as such each feature is selected by an unsupervised approach. Thus, the trained encoder theoretically could be used to generate features for any subsequent IMS spectrum of similar shape.

The distribution of the latent space was compared to another common unsupervised feature selection technique Principal Component Analysis (PCA). The distribution of the PC's is shown in figure 8 and can be visually compared to figure 8. To compare the feature selection techniques, various classifiers selected from the scikit-learn package were trained and tested on the VAE features, PCA features, and all spectral features.

The results shown in figure 9 that the VAE selected feature(s) have improved classification accuracy over PCA based feature selection. Additionally the 2 features of the VAE latent space outperform all features with 5 out of 8 class

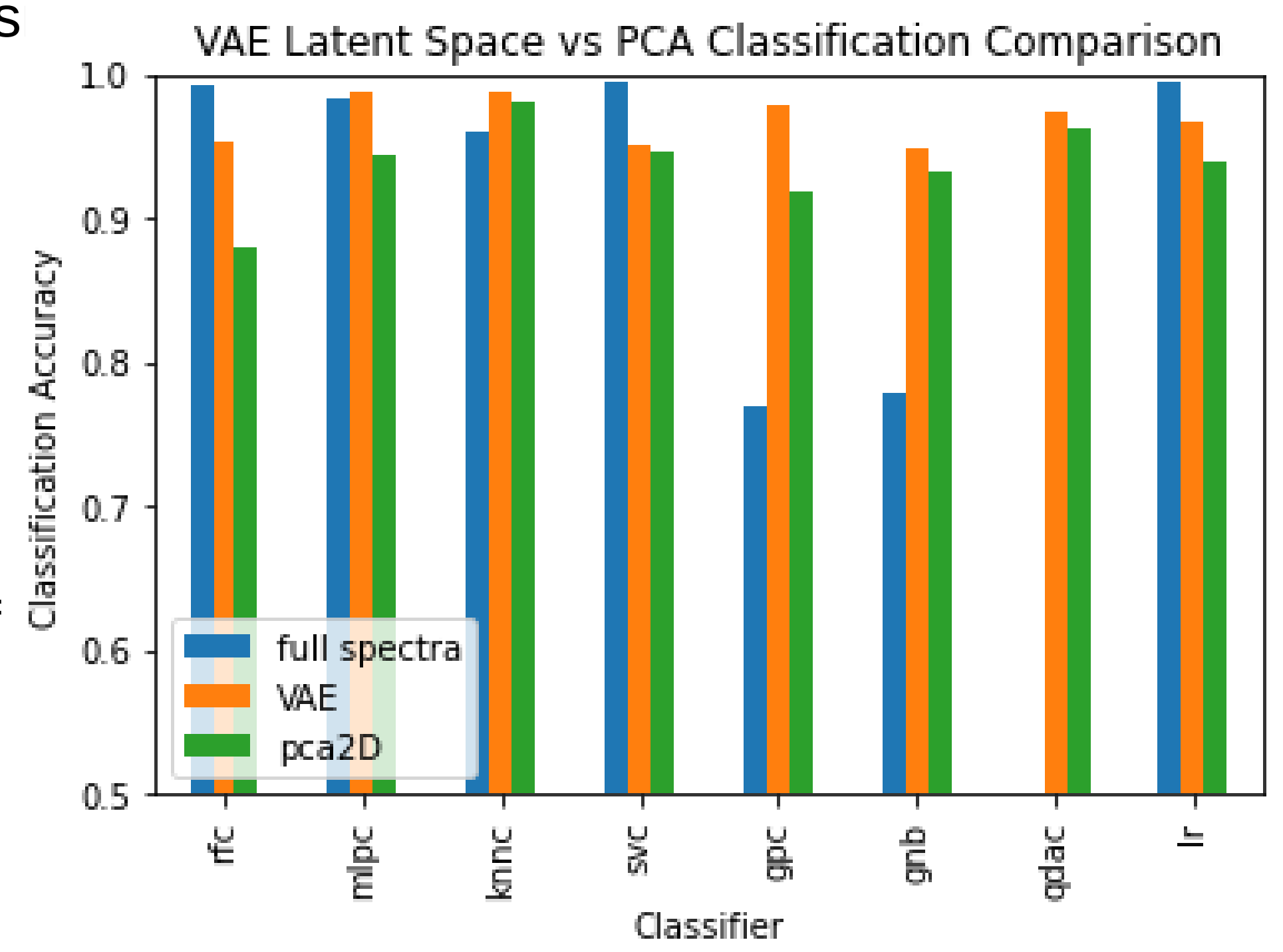


Figure 9. Comparison of classifier accuracy on trained on the full spectra (blue), VAE Features (orange), and PCA features (green).

Conclusion

Including our data preparation steps, this method of feature selection reduces the original +ve and -ve mode spectrum of 1676 features down to only 2 features. The similarity to accuracy in classifiers with the full spectrum feature list and low probability range of a predicted spectrum, shows that the compression is relatively lossless. Future work would seek to incorporate this type of VAE model into generating synthetic IMS spectra.

Acknowledgements: Funding for Unsupervised Spectral Compression was provided by the Defense Threat Reduction Agency, Research and Development Directorate, Chemical and Biological Technologies Department (CB10792, Dr. Tsai Chia-Wei, Ms. Nicole Farrall). The authors would like to express our thanks to those nameless scientists who contributed to the data collection over the many years of IMS product development. We also extend our gratitude to Charlie Davidson for his intellectual contributions made in the development of this VAE mode. The views expressed in this poster are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.