Utilising Graph Databases and Proper Cross-Validation Approaches to Improve the Machine Classification of Exosomes

Edward Bolger and Stan Goodwin

Dublin City University

Abstract. Current studies into the use of machine learning for the classification of disease, based on the Raman spectra of exosomes have claimed near perfect classification accuracy. In this paper we explain a simple validation error that can result in dramatically overestimated model performance. We then outline some novel spectral processing techniques utilising graph databases and feature selection, that improve the actual generalisation ability of these models.

Keywords: Cross-Validation · PageRank Outlier Detection · Feature Selection · Exosomes · Raman Spectroscopy.

1 Introduction

1.1 Exosomes

Traditional diagnosis of cellular diseases such as cancer, diabetes, hyperglycemia and hypoglycemia can be a complex and invasive task. Cancer diagnosis is typically carried out using tissue biopsies, where tissue is extracted from a suspected tumour and analysed in a lab [1]. This process is time consuming and expensive and can potentially pose risks to the patient. Likewise, current approaches for early diabetes screening are reliant on blood tests, which reduces the likelihood of patients receiving optimal treatment plans [2]. As a result, there is a significant drive to discover faster, less invasive methods for disease screening.

Exosomes are microscopic extracellular vesicles involved in cell-to-cell communication. They contain proteins, DNA, RNA and other biological material of the cells that secrete them [3]. Exosomes are found within almost all bodily fluids, including blood, saliva and urine [4]. Exosomes can be used as biomarkers which indicate the existence of disease within their parent cells [2]. Due to their presence within common bodily fluids, exosomes gathered via non-invasive, liquid biopsies provide a great advantage over traditional tissue biopsies for disease diagnosis, saving time and reducing the stress on patients. These liquid biopsies of exosomes have been used for the diagnosis of diseases such as prostate, pancreas, breast, and ovarian cancers as well as diabetes and hyperglycemia [5, 6, 7].

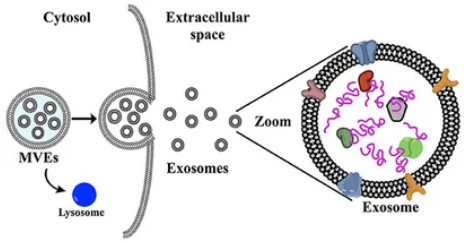

Fig. 1. The Emission of Exosomes from a Cell [8]

1.2 Raman Spectroscopy

As a result of their microscopic size (40–150 nm) and presence within complex biological samples, exosomes require specific techniques in order to be measured and analysed [9]. One method that is particularly effective at measuring the structure of exosomes is Raman Spectroscopy. Raman is a non-destructive vibrational spectroscopic technique based on inelastic photon-scattering [10]. A monochromatic light, usually a laser, is used to excite the chemical bonds that make up a particle causing it to emit radiation. A small fraction of the scattered radiation is observed to have a slightly different frequency from that of the incident light. The observed positive frequency shift, known as the Raman shift, is used as the x-axis in a Raman spectrum.

The emitted Raman spectrum features a number of peaks, showing the intensity and wavenumber position of the Raman scattered light. Each peak corresponds to a specific molecular bond vibration, giving every substance a unique Raman spectral "fingerprint" that describes its shape and composition. This fingerprint can be used for the identification of a wide range of organic and inorganic materials. Surface-Enhanced Raman Spectroscopy (SERS) is an advanced approach that places the material on roughened metal surfaces, causing a large enhancement to the strength of the Raman scattering signal.

1.3 The Goal of the Project

This project focuses on using machine learning for the multi-class classification of exosomes derived from normal, hyperglycemic, and hypoglycemic cells, based on their Raman spectra. The dataset, compiled by John O'Sullivan for his master's thesis [11], consists of 3,045 Raman spectra. The dataset contains the fields SpecID, WaveNumber, Absorbance, SurID, and Status, outlining the intensity of each spectra across wavenumbers from 200 to 2000 cm^{-1}. The dataset describes 63 unique SERS surfaces, and categorises the spectra into 915 hyperglycemic, 1065 hypoglycemic, and 1065 normal samples.

As the spectral peaks correspond to specific exosomal biomarkers [11, pp.112-121], our initial approach utilised peak profiling and graph databases to create new features that better represent the distinguishing characteristics of each

spectra. Model performance was evaluated using both random shuffling of samples with K-Fold cross-validation and more robust surface-based splits using Group K-Fold cross-validation. To better identify representative spectra from each SERS surface, we developed a novel spectral filtering process utilising PageRank centrality. Finally, dimensionality reduction was carried out through a forward sequential feature selection process.

2 Related Work

2.1 Machine Learning and Raman Spectroscopy

Most studies into using machine learning with Raman spectra tend to follow the same set of spectral preprocessing steps; first a despiking algorithm is used to remove artefacts caused by cosmic rays, then background fluorescence is removed through baseline correction, which leaves behind the underlying Raman signal, the spectra are then smoothed with a noise removal filter, and normalised to remove the effect of the overall intensity of the spectrum [12, 13, 14]. To reduce the impact of the curse of dimensionality, where large feature spaces harm classifier performance, feature set reduction techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) or partial least squares (PLS) are employed [15]. Then the machine learning task is carried out with a traditional algorithm such as a Support Vector Machine (SVM) [14, 16].

One paper by Shin et al. [17], explores the use of machine learning to classify the Raman spectra of exosomes. In the study they used a Residual Neural Network to classify the Raman spectra of exosomes derived from normal and lung cancer cell lines. As in the previously mentioned papers, the spectra were baseline corrected, denoised and normalised. Neural networks have the added advantage of discovering inherent features within the data on their own, so they do not rely on specialised feature engineering techniques. Model training and evaluation was done using an 80-20 split of shuffled exosome spectra. The approach also achieved 94.8% 5-Fold cross-validation accuracy. According to the authors, the relationship between the training accuracy and the 5-Fold cross-validation scores indicates that "the optimized model was free of the overfitting".

A similar investigation by Xie et al. [18], used an Artificial Neural Network (ANN) and the SERS spectra of exosomes to detect 4 different types of breast cancer. The same preprocessing steps were carried out and the data was randomly separated into an 80-20 training-validation split. The ANN reached an accuracy of 95% outperforming a PCA-SVM model which achieved 81.3% and a convolutional neural network with 69.2% accuracy.

2.2 Utilising Graph Representations with Machine Learning

Graph databases have been successfully used to develop additional features for machine learning in relation to disease classification. In a paper by Alqaissi, Alotaibi and Ramzan [19], they were able to use a knowledge graph constructed

on COVID-19 literature to train a random forest model to detect COVID-19 based on symptoms. This graph-based random forest model outperformed other models that used the same dataset. Furthermore, they found that incorporating Fast Random Projection (FastRP) node embeddings as features, further improved the performance of the model. According to the researcher, the study "demonstrates that graph algorithms support extracting essential features from the COVID-19 dataset".

Graph structures have also been used to classify Raman spectra. Another study by Wang et al. [20] utilised graph representations of Raman spectra to perform classification of oil paper. In order to construct a graph from the spectral data, they took advantage of the fact that the wavenumber intervals between intensity data points remained unchanged across every spectrum, this allowed them to represent each spectrum as a 1023-dimensional vector. They calculated the Euclidean distance between different spectra samples before passing this through a Gaussian kernel function to get a similarity measure between each spectra. Based on this similarity metric, they constructed a graph where each node is a spectrum that is connected to every other spectrum and the weight of the edge between two spectra is the similarity score. Using this structure they constructed a Graph Convolutional Network that was able to classify spectra with accuracies as high as 95.5%.

3 Methodology

3.1 Spectra Cleaning

As was seen in the other studies, before any classification approaches can be carried out, the spectra should be cleaned, removing background noise that can drown out the Raman signal. The four main steps are cosmic spike removal, baseline correction, smoothing and scaling.

Despiking was carried out using an algorithm developed by Whitaker and Hayes [21] to remove the effect of cosmic ray interference from the spectra.

Baseline Correction was performed using the asymmetric least squares algorithm developed by Eilers and Boelens [22]. Asymmetric least squares calculates the baseline using "asymmetric weighting of deviations from the (smooth) trend". It has the advantage of working without any prior knowledge of peak shapes.

Noise Removal was implemented using a Savistky-Golay filter. Savitsky-Golay filtering is a digital signal processing technique that can reduce high-frequency noise in a signal [23].

Normalisation the final step, was done to remove the impact of the absolute Raman intensity across each spectrum. The absolute intensity of a spectrum within our dataset intensity is related to the quantity of the exosome within the sample, rather than the underlying composition of the exosome [11]. Scaling also removes spectral variance caused by other external factors such as fluctuating laser intensity [24]. Three normalisation approaches were compared, scaling to the max peak, scaling to the vector norm of the spectra and Standard Normal Variate (SNV) scaling.

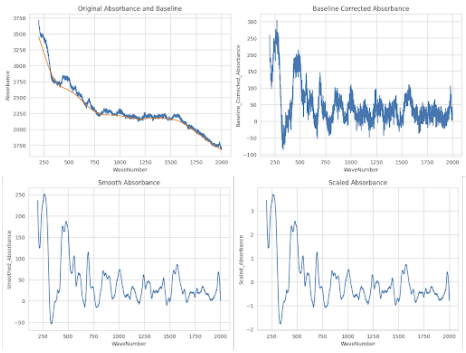

Fig. 2. The Spectra Cleaning Process Using Standard Cleaning Parameters

3.2 Model Evaluation Strategies

To evaluate the performance of our models we made use of 10-Fold cross validation. Cross-validation ensures the entirety of the dataset is used for performance evaluation and can help to measure the generalisation performance of the model by estimating how it will perform on unseen data [25].

However, in certain contexts randomly shuffling the data with K-Fold Cross-Validation can be misleading. A study by Guo et al. [26], explores the dangers of using K-Fold cross-validation with Raman spectra. In the study they compare both K-Fold and K-Replicate cross-validation against a held out test set to compare how each approach estimates their model's actual generalisation ability. K-Replicate validation accuracy was a much closer reflection of the test accuracy, while the K-Fold performance was drastically over-exaggerated. They state "k-fold CV doesn't reflect the real applications, where the dataset to be predicted is usually a different replicate compared with the training datasets". The researchers reach the conclusion that when training machine learning models to work with Raman spectra, the data should be split at the highest hierarchical level, such as separate biological or technical replicates in order to get an unbiased estimate of the generalisation error.

Based on these findings we decided to evaluate our models with both K-Fold cross-validation and Group K-Fold cross-validation using the SERS surfaces as groups. Group K-Fold, like K-Replicate cross-validation, ensures that spectra from the same surface are not contained within both the training and validation folds.

3.3 Baseline Approach

In order to create a baseline approach to compare the graph representations with, we created feature sets using both the full spectrum and peak feature sets based on peak properties.

Full Spectrum: Representing the full spectrum involved pivoting the dataset, so that each wavenumber becomes a separate column with the absorbance as its value. The 2,049 wavenumber values are then used as input features for the machine learning algorithm.

Peak Features: The reduced peak feature sets were developed using the *find_peaks* method in *SciPy* [27]. One problem faced was the varying number of peaks per spectrum. Machine learning algorithms generally require a uniform input feature set. The first method we used to create a uniform peak feature set involved calculating statistical properties such as the mean, standard deviation, upper and lower quartiles and the max and min values of the peak intensity, prominence and width within a spectrum. An alternative approach segmented the spectra into wavenumber intervals of equal length. The intensity, prominence and width of the peaks within each interval were recorded, aggregating the values if multiple peaks occurred.

3.4 Graph Feature Engineering

Another way of reflecting the non-uniform nature of the peaks used graph databases. Centrality algorithms were run on a variety of graphs constructed in Neo4j through the Python API [28]. Centrality was used to measure the importance of nodes within the networks. These algorithms consisted of PageRank, Degree, Eigenvector and Article Rank centrality [29, 30, 31, 32]. In addition to this, we ran a node embedding algorithm to capture high-dimensional aspects of the graph structure in a dense, lower dimensional vector [33]. The node embedding algorithm ran was FastRP [34].

Peak Grids, the first approach used every peak in each spectral sample as a node. Nodes were connected if two peaks were either in that same spectrum or if two peaks were close together as their proximity should indicate a common biomarker [10].

Biomarker Ranges was the next graph structure. Here we used spectral peak ranges of significance detailed in the thesis [11, pp.112-121] and shown in figure 3, to build a knowledge graph. There were two types of nodes in this graph, *PeakRange* nodes which represent specific biomarker regions outlined in the thesis. As well as *Spectra* nodes that represent each spectral sample.A relationship called *HasPeak* would exist between a *Spectra* node and a *PeakRange* node if a spectral sample has an identified peak in that WaveNumber range. The weight attached to the relationship would be the scaled absorbance of that peak.

Gaussian Kernel, the final approach attempted to replicate the graph constructed in Wang et al. [20]. This approach, as described in related works, uses each spectra as a vector and calculates the similarity between every spectra

Raman Shift (cm^{-1})	Corresponding Spectrum on Figure 5.13	Presumed Origin	References
402-562	1-10	Glycogen Fingerprint Region.	85.
510-550	3, 4, 5, 6, 9, 10	ν (S-S) in protein.	11,87.
618-624	1, 2, 5, 8	aromatic ring tw (C-C), most likely due to F.	89-93.
634-795	1-10	Nucleotide structure(s).	89.
808-812	3, 4, 5, 9	Phosphodiester (Z-marker).	95.
844-861	1, 2, 3, 8, 9	Polysaccharide structure/ Protein/ Carbohydrate fingerprint region.	117.
1001-1006	1-10	aromatic ring ν (C-C), usually Amino Acid F.	91,99,101.

Fig. 3. Example of presumed biomarker origins of peaks in different wavenumber ranges [11, pp.112-121]

through a Gaussian kernel function. This creates a fully connected graph with the Gaussian kernel function as a relationship weight between each node.

Centrality scores and node embeddings were calculated using each of the outlined graph structures and were then used as input features for machine classification.

3.5 PageRank Filter

When visualising spectra from the same surface, we noticed a larger amount of spectral variance than expected. Each surface contains the same type of exosome, so in theory the Raman spectra should be very similar. In figure 4, you can see a clear overlap of several spectra that seem to represent the true underlying Raman fingerprint. However, you can also see several samples that deviate from what seems to be the representative spectra, possibly representing interfering cosmic rays or surface noise. We wanted to check if these samples were negatively impacting the performance of our model and started investigating methods for removing these outlier samples.

Upon further investigation, we discovered a common step during the spectra gathering process involves collecting multiple spectral readings at the same point on a surface, and then averaging the readings to reduce background interference [12, 13, 35]. Based on the thesis it appears that this step was not undertaken during the collection of the spectra in our dataset leading to some of the samples being noisy and unrepresentative of the overall surface. We began to experiment with ways of extracting the relevant spectra from each surface.

To find outliers we first used the interquartile range of the spectral intensity within each surface and then defined an outlier as any value outside 1.5 times the range. Spectra were dropped if a chosen proportion of their values were defined as outlier. We systematically checked a range of proportion cutoffs. This method caused a slight improvement in performance, which led us to investigating more outlier removal techniques.

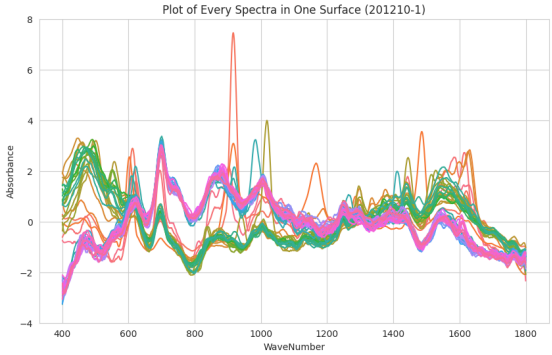

Fig. 4. An Example of Normalised Spectra from the Same Surface with a Dense Collection of Representative Spectra and a few Noisier Outliers

We then tried an alternative approach using the Gaussian Kernel method described earlier. Using spectra that had been normalised using SNV scaling in order to ensure that spectra of the same shape were given a high similarity, even if their absolute absorbance was different, a Gaussian kernel subgraph was created separately for each surface. PageRank centrality was then calculated for each spectra node in these subgraphs. The idea was that the more central spectra in each surface would better represent the underlying Raman signal while spectra with a low centrality would indicate noisy or unrepresentative samples. This idea can be visualised in figure 5, where the most and least central spectra in a surface are compared. We then introduced a PageRank cutoff to drop any spectra from the dataset below a certain centrality threshold.

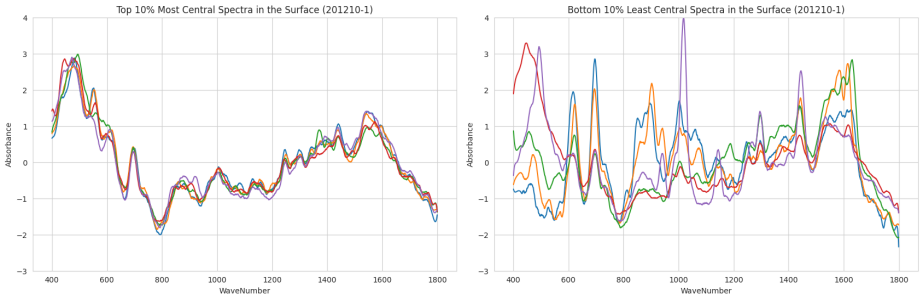

Fig. 5. The Most and Least Central Spectra Within a Surface Based on PageRank

3.6 Dimensionality Reduction

Several feature extraction and selection techniques were implemented in order to combat the curse of dimensionality. PCA, LDA and limiting the wavenumbers to the regions of importance outlined in the thesis were tried first, but each of these approaches yielded worse results. However, passing the models through a forward sequential feature selection process, where the wavenumber which contributed the most to performance was selected one at a time resulted in a reduced feature set and greatly improved model performance. Due to the extreme computational cost of checking every feature combination, we limited this search up to the 50 most influential features.

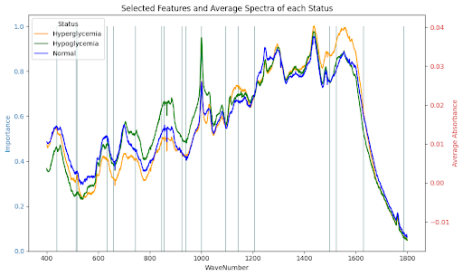

Fig. 6. Average Spectra by Status and Wavenumbers Selected by Sequential Feature Selection

4 Results and Discussion

4.1 Preliminary Results and the Discovery of Data Leakage

A variety of algorithms were tested with each approach. Consistently, the best performing models were Random Forest, Extra Trees and Support Vector Classifier. Initial performance, evaluated using 10-Fold cross-validation, was suspiciously high with accuracies of up to **95.5%**. To verify these models we plotted the feature importance of each wavenumber to see the spectral regions that were the most important for classification, this can be seen in figure 7. This indicated that the models were mainly utilising regions at either end of the spectrum which was odd, considering that these were completely unmentioned in the thesis.

Upon inquiry, we were told that the large peak at 250 cm^{-1} was caused by the Raman scattering of the background substrate and were advised that the analysis should be limited to the 400 to 1800 cm^{-1} regions, as these areas describe the core exosome biomarkers. Limiting the spectra to this range caused a slight degradation to performance, which made us concerned that the model had been using surface information instead of the exosome signal. To account for this possibility we implemented GroupK-Fold cross-validation, as in Guo et al. [26].

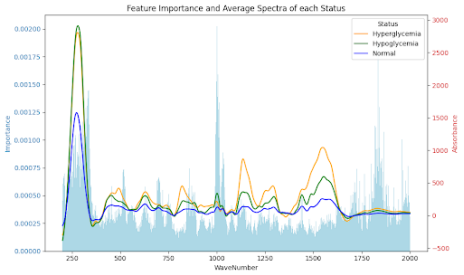

Fig. 7. Average Spectra of each Status alongside Cross-Validated Model Feature Importance.

Optimal cleaning parameters were initially found using a grid-search, separately for both K-Fold and Group K-Fold. Table 1 showcases the best results for each approach.

Table 1. 10-Fold cross-validation Results

	KFold		GroupKFold	
Approach	Accuracy	F1 Score	Accuracy	F1 Score
Full Wavelength	0.9366	0.9357	0.6122	0.5948
Peak Stats	0.6355	0.6312	0.4168	0.3892
Peak Bins	0.8407	0.8390	0.5519	0.5314
Peak Grids Centrality	0.7113	0.7092	0.4837	0.4701
Peak Grids FastRP	0.6509	0.6484	0.4533	0.4332
Biomarker Ranges Centrality	0.4810	0.4784	0.4010	0.3813
Biomarker Ranges FastRP	0.6126	0.6101	0.4106	0.3879
Gaussian Kernel Centrality	0.4956	0.4943	0.3573	0.3411
Gaussian Kernel FastRP	0.7218	0.7201	0.4403	0.4257

The Group K-Fold results were striking, depicting a stark drop in model performance across every approach, aligning with the aforementioned study. K-Fold cross-validation placed spectra from the same surface in both train and test folds, which leads to the model "cheating" by learning background information describing each surface instead of the exosome biomarkers from the Raman spectrum. When this leakage is prevented via Group K-Fold the model has no prior exposure to the surface information in the testing folds and has to rely on its knowledge of the Raman spectral structure of the exosomes instead, which results in dramatically worse performance. This splitting strategy is a much closer simulation of how the model would actually be used in practice, where new unseen surfaces of exosomes are introduced to be classified.

Given the performance of each approach in the preliminary results, we decided to utilise the full spectra as model input going forward as it was by far the most performant method under both validation strategies.

4.2 Final Results

With the true ability of the models unveiled, we searched for ways to improve the Group K-Fold performance. We first performed a wider cleaning and model parameter search using a Bayesian approach through *Optuna* [36], optimising for both accuracy and F1-score. We then implemented the PageRank filter and ran another Bayesian parameter search, this time with the PageRank threshold included. Finally, the feed forward feature selection was performed on the PageRank filter approach. Table 2 details the macro-average metrics of these processes.

Table 2. Group 10-Fold cross-validation Results Tuned for Accuracy and F1 Score

Approach	Best Model	Accuracy	Precision	Recall	F1 Score
Optimal Cleaning	Random Forest	0.6352	0.6410	0.6481	0.6207
PageRank Filter	Extra Trees	0.6963	0.6088	0.6440	0.6059
Feature Selection	Extra Trees	0.7595	0.6750	0.7078	0.6751
F1 Optimal Cleaning	Random Forest	0.6308	0.6358	0.6454	0.6158
F1 PageRank Filter	Random Forest	0.6643	0.6673	0.6727	0.6401
F1 Feature Selection	Random Forest	0.7202	0.7109	0.7213	0.6990

As seen above, introducing the Bayesian search alone was able to marginally increase all of the performance metrics. Although it also highlights the fact that the probabilistic nature of Bayesian searches can cause them to miss optimal parameter combinations, as the accuracy search found parameters that yielded a higher F1-score than the F1 search.

When tuned for accuracy, the PageRank filter approach's accuracy increased but all of the other metrics dropped. This is because the PageRank filter is increasing the class imbalance, and the accuracy model seems to be prioritising the majority Hypoglycemic class, as seen in table 3. However, when tuned to F1-Score, the PageRank filter does successfully raise all of the metrics, indicating that the benefits it provides are not solely caused by class imbalance. This backs up our idea that the PageRank filter is identifying unrepresentative samples.

Table 3. Best Accuracy Model Performance By Class

Class	Precision	Recall	F1 Score
Normal	0.6396	0.5735	0.5907
Hyperglycemic	0.6737	0.7758	0.7097
Hypoglycemic	0.7118	0.7741	0.7250

Finally, feature selection seems to successfully reduce the curse of dimensionality as it yields large increases to every performance metric. When investigating the features selected by this process, we found that 70% of features, when tuned

for accuracy, and 65.71% of features, when tuned for F1-score, occur within the ranges of significance outlined in the thesis [11, pp.112-121]. Given that these ranges make up 52.22% of the entire spectrum, the feature selection process seems to be slightly predisposed to the important spectral bands.

To further validate the effectiveness of our approaches, we tested their surface classification ability. SERS surfaces were classified based on the most frequently predicted status of their spectra. This shows the effectiveness of the PageRank filter at extracting the relevant spectra, while also removing the impact of the increased class imbalance, as no surfaces are dropped.

Table 4. Performance when predicting Surfaces by their most frequent Status

Approach	Accuracy	Precision	Recall	F1 Score
Optimal Cleaning	0.6825	0.7002	0.6850	0.6870
PageRank Filter	0.8254	0.8379	0.8309	0.8237
Feature Selection	0.8571	0.8726	0.8636	0.8568
F1 Optimal Cleaning	0.6667	0.6834	0.6699	0.6712
F1 PageRank Filter	0.6825	0.6976	0.6826	0.6784
F1 Feature Selection	0.7778	0.7818	0.7831	0.7758

5 Conclusion

From our experiments, we found that performing K-Fold cross-validation when evaluating a model to classify Raman spectra samples leads to overly optimistic performance metrics due to data leakage between spectra from the same surface. Performing Group K-Fold validation on the highest hierarchical layer of the data more accurately evaluates how the model will perform when exposed to new unlabeled data. This validation mistake appears to have been made in two prominent papers into machine classification of exosomes by Shin et al. [17] and Xie et al. [18], where spectra from the same source were shuffled and used to claim close to 100% predictive performance.

The PageRank filter was successfully able to identify representative spectra in each surface. This can be validated by visually inspecting that data as well as by the observed performance increase of the models when this filter is applied. Additionally, forward sequential feature selection successfully reduced the curse of dimensionality and improved classification performance.

It is impossible to discount the idea that our models are still overfitting to specific laboratory or spectrometer information, as each sample was sourced from the same machine. More samples acquired under different conditions would be needed to properly evaluate how this pipeline would classify new independent data. As well as this, future work could perform nested cross-validation to further validate our PageRank filtering and feature selection process, providing a more trustworthy evaluation of our pipeline's generalisation ability.

Acknowledgments. We would like to thank our supervisor, Prof. Mark Roantree, and his PhD student, Thomas Keogh, for their helpful discussion and guidance throughout this project.

References

1. J. Sierra, J. Marrugo-Ramírez, R. Rodriguez-Trujillo, M. Mir, and J. Samitier, "Sensor-Integrated Microfluidic Approaches for Liquid Biopsies Applications in Early Detection of Cancer," Sensors, vol. 20, no. 5, p. 1317, Feb. 2020, doi: https://doi.org/10.3390/s20051317.
2. Y. Sun, Q. Tao, X. Wu, L. Zhang, Q. Liu, and L. Wang, "The Utility of Exosomes in Diagnosis and Therapy of Diabetes Mellitus and Associated Complications," Frontiers in Endocrinology, vol. 12, Oct. 2021, doi: https://doi.org/10.3389/fendo.2021.756581.
3. R. Kalluri and V. S. LeBleu, "The biology, function, and Biomedical Applications of Exosomes," Science, vol. 367, no. 6478, Feb. 2020, doi: https://doi.org/10.1126/science.aau6977.
4. J. Qin and Q. Xu, "Drug Research ACTA POLONIAE PHARMACEUTICA," 2014. Available: https://ptfarm.pl/pub/File/Acta_Poloniae/1998/acta4-2014.pdfpage=17
5. R. Kalluri, "The biology and function of exosomes in cancer," Journal of Clinical Investigation, vol. 126, no. 4, pp. 1208–1215, Apr. 2016, doi: https://doi.org/10.1172/jci81135.
6. M. Garcia-Contreras, R. W. Brooks, L. Boccuzzi, P. D. Robbins, and C. Ricordi, "Exosomes as biomarkers and therapeutic tools for type 1 diabetes mellitus," 2017. https://www.europeanreview.org/wp/wp-content/uploads/2940-2956-Exosomes-as-biomarkers-and-therapeutic-tools-for-type-1-diabetes-mellitus.pdf
7. M. Kopeć, K. Beton, K. Jarczewska, and H. Abramczyk, "Hyperglycemia and cancer in human lung carcinoma by means of Raman spectroscopy and imaging," Scientific Reports, vol. 12, no. 1, p. 18561, Nov. 2022, doi: https://doi.org/10.1038/s41598-022-21483-y.
8. C. de la Torre Gomez, R. V. Goreham, J. J. Bech Serra, T. Nann, and M. Kussmann, "'Exosomics'—A Review of Biophysics, Biology and Biochemistry of Exosomes With a Focus on Human Breast Milk," Frontiers in Genetics, vol. 9, Mar. 2018, doi: https://doi.org/10.3389/fgene.2018.00092.
9. J. Li et al., "Exosome detection via surface-enhanced Raman spectroscopy for cancer diagnosis," Acta Biomaterialia, vol. 144, pp. 1–14, May 2022, doi: https://doi.org/10.1016/j.actbio.2022.03.036.
10. R. S. Das and Y. K. Agrawal, "Raman spectroscopy: Recent advancements, techniques and applications," Vibrational Spectroscopy, vol. 57, no. 2, pp. 163–176, Nov. 2011, doi: https://doi.org/10.1016/j.vibspec.2011.08.003.
11. J. O'Sullivan, "Development of Bioplasmonic Platforms for Extracellular Vesicle Capture and Analysis," May 2022. Accessed: Apr. 13, 2024. [Online]. Available: https://doras.dcu.ie/27239/1/ethesis_O%27Sullivan_John_2022.pdf
12. S. Hu et al., "Raman spectroscopy combined with machine learning algorithms to detect adulterated Suichang native honey," Scientific Reports, vol. 12, p. 3456, Mar. 2022, doi: https://doi.org/10.1038/s41598-022-07222-3.
13. X. Chen et al., "Raman spectroscopy combined with a support vector machine algorithm as a diagnostic technique for primary Sjögren's syndrome," Scientific Reports, vol. 13, no. 1, Mar. 2023, doi: https://doi.org/10.1038/s41598-023-29943-9.

14. L. Zhang et al., “Raman spectroscopy and machine learning for the classification of breast cancers,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 264, p. 120300, Jan. 2022, doi: https://doi.org/10.1016/j.saa.2021.120300.
15. W. Schumacher, S. Stöckel, P. Rösch, and J. Popp, “Improving chemometric results by optimizing the dimension reduction for Raman spectral data sets,” Journal of Raman Spectroscopy, vol. 45, no. 10, pp. 930–940, Sep. 2014, doi: https://doi.org/10.1002/jrs.4568.
16. A. Amjad, R. Ullah, S. Khan, M. Bilal, and A. Khan, “Raman spectroscopy based analysis of milk using random forest classification,” Vibrational Spectroscopy, vol. 99, pp. 124–129, Nov. 2018, doi: https://doi.org/10.1016/j.vibspec.2018.09.003.
17. H. Shin et al., “Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes,” ACS Nano, vol. 14, no. 5, pp. 5435–5444, Apr. 2020, doi: https://doi.org/10.1021/acsnano.9b09119.
18. Y. Xie, X. Su, Y. Wen, C. Zheng, and M. Li, “Artificial Intelligent Label-Free SERS Profiling of Serum Exosomes for Breast Cancer Diagnosis and Postoperative Assessment,” Nano Letters, vol. 22, no. 19, pp. 7910–7918, Sep. 2022, doi: https://doi.org/10.1021/acs.nanolett.2c02928.
19. E. Alqaissi, F. Alotaibi, and M. S. Ramzan, “Graph data science and machine learning for the detection of COVID-19 infection from symptoms,” PeerJ Computer Science, vol. 9, p. e1333, Apr. 2023, doi: https://doi.org/10.7717/peerj-cs.1333.
20. Z. Wang, W. Chen, W. Zhou, R. Zhang, R. Song, and D. Yang, “A Few-shot Learning Method for Aging Diagnosis of Oil-paper Insulation by Raman Spectroscopy Based on Graph Theory,” IEEE Transactions on Dielectrics and Electrical Insulation, vol. 28, no. 6, pp. 1892–1900, Dec. 2021, doi: https://doi.org/10.1109/tdei.2021.009638.
21. D. A. Whitaker and K. Hayes, “A simple algorithm for despiking Raman spectra,” Chemometrics and Intelligent Laboratory Systems, vol. 179, pp. 82–84, Aug. 2018, doi: https://doi.org/10.1016/j.chemolab.2018.06.009.
22. P. Eilers and H. Boelens, “Baseline Correction with Asymmetric Least Squares Smoothing,” 2005.
23. N. Gallagher, “Savitzky-Golay Smoothing and Differentiation Filter,” 2020, doi: https://doi.org/10.13140/RG.2.2.20339.50725.
24. A. Martyna et al., “Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components,” Chemometrics and Intelligent Laboratory Systems, vol. 202, p. 104029, Jul. 2020, doi: https://doi.org/10.1016/j.chemolab.2020.104029.
25. D. Berrar, “Cross-Validation,” Encyclopedia of Bioinformatics and Computational Biology, pp. 542–545, 2019, doi: https://doi.org/10.1016/b978-0-12-809633-8.20349-x.
26. S. Guo, T. Bocklitz, U. Neugebauer, and J. Popp, “Common mistakes in cross-validating classification models,” Analytical Methods, vol. 9, no. 30, pp. 4410–4417, 2017, doi: https://doi.org/10.1039/c7ay01363a.
27. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and

SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.
28. "Neo4j Graph Platform – The Leader in Graph Databases," Neo4j Graph Database Platform. https://neo4j.com
29. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks, vol. 56, no. 18, pp. 3825–3833, Dec. 2012, doi: https://doi.org/10.1016/j.comnet.2012.10.007.
30. D. F. Gleich, "PageRank Beyond the Web," SIAM Review, vol. 57, no. 3, pp. 321–363, Jan. 2015, doi: https://doi.org/10.1137/140976649.
31. J. Zhang and Y. Luo, "Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network," download.atlantis-press.com, Mar. 01, 2017. https://download.atlantis-press.com/proceedings/msam-17/25874733
32. B. Ruhnau, "Eigenvector-centrality — a node-centrality?," Social Networks, vol. 22, no. 4, pp. 357–365, Oct. 2000, doi: https://doi.org/10.1016/s0378-8733(00)00031-9.
33. J. Zhou, L. Liu, W. Wei, and J. Fan, "Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding," ACM Computing Surveys, vol. 55, no. 2, pp. 1–35, Jan. 2022, doi: https://doi.org/10.1145/3491206.
34. H. Chen, Syed Fahad Sultan, Y. Tian, M. Chen, and S. Skiena, "Fast and Accurate Network Embeddings via Very Sparse Random Projection," arXiv (Cornell University), Nov. 2019, doi: https://doi.org/10.1145/3357384.3357879.
35. vR. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for Raman and infrared spectroscopy," EPJ Techniques and Instrumentation, vol. 2, no. 1, Jun. 2015, doi: https://doi.org/10.1140/epjti/s40485-015-0018-6.
36. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.