Testing the Efficacy of Peak Profiling and Graph Databases to Improve the Machine Classification of Disease in the Raman Spectrum of Exosomes

Student ID: 20364133, 20449042
Student Names: Edward Bolger, Stan Goodwin
Student email: edward.bolger25@mail.dcu.ie, stan.goodwin6@mail.dcu.ie
Supervisor: Mark Roantree (Mark.Roantree@dcu.ie)

Table of Contents

Executive Summary

This proposal highlights the potential of exosomes, microscopic extracellular vesicles containing biomarkers, as a non-invasive source of information for disease screening. Exosomes can be analysed using a technique called Raman spectroscopy which identifies the molecular ‘fingerprint’ of a tested sample. These spectral fingerprints are complex and noisy which opens avenues for machine learning to automate their analysis.

The project aims to test the effectiveness of novel techniques, like peak profiling, feature engineering and graph data structures against traditional machine learning methods in the task of classifying exosome spectral samples into hypoglycemic, hyperglycemic, or normal categories.

The phases of the project consist of peak detection of the Raman spectroscopy data, feature engineering to create a uniform feature set for our baseline machine learning tests, converting the irregular feature set to a graph database and then proceeding to use techniques such as community detection, centrality and graph neural networks to create a new type of classifier that can be compared with the baseline test.

The goal of this project is to create a better understanding of which techniques are more effective for classifying exosome spectral data, potentially improving disease screening methods.

1. Motivation and Background

Diagnosing diseases such as cancer or diabetes can be a difficult task. Cancer diagnosis is usually achieved using tissue biopsies, where tissue is extracted from a suspected tumour and analysed in a lab (Sierra et al., 2020). This process is time consuming and expensive and can potentially pose risks to the patient. Likewise, early diabetes screening methods are lacking due to the requirement of blood tests, reducing the patient's chance of having a full treatment plan (Sun et al., 2021). As a result there is a large push to find faster and less intrusive methods of disease screening.

One such approach is through the analysis of exosomes. Exosomes are microscopic extracellular vesicles involved in cell-to-cell communication. They contain DNA, RNA and proteins. Exosomes are present in the majority of liquid that the body produces including blood, urine and saliva (Kalluri and LeBleu, 2020). According to Sun et al. (2021), exosomes carry biomarkers that can indicate the presence of disease in their parent cells. As they are found in common bodily fluids, exosomes are a useful indicator of disease due to the fact that they can be easily collected in a non-invasive manner. This provides a great advantage over traditional tissue biopsies, saving time while causing less stress for patients.

However, exosomes are difficult to analyse as they exist within complex biological samples and their small size of about 100 nanometres requires very precise measuring methods such as Raman spectroscopy (Crosby et al., 2022). Raman spectroscopy is an analytical technique where scattered light is used to measure the vibrational energy of the chemical bonds that make up a material. This results in a spectrum known as a Raman “fingerprint” which holds chemical and structural information that is intrinsic to the analysed sample, enabling its identification and classification (Qi et al., 2023). The spectral data yielded by Raman spectroscopy is often complex and contains background noise, requiring intense data processing to yield valuable results. This creates an opportunity to use machine learning to automate the analysis and classification of these spectral results.

Machine learning can be applied to analyse large amounts of Raman spectroscopy data, finding relationships, patterns, and connections in the spectral datasets. Furthermore, it can be applied to to classify materials based on their Raman fingerprints. In this case, the biomarkers which indicate disease can be identified in the spectrum of an exosome that has undergone Raman spectroscopy. Training a machine learning model to identify these biomarkers has the potential to let exosome analysis be used as a non-intrusive screening tool for many diseases.

2. Problem Statement

This project aims to measure the effectiveness of novel techniques, such as peak detection, feature engineering and graph data structures against traditional machine learning methods at classifying the status of exosome spectral samples. Our objective is to create a supervised machine learning model using these techniques that can classify whether an exosome sample is hypoglycemic, hyperglycemic or normal. The model will be trained using features extracted from 3046 labelled exosome spectra.

A PhD student at DCU has already attempted this classification task using a random forest algorithm, achieving an overall accuracy of 60% when trained on exosome spectral data. This approach was particularly successful at identifying the hypoglycemic samples, however it struggled at distinguishing between normal and hyperglycemic cells hurting the total accuracy.

In a Raman spectrum, the size and intensity of peaks serve as indicators of the chemical composition of the material under examination and have been effectively utilised for representing exosomes (Li et al., 2022). By analysing the properties of the peaks in each sample, we aim to engineer a feature-set based on the peaks of each spectrum, which can then be used to train the model. Due to the varying shape of each spectrum the feature sets will not be uniform. This is where graph data structures will be utilised as unlike relational databases, they do not require uniform schemas, instead they are based on the structural relationships which can have irregular shapes, which may enable us to more effectively represent the complex variations in exosome spectra. (Ghrab et al., 2016, p. 2).

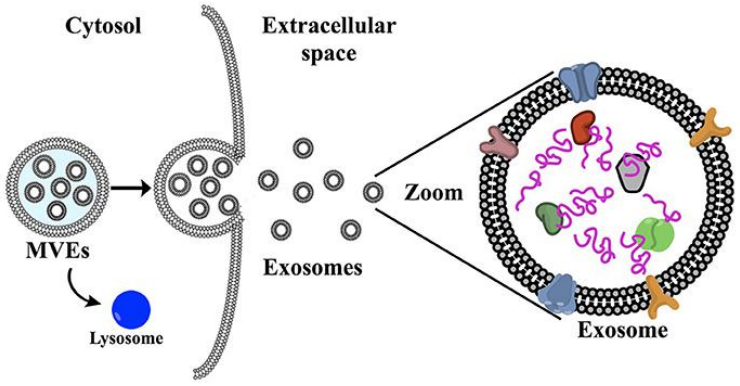
The ideal outcome for this project would be a fully functional machine classification model based on these techniques that can be evaluated against the labelled data and compared with the traditional machine learning approaches to find the advantages and disadvantages of using these techniques for exosome classification, leading to better disease screening.

3. State of the Art

There has been a variety of studies related to both the use of Raman spectroscopy on exosome samples to detect disease, as well as investigations into the use of machine learning to extract and classify Raman spectroscopy data. Furthermore, other research has shown the effectiveness of peak detection feature engineering and graph machine learning when used in certain spectral classification tasks.

Exosomes

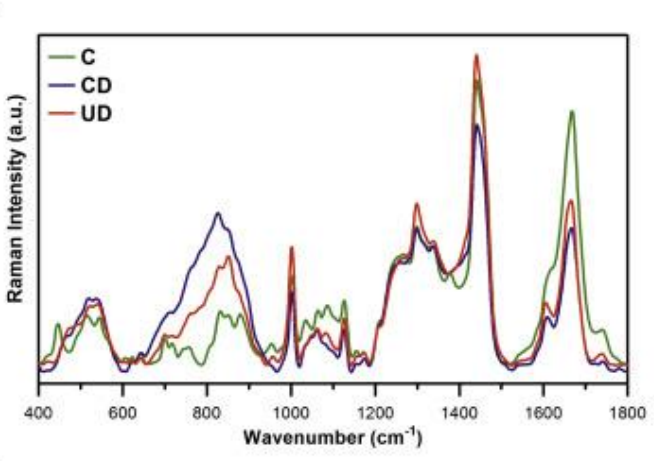
As mentioned above, exosomes are extracellular vesicles, nanoparticles excreted by cells, containing bioactive molecules which can function as biomarkers to be used to aid early diagnosis of diseases such as diabetes and cancer (Sun et al., 2021; Li et al., 2020). The benefit of using exosomes to aid this diagnosis is that exosomes are present in most liquids that the body produces, thus they can be retrieved via a non-invasive liquid biopsy. The data we will be working with provides samples of exosome data from Raman spectroscopy.



A diagram exosomes being emitted from a cell and the molecules contained within (De La Torre Gomez et al., 2018)

The samples we will be working with indicate whether the parent cell of the exosome is hyperglycemic, hypoglycemic, or normal. Hypoglycemia is a condition in which a person's blood sugar level is lower than the standard range, reducing the body's ability to feed its cells. (Mayo Clinic, 2020). Hypoglycemia can cause recurrent illness, or even death in most people with type 1 diabetes and many people with type 2 diabetes (Cryer, Davis and Shamoon, 2003). Conversely, hyperglycemia is when the level of blood sugar in your body is too high. Again, this affects people with diabetes and can become dangerous if gone untreated (NHS, 2017).

A study by Roman et al. (2019) found that “Raman spectroscopy is a powerful tool in UEVs studies for fast differentiation between healthy and diabetic patients”. In this study they used Raman spectroscopy on urinary extracellular vesicles to differentiate between diabetic and non-diabetic patients. They also grouped diabetic patients by their level of glycemic control, dividing them based on how well they maintain their blood sugar levels. The spectra were able to show the structural differences of the vesicles produced by both the diabetic and non-diabetic patients, while also differentiating samples based on the level of glycemic control. The largest indicators of hyperglycemia were found along the 600-950 cm^{-1} region, as well as a variance in the peak at 1670 cm^{-1}. This indicates that peak intensity might be a key input feature when predicting whether a sample is hyperglycemic or normal.



The Raman spectra of urinary extracellular vesicles grouped by good and bad glycemic control (CD and UD) and normal (C) (Roman et al., 2019).

Traditional Machine Learning with Raman Spectroscopy

Machine classification is a supervised machine learning method where a model tries to predict the correct label of a given input data (Keita, 2022). It involves training the model on labelled training data, then evaluating the results on a test dataset.

One study by (Qi et al., 2023) outlines the role of machine learning at extracting valuable information from Raman spectroscopy data across a wide range of fields, including medical diagnosis. The most common use cases of machine learning are in the classification and identification of Raman spectra. The paper explores the roles of traditional statistical based methods of machine learning including, K-Nearest Neighbour, Decision Trees and Random Forests, as well as deep-learning approaches used by artificial neural networks. K-Nearest Neighbour has no training process, instead it maps samples to n-dimensional space, then uses a distance metric to select the k nearest points to the unknown samples. The unknown samples are then classified by the categories of the nearest neighbour samples. Decision trees use a series of nested decision rules to learn the features of the dataset allowing it to be partitioned. Decision trees are prone to overfitting, where they accurately predict the training set, but are less accurate against real data. Random forests are more resistant to overfitting as

they combine the output of multiple decision trees to achieve a single pooled result. On the other hand, deep learning does not require knowledge of the relationship between inputs and outputs, instead neural networks learn inherent features of the data autonomously by optimising the weights of neural layers to learn intrinsic information from sample data, making them useful for processing noisy, nonlinear signals.

Promising results were found in a study by Shin et al. (2020) where they trained a Residual Neural Network on the Raman spectra of exosomes derived from normal and lung cancer cell lines. The model was able classify the cancerous and non cancerous samples with an accuracy of 95%. However, as deep learning methods discover inherent features on their own when testing other statistical algorithms, such as random forests, we will have to manually build input feature sets from the spectral data.

Feature Engineering from Spectral Peaks

Peak detection is the process of finding local maxima in a series, in this case a spectrum. In Raman spectroscopy, peaks represent the specific vibrational mode of the molecule, so the combined series of peaks represent the molecule's fingerprint (Unruh and Meyers, 2016). The use of peak detection with Raman spectroscopy was explored by Li, Shen and Zhou (2022). They applied a peak detection method to Raman spectroscopy data of Glioma. "Glioma is the most common tumour of the central nervous system" (Li, Shen and Zhou, 2022). In that paper, they extract variables such as peak position, intensity, and half-wave width to use as features for a weight fuzzy-rough nearest neighbour algorithm. Given that they were able to achieve an accuracy of 86.99%, this could prove as a useful starting point for what features we should extract the peaks in our exosome samples.

Traditional machine learning methods tend to rely on tabular representations of data as input features, usually columns in a relational database (Assareh, 2022). The use of tabular data requires a uniform schema where each sample is represented by a fixed number of attributes. As each exosome possesses a unique Raman spectral "fingerprint", it is likely that the different exosome samples exhibit varying peak numbers and intensities generating an irregular shaped feature set which is unsuitable for our baseline tests using traditional machine learning. This means that we will have to use feature engineering in order to create a uniform feature set from this spectral data that can be used as input in our baseline test. "Feature engineering is the task of improving predictive modelling performance on a dataset by transforming its feature space" (Nargesian et al., 2017).

Feature selection is a key part of feature engineering. Methods of selecting features are usually classified into three groups: filters, embedded methods and wrappers (Remeseiro and Bolon-Canedo, 2019). Filters are independent of any learning method as they are centred around selecting features based on general characteristics of the data, for example coding a rule that discards a feature if it crosses a predefined threshold. Wrappers are an induction method which evaluates candidate subsets of features, mathematically defining feature

importance. Embedded methods combine the techniques of filters and wrappers (Remeseiro and Bolon-Canedo, 2019).

Raman spectroscopy falls under biomedical signal processing. Analysing feature selection in other areas of biomedical signals could lead to successful feature engineering in relation to the Raman spectra of our exosome data. One other area of biomedical signals that also uses spectral data is brain-computer interfaces (BCI). Here they use near-infrared spectroscopy. These BCI techniques allow “communication between users and systems without intervention from muscles or external devices” (Remeseiro and Bolon-Canedo, 2019). According to them, a new hybrid method based on particle swarm optimisation and novel neighbourhood rough set classifier has been successfully applied to multiclass classification of brain signals, resulting in a feature set that produces a high classification accuracy. Particle swarm optimisation is “a concept for the optimization of nonlinear functions using particle swarm methodology” (Kennedy and Eberhart, 1995) and novel neighbourhood rough set classifier is based on a distance metric function and a value of a neighbourhood relation all of which is described in Kumar and Inbarani (2015). These machine learning developments in the biomedical field may have use when building the exosome classifier.

In a study by Rubattu, Maroni and Corani (2023), they used clustered permutation feature importance. This allows them to group together highly correlated features to form a correlation matrix. They then shuffle the clustered variables and attempt to find the most orthogonal information in the different clusters as this leads to the most reliable estimate of importance. They then drop the non-informative features. This could prove useful to us if we are able to compare clusters of different lengths in order to identify a core subset of features which will be uniform in length across all samples.

Applications of Graph-Based Machine Learning

As was previously mentioned, traditional machine classification models tend to require regular feature sets utilising uniform schemas. This can be a disadvantage when the data being modelled is represented in a non-uniform way, which is the case with Raman spectra. This is where graph structures have an advantage. Graph databases use a non-strict schema design based around modelling structural relationships (Ghrab et al., 2016). As a result, using a graph structure to represent the Raman spectra as a series of nodes and relationships does not require the use of a uniform feature set, and may create a more detailed representation of the exosome structure that can be used to train the classifier as well as providing more information on the relationships between the data.

Graph-based classification of Raman spectroscopy has been used before. A study by Wang et al. (2021) attempted to build a graph convolutional neural network (GCN) for diagnosing oil paper insulation. Raman spectra were represented using an identity matrix. These matrices acted as nodes in the graph, then the gaussian kernel function was used to calculate the level of similarity between each sample using the euclidean distance, creating a graph structure for the entire Raman spectroscopy library. The GCN is a semi-supervised deep learning method

or processing graph domain information. In this study it was made up of an input layer of 200 neurons, a hidden layer of 10 neurons and an output layer with 4 neurons The nodes on the graph were divided into known and unknown groups equivalent to a train-test split in traditional machine learning. The unknown nodes were then classified by the model based on topological relationship between the nodes on the graph. The error between the output value of the training node and the real value was calculated and the weights of the neurons were adjusted accordingly. This process repeated until the output stabilised, yielding accuracies as high as 95.5%.

Graphs can also be used to improve the performance of traditional machine learning models by providing new metrics describing the relationships hidden within tabular data that can be used as input features by the model (Hodler, 2019). One metric that is commonly used to improve model performance is node centrality. Centrality is calculated by algorithms such as PageRank which describe the importance of nodes within the network based on the influence of neighbouring nodes (Amine, 2020). Community detection is another graph metric that can aid in machine classification tasks. Methods such as Louvain community detection are used to detect localised groups of nodes in a network (Jayawickrama, 2021). The groups identified by community detection may be useful identifiers that the traditional classification model can use to predict the status of a sample.

4. Methodology

Investigate the Samples

This project is building on the research of a PhD student at DCU who investigated the possibility of classifying exosomes as normal, hyperglycemic or hypoglycemic based on Raman spectroscopy data of those exosomes. The main dataset we will be working with consists of anonymised exosome spectra data. The fields in this dataset are indicate the spectrum sample (SpecID), the sequence number of the values representing the spectrum (Seq), the wavenumber in cm^{-1} (Wavenumber), the absorbance of the sample at that wavenumber (Absorbance), the surface ID which is a foreign key related to in a table that is less relevant to us (SurID) and the status of the exosome (Status).

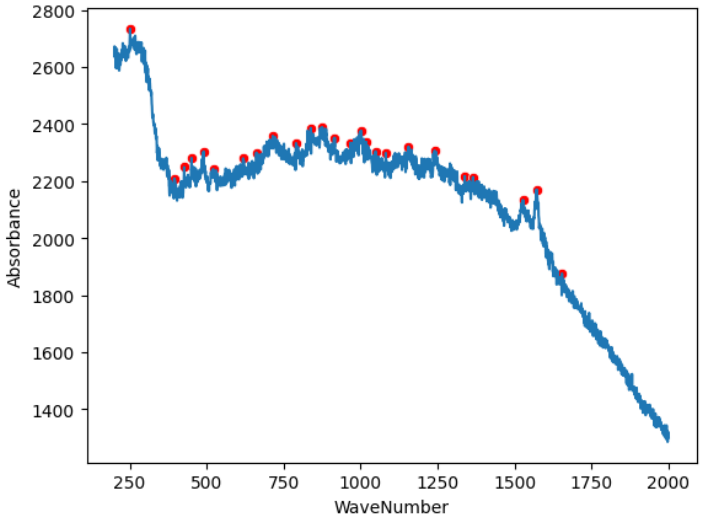
	SpecID	Seq	WaveNumber	Absorbance	SurID	Status
0	201210-1-00	0	200.00000	2709.3699	201210-1	Normal
1	201210-1-00	1	200.68336	2697.1318	201210-1	Normal
2	201210-1-00	2	201.36674	2696.0413	201210-1	Normal
3	201210-1-00	3	202.05011	2678.5925	201210-1	Normal
4	201210-1-00	4	202.73349	2670.8928	201210-1	Normal

The 5 rows of of the exosome spectroscopy dataset

This data is an example of the spectral fingerprint of each exosome and a label that represents its status. We have been given this data in the form of a CSV file. We will import the data into a jupyter notebook using the pandas data manipulation library in the python programming language (Pandas, 2018; Python, 2019). We will then proceed to use pandas to clean, process and gain familiarity with the data. By visualising the spectra of the different sample groups, we might be able to identify the key spectral properties that differentiate the three statuses.

Using Peak Detection to Generate a Non-Uniform Featureset

In order to make any accurate classification with either traditional machine learning or graph techniques we first need to extract features to feed to these methods. To do this we intend to extract the peaks from the spectral fingerprint of each exosome. To do this we will use the SciPy library for python. We can use the 'find_peaks' function from SciPy to identify the peaks in the Raman spectra (docs.scipy.org, n.d.). The find_peaks function contains parameters such as peak prominence, peak width and distance between peaks. Defining these properties will allow us to fine-tune the function to return a meaningful selection of peaks that properly represent the underlying sample. Our target is to fine tune our peak detection function to return between 20 and 35 peaks. Below is a plot of the spectral data from a sample exosome and its peaks plotted with the matplotlib (Matplotlib, 2012).



Example of peak detection on a Raman spectroscopy sample

Creating a Uniform Featureset using Feature Engineering

Each peak extracted from the spectral data will be represented as a set of characteristic variables. In order to determine what characteristics to use in our uniform feature set, we will manually determine what characteristics we use like in Li, Shen and Zhou (2022) where they

used peak position, intensity and half-wave width as variables. We may also use some kind of automatic feature importance technique such as clustered permutation feature importance used in Rubattu, Maroni and Corani (2023). Normalisation might need to be performed on the characteristics to ensure that our model isn't biassed towards features that tend to have larger numbers.

To create a baseline test using techniques such as random forest, we need to transform our irregular feature set into a uniform feature set. This then calls into question how we will produce a uniform feature set and what features we will keep or drop. With the feature set we receive from the peaks, one method of approaching this problem is using the top-n peaks based on either importance or some user defined rules such as the top -n widest peaks or the top n-most prominent peaks. Another technique we may employ is gathering several different peaks to include in the feature set based on that peak's property within a sample. For example we may choose to include the highest peak, the lowest peak, the widest peak and the most prominent peak from each sample and just use the characteristic variables of those peaks. Either way, when this is complete we will have a uniform feature set that we can use to train our baseline machine learning algorithm.

Run a Baseline Experiment Using Traditional Machine Learning Algorithms

The first baseline test we intend to run on our feature set is Random Forest. As in Qi et al. (2023) it proved to be effective over decision trees as it was less resistant to overfitting and subsequently we believe it will provide a suitable baseline. We will use the scikit-learn library for python to perform our baseline machine learning test (scikit-learn, 2019). We will first split our dataset into a training and test set with an 80-20 split. After we train our model we will then proceed to evaluate it based on several metrics such as accuracy, precision, recall and f-score. We may also decide to attempt a baseline test using logistic regression, although if we had any categorical features, this would require further feature engineering.

Convert the spectrum features to a graph data structure

After the baseline test is complete, we will work on representing the spectral data in a graph database, testing a variety of graph schemas that can more accurately represent the non-uniform structure of the spectra. One approach will be to represent the spectrum peaks as nodes, while the relationships will be defined by the similarity of peak features and their concurrent frequency within the sample. For example, peaks that appear within a wavenumber window will be connected together. Other measures could calculate the euclidean distance between peaks in the spectrum to measure the strength of the relationships. We will also explore the approach used by Wang et al. where each spectrum is represented by a single node with the relationships between nodes calculated using an adjacency matrix based on the level of similarity between each spectrum creating a Raman spectroscopy library which can then be used to analyse the relationships between the different exosome samples. These graphs will be created using Neo4j where they will be analysed (Neo4j, 2017).

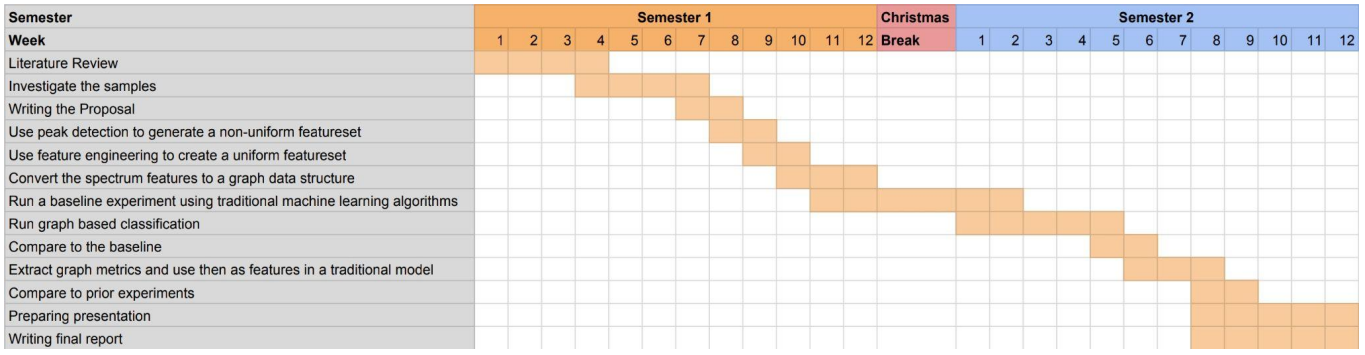
Graph Based Classification

Initial testing for graph classification will consist of utilising several different community detection algorithms such as Louvain modularity in order to see if the graph database structure alone is able to reveal the difference between the status of the exosome samples. Running centrality algorithms may yield information about the importance of different spectral features. A combination of these metrics will need to be utilised to create an accurate classifier so we will look into running a graph neural network that is trained to predict the exosome status based on the total graph structure. Additionally, we will extract the graph metrics and use them as features in a traditional machine learning model to see how they influence performance.

Evaluation of results against the baseline example

Once we have finalised the graph classifier and have run a traditional machine learning model using the extracted graph features we will compare the results with the baseline model when evaluated against the test set. Examining the precision, recall, accuracy and f-score of each model will indicate how many samples the models were able to correctly predict (Mishra, 2018). Particular focus will be to the predictive accuracy of the 3 statuses, hypoglycemic, hyperglycemic and normal. Even if a model has lower overall accuracy, the ability to more accurately predict one of the statuses may provide merit to its use.

5. Project Plan



Project Gantt Chart

In the initial four weeks, we conducted a literature review to gain insights into the real-world applications of exosomes, state-of-the-art machine learning techniques in biomedical signal processing, and graph databases. We then explored our sample data between weeks 4 to 7 to understand its relevance to the project. Following this, we began drafting our project proposal.

Our next steps involve implementing a peak detection algorithm to create a non-uniform feature set, expected to take two weeks. Subsequently, we'll perform feature engineering to transform this non-uniform feature set into a uniform one for training our baseline test. Converting our feature set into a graph database, a significant project milestone, will require approximately three weeks of work. Simultaneously, we'll work on building our machine learning classifier as the baseline test, which includes the Christmas break, the last two weeks of semester one as well as the first two weeks of semester two. We will consider this another milestone in the project.

The next phase of the project involves running our graph-based classification for the first five weeks of semester two, marking the third major milestone. Following this, we'll conduct a comparative analysis between our baseline and graph classifier results. Subsequently, we'll extract features from the graph and re-run our baseline model to potentially enhance results. Finally, we'll compile all results and document our entire project process in the final report and project presentation.

6. Conclusion

To summarise, this project explores the use of Raman spectroscopy and machine learning to classify exosome spectral data for non-invasive disease screening. The objective is to compare peak detection and graph databases with traditional machine learning methods. By investigating the spectral data, engineering features, running machine learning tests, and employing graph-based techniques, the project aims to see if these methods have any advantages in the field of exosome classification.

Bibliography

Amine, A. (2020). *PageRank algorithm, fully explained*. [online] Medium. Available at: https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af.

Assareh, A. (2022). *Demystifying Graph based Machine Learning*. [online] MLearning.ai. Available at: https://medium.com/mlearning-ai/demystifying-graph-based-machine-learning-ed6b6b7c4081 [Accessed 3 Nov. 2023].

Aurai (2022). *Traversing Information Networks by Graph Databases*. [online] Aurai. Available at: https://aurai.com/traversing-information-networks-by-graph-databases/ [Accessed 3 Nov. 2023].

Crosby, D., Bhatia, S., Brindle, K.M., Coussens, L.M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R.C., Gambhir, S.S., Kuhn, P., Rebbeck, T.R. and Balasubramanian, S. (2022). Early detection of cancer. *Science*, 375(6586). doi:https://doi.org/10.1126/science.aay9040.

Cryer, P.E., Davis, S.N. and Shamoon, H. (2003). Hypoglycemia in Diabetes. *Diabetes Care*, [online] 26(6), pp.1902–1912. doi:https://doi.org/10.2337/diacare.26.6.1902.

De La Torre Gomez, C., V. Goreham, R., J. Bech Serra, J., Nann, T. and Kussman, M. (2018). *Schematic Representation of Exosome Biogenesis*. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00092/full [Accessed 2 Nov. 2023].

docs.scipy.org. (n.d.). *scipy.signal.find_peaks — SciPy v1.6.0 Reference Guide*. [online] Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html.

domino.ai. (n.d.). *What is Feature Engineering? | Domino Data Science Dictionary*. [online] Available at: https://domino.ai/data-science-dictionary/feature-engineering.

Fourrier, C. (2023). *Introduction to Graph Machine Learning*. [online] huggingface.co. Available at: https://huggingface.co/blog/intro-graphml.

Ghrab, A., Romero, O., Skhiri, S., Vaisman, A. and Zimányi, E. (2016). *GRAD: On Graph Database Modeling*. [online] Available at: https://arxiv.org/pdf/1602.00503.pdf [Accessed 1 Nov. 2023].

Hodler, M.N., Amy (2019). *How graph algorithms improve machine learning*. [online] O'Reilly Media. Available at: https://www.oreilly.com/content/how-graph-algorithms-improve-machine-learning/ [Accessed 3 Nov. 2023].

Jayawickrama, T.D. (2021). *Community Detection Algorithms*. [online] Medium. Available at: https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae.

Kalluri, R. and LeBleu, V.S. (2020). The biology, function, and Biomedical Applications of Exosomes. *Science*, 367(6478). doi:https://doi.org/10.1126/science.aau6977.

Keita, Z. (2022). *Classification in Machine Learning: A Guide for Beginners*. [online] www.datacamp.com. Available at: https://www.datacamp.com/blog/classification-machine-learning.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, [online] 4, pp.1942–1948. doi:https://doi.org/10.1109/icnn.1995.488968.

kumar, S.U. and Inbarani, H.H. (2015). A Novel Neighborhood Rough Set Based Classification Approach for Medical Diagnosis. *Procedia Computer Science*, 47, pp.351–359. doi:https://doi.org/10.1016/j.procs.2015.03.216.

Li, J., Li, Y., Li, P., Zhang, Y., Du, L., Wang, Y., Zhang, C. and Wang, C. (2022). Exosome detection via surface-enhanced Raman spectroscopy for cancer diagnosis. *Acta Biomaterialia*, 144, pp.1–14. doi:https://doi.org/10.1016/j.actbio.2022.03.036.

Li, Q., Shen, J. and Zhou, Y. (2022). Diagnosis of Glioma Using Raman Spectroscopy and the Entropy Weight Fuzzy-Rough Nearest Neighbor (EFRNN) Algorithm on Fresh Tissue. *Analytical Letters*, 56(6), pp.895–905. doi:https://doi.org/10.1080/00032719.2022.2107660.

Li, S., Yi, M., Dong, B., Tan, X., Luo, S. and Wu, K. (2020). The role of exosomes in liquid biopsy for cancer diagnosis and prognosis prediction. *International Journal of Cancer*. doi:https://doi.org/10.1002/ijc.33386.

Li, Y., Yang, Y., Zhang, J., Yuan, Q. and Liang, Y. (2023). *Optica Publishing Group*. [online] opg.optica.org. Available at: https://opg.optica.org/optcon/fulltext.cfm?uri=optcon-2-8-1875&id=536450 [Accessed 1 Nov. 2023].

Matplotlib (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation*. [online] Matplotlib.org. Available at: https://matplotlib.org/.

Mayo Clinic (2020). *Hypoglycemia - Symptoms and causes*. [online] Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/hypoglycemia/symptoms-causes/syc-20373685.

Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. [online] Medium. Available at: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E.B. and Turaga, D. (2017). Learning Feature Engineering for Classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. [online] doi:https://doi.org/10.24963/ijcai.2017/352.

Neo4j (2017). *Neo4j Graph Platform – The Leader in Graph Databases*. [online] Neo4j Graph Database Platform. Available at: https://neo4j.com/.

NHS (2017). *High blood sugar (hyperglycaemia)*. [online] nhs.uk. Available at: https://www.nhs.uk/conditions/high-blood-sugar-hyperglycaemia/#:~:text=High%20blood%20sugar%20(hyperglycaemia)%20is.

Pandas (2018). *Python Data Analysis Library — pandas: Python Data Analysis Library*. [online] Pydata.org. Available at: https://pandas.pydata.org/.

Python (2019). *Python*. [online] Python.org. Available at: https://www.python.org/.

Qi, Y., Hu, D., Jiang, Y., Wu, Z., Zheng, M., Chen, E.X., Liang, Y., Sadi, M.A., Zhang, K. and Chen, Y.P. (2023). Recent Progresses in Machine Learning Assisted Raman Spectroscopy. *Advanced Optical Materials*, 11(14). doi:https://doi.org/10.1002/adom.202203104.

Remeseiro, B. and Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112, p.103375. doi:https://doi.org/10.1016/j.compbiomed.2019.103375.

Roman, M., Kamińska, A., Drożdż, A., Platt, M., Kuźniewski, M., Małecki, M.T., Kwiatek, W.M., Paluszkiewicz, C. and Stępień, E.Ł. (2019). Raman spectral signatures of urinary extracellular vesicles from diabetic patients and hyperglycemic endothelial cells as potential biomarkers in diabetes. *Nanomedicine: Nanotechnology, Biology and Medicine*, [online] 17, pp.137–149. doi:https://doi.org/10.1016/j.nano.2019.01.011.

Rubattu, N., Maroni, G. and Corani, G. (2023). *Electricity Load and Peak Forecasting: Feature Engineering, Probabilistic LightGBM and Temporal Hierarchies*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2305.05575.

scikit-learn (2019). *scikit-learn: machine learning in Python*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/.

Shin, H., Oh, S., Hong, S., Kang, M., Kang, D., Ji, Y., Choi, B.H., Kang, K.-W., Jeong, H., Park, Y., Hong, S., Kim, H.K. and Choi, Y. (2020). Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes. *ACS Nano*, 14(5), pp.5435–5444. doi:https://doi.org/10.1021/acsnano.9b09119.

Sierra, J., Marrugo-Ramírez, J., Rodriguez-Trujillo, R., Mir, M. and Samitier, J. (2020). Sensor-Integrated Microfluidic Approaches for Liquid Biopsies Applications in Early Detection of Cancer. *Sensors*, 20(5), p.1317. doi:https://doi.org/10.3390/s20051317.

Sun, Y., Tao, Q., Wu, X., Zhang, L., Liu, Q. and Wang, L. (2021). The Utility of Exosomes in Diagnosis and Therapy of Diabetes Mellitus and Associated Complications. *Frontiers in Endocrinology*, 12. doi:https://doi.org/10.3389/fendo.2021.756581.

Unruh, A. and Meyers, K. (2016). *4.3: Raman Spectroscopy*. [online] Chemistry LibreTexts. Available at: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Physical_Methods_in_Chemistry_and_Nano_Science_%28Barron%29/04%3A_Chemical_Speciation/4.03%3A_Raman_Spectroscopy.

Wang, Z., Chen, W., Zhou, W., Zhang, R., Song, R. and Yang, D. (2021). 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*, 28(6), pp.1892–1900. doi:https://doi.org/10.1109/tdei.2021.009638.

Xu, O., Liu, J., Tong, X., Zhang, C., Deng, C., Mao, Y., Yin, R., Jin, C. and Fang, D. (2020). A multi-peak detection algorithm for Fiber Bragg Grating sensing systems. *Optical Fiber Technology*, 58, pp.102311–102311. doi:https://doi.org/10.1016/j.yofte.2020.102311.