A Graph Based Raman Spectral Processing Technique for Exosome Classification

Vuong M. Ngo[1,4] ✉, Edward Bolger[2], Stan Goodwin[2],
John O'Sullivan[3], Dinh Viet Cuong[1], and Mark Roantree[1]

[1] Insight Centre for Data Analytics, Dublin City University, Ireland
[2] School of Computing, Dublin City University, Ireland
[3] School of Chemical Sciences, Dublin City University, Ireland
[4] Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
vuong.ngo@dcu.ie or vuong.nm@ou.edu.vn, eddie.bolger62@gmail.com,
stangoodwin02@gmail.com, johns@gmail.com, mark.roantree@dcu.ie

Abstract. Exosomes are small vesicles crucial for cell signaling and disease biomarkers. Due to their complexity, an "omics" approach is preferable to individual biomarkers. While Raman spectroscopy is effective for exosome analysis, it requires high sample concentrations and has limited sensitivity to lipids and proteins. Surface-enhanced Raman spectroscopy helps overcome these challenges. In this study, we leverage Neo4j graph databases to organize 3,045 Raman spectra of exosomes, enhancing data generalization. To further refine spectral analysis, we introduce a novel spectral filtering process that integrates the PageRank Filter with optimal Dimensionality Reduction. This method improves feature selection, resulting in superior classification performance. Specifically, the Extra Trees model, using our spectral processing approach, achieves 0.76 and 0.857 accuracy in classifying hyperglycemic, hypoglycemic, and normal exosome samples based on Raman spectra and surface, respectively, with group 10-fold cross-validation. Our results show that graph-based spectral filtering combined with optimal dimensionality reduction significantly improves classification accuracy by reducing noise while preserving key biomarker signals. This novel framework enhances Raman-based exosome analysis, expanding its potential for biomedical applications, disease diagnostics, and biomarker discovery.

Keywords: Exosome· Raman Spectroscopy· Graph Database · Machine Learning · PageRank · Linear Discriminant Analysis

1 Introduction

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 [25]. 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 [26]. As a result, there is a significant drive to discover faster, less invasive methods of disease screening.

Exosomes are microscopic extracellular vesicles that play a crucial role in cell-to-cell communication, as illustrated in Figure 1. They carry proteins, DNA, RNA, and other biological materials derived from the cells that produce them [15]. Exosomes are present in nearly all bodily fluids, including blood, saliva, and urine [21]. Their composition makes them valuable as biomarkers, revealing the presence of diseases in their parent cells [26]. Because exosomes can be obtained from common bodily fluids, non-invasive liquid biopsies offer significant advantages over traditional tissue biopsies, such as reduced patient stress and quicker diagnostic processes. These liquid biopsies have been successfully applied to diagnose various diseases, including prostate, pancreatic, breast, and ovarian cancers, as well as diabetes and hyperglycemia [14], [16].

Due to their small size (40–150 nm) and presence in complex biological samples, exosomes require specialized techniques for measurement and analysis [17]. Raman Spectroscopy, a non-destructive vibrational spectroscopic technique based on inelastic photon scattering, is particularly effective for analyzing exosome structures [7]. Using a monochromatic laser, it excites chemical bonds within the particle, causing them to emit radiation with a slight frequency shift, known as the Raman shift, which forms the x-axis of a

Fig. 1: The Emission of Exosomes from a Cell [27]

Raman spectrum. The spectrum's peaks represent specific molecular bond vibrations, creating a unique "fingerprint" that identifies a substance's shape and composition. Surface-Enhanced Raman Spectroscopy (SERS) further amplifies the Raman signal by placing the sample on roughened metal surfaces, enhancing its detection sensitivity.
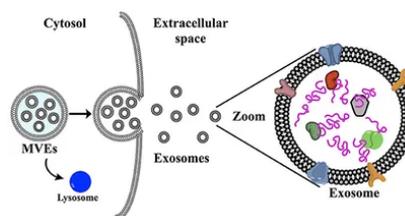
Contribution. The contribution in this paper can be articulated as follows:

- To enhance the identification of representative spectra, we developed a novel spectral filtering process that integrates the PageRank Filter (RDF) with optimal Dimensionality Reduction (DR) centrality. This method removes noisy or unrepresentative spectra while preserving the most informative spectral features, leading to a more robust dataset for classification.
- We applied a range of machine learning models to classify exosome samples based on processed Raman spectra or surfaces. The models were chosen for their strong performance in high-dimensional data and their ability to capture complex relationships within spectral features.
- To ensure the generalizability of our approach, we employed group k-fold cross-validation, which prevents data leakage by ensuring that spectra from the same experimental surface do not appear in both training and test sets.

2   Related Work

Most studies applying machine learning to Raman spectra follow a standard pre-processing sequence: despiking to remove cosmic ray artifacts, baseline correction to isolate the Raman signal, smoothing to reduce noise, and normalization to adjust for intensity variations [13], [5], [32]. To mitigate dimensionality where large feature spaces can degrade classifier performance, feature reduction techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), or partial least squares (PLS) are applied [23]. Machine learning is typically performed using algorithms like Support Vector Machines (SVM) [32], [2].

Shin et al. [24] employed a Residual Neural Network to classify Raman spectra of exosomes derived from normal and lung cancer cell lines. Their study utilized standard preprocessing steps, including baseline correction, denoising, and normalization. Neural networks, unlike traditional models, can automatically extract features from the data, eliminating the need for manual feature engineering. Similarly, Xie et al. [30] used an Artificial Neural Network (ANN) with SERS spectra of exosomes to successfully detect four types of breast cancer.

Utilising Graph Representations with Machine Learning, Graph databases have been successfully used to develop additional machine learning features in relation to disease classification. In a study by Alqaissi et al. [1], a knowledge graph constructed from COVID-19 literature was used to train a random forest model for symptom-based COVID-19 detection. This graph-based model outperformed other models trained on 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 representations have been effectively utilized to enhance machine learning features for disease classification and spectral analysis. Alqaissi et al. [1] used a knowledge graph derived from COVID-19 literature to train a random forest model for detecting COVID-19 based on symptoms,incorporating Fast Random Projection (FastRP) node embeddings to improve performance. Similarly, Wang et al. [28] applied graph representations to classify Raman spectra of oil paper. They constructed graphs by representing spectra as 1023-dimensional vectors based on consistent wave number intervals, calculating Euclidean distances, and applying a Gaussian kernel to measure similarity. This enabled the creation of a Graph Convolutional Network, which successfully classified spectra samples.

3   Methodology

Raman spectra were generated to classify exosomes from normal, hyperglycemic, and hypoglycemic cells using a Horiba LabRAM HR spectrophotometer with a Leica 50X (0.55 N.A.) long-distance objective lens. Spectra were recorded in the 200 to 2000 cm^{-1} range using a 785 nm laser (60.2 mW) with a 10% O.D. filter.

Each acquisition lasted 1 second, with a 4-second exposure time and 9 frame accumulations. The instrument was calibrated according to the manufacturer's guidelines using the Rayleigh line of a silicon wafer [20]. Our dataset includes 3,045 Raman spectra, with fields SpecID, WaveNumber, Absorbance, SurID, and Status. It contains 63 unique SERS surfaces and categorizes the spectra into 915 hyperglycemic, 1,065 hypoglycemic, and 1,065 normal samples.

3.1 Graph Data Modelling

Our novel method for reflecting the uniform nature of the peaks used graph databases. Centrality algorithms were run on many graphs constructed using Neo4j [19] through Python, using the approach in [6] to quantitatively measure the importance of a node in the exosome network. These measures used were: PageRank [3], Degree [11], Eigenvector [31] and Article Rank centrality [22]. We then ran a node embedding algorithm, FastRP [4], to capture high-dimensional aspects of the graph structure in a dense, lower dimensional vector [33].

Peak Grids were used in designing the graph. Specifically, each peak within a spectral sample is represented as a node. Connections (edges) between nodes are established under two conditions: (1) if two peaks originate from the same spectrum or (2) if two peaks are located close to one another in proximity, suggesting that they might correspond to the same biomarker [7]. Our final graph database consists of 27,666 nodes, 4,142,814 edges, and 4,336,476 properties. This approach leverages the spatial relationship and co-occurrence of peaks to identify potential biomarkers, ensuring a more comprehensive representation of the underlying biological patterns within the data.

3.2 OSC: Optimal Spectra Cleaning Method

Before classification, preprocessing spectral data is crucial to eliminate background noise that may obscure the Raman signal. Our OSC process consists of four main steps: cosmic spike removal, baseline correction, smoothing, and scaling. First, we applied the algorithm by Whitaker and Hayes [29] to remove artifacts caused by cosmic ray interference, preserving spectral integrity. Next, baseline correction was performed using the asymmetric least squares algorithm by Eilers and Boelens [8], which effectively removes background fluorescence through asymmetric weighting of deviations from the smooth graph. This method is particularly advantageous as it requires no prior knowledge of peak shapes.

To minimize high-frequency noise, the Savitzky-Golay filter—a robust digital signal processing technique commonly used in spectral analysis—is applied to preserve signal characteristics while effectively reducing noise [9]. Finally, scaling was applied to account for variations in absolute Raman intensity, ensuring the focus remains on the intrinsic composition of the exosome rather than its quantity in the sample [20]. This step also mitigates variability from external factors, such as fluctuations in laser intensity [18]. Three normalization methods were evaluated: scaling to the maximum peak, vector norm scaling, and Standard Normal Variate (SNV) scaling.

Figure 2a presents the raw confocal SERS data (blue) obtained from an exosome-captured gold cavity array, alongside its smoothed version (orange) and baseline-corrected version (green). Meanwhile, Figure 2b displays normalized spectra from the same surface, highlighting both representative spectra and a few noisier outliers.
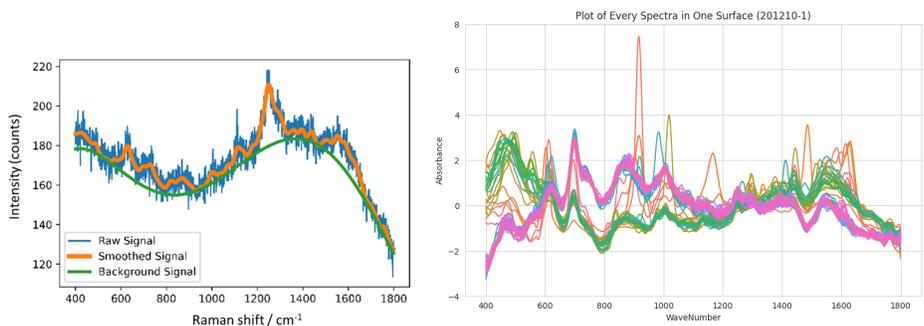
(a) A Raman spectrum at steps (b) Normalized spectra from the same surface

Fig. 2: Examples of Raman spectra after OSC proccesing

However, as seen in Figure 2b, several spectra overlap, suggesting a shared underlying Raman fingerprint. Additionally, some samples deviate from the expected representative spectra, likely due to surface noise or contamination. To assess whether these outliers negatively impacted model performance, we developed methods to identify and remove them.

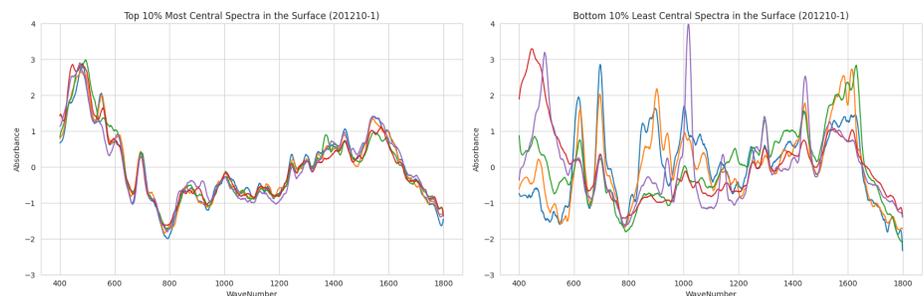### 3.3 OSC+PRF Method: Enhance OSC using a PageRank Filter

Fig. 3: The 10% most and least central spectra within a surface according to OSC+PRF processing

Spectral averaging is commonly used to reduce background interference [13], [10], but this step was not performed in our dataset, resulting in the inclusion of noisy and unrepresentative samples. To address this, we first applied an interquartile range (IQR) method to identify and discard spectra with a high proportion of outlier intensity values (defined as values beyond 1.5 times the IQR). Varying

the outlier proportion threshold led to slight performance improvements and prompted further exploration of noise reduction techniques.

We then analyzed the spectra using a Gaussian kernel-based approach. After applying SNV scaling to enhance shape-based similarity, we constructed Gaussian kernel subgraphs for each surface and computed PageRank centrality scores. Spectra with higher centrality were assumed to better represent the Raman signal, while lower-scoring spectra were treated as noise. A centrality cutoff was used to filter out less relevant spectra, as illustrated in Figure 3.

3.4 OSC+PRF+DR Method: Applying Optimal Dimensionality Reduction to OSC+PRF

To address dimensionality, we explored various feature extraction and selection techniques. Initially, methods such as PCA, LDA, and limiting wave numbers to predefined regions of interest were tested. As each approach resulted in suboptimal model performance, it was necessary to refine the feature set by employing a forward sequential feature selection process, iteratively selecting the wave numbers that contributed most to model performance. Figure 4 presents the average spectra for hyperglycemia, hypoglycemia, and normal conditions after OSC+PRF+DR processing. This approach significantly improved model performance by focusing on the most relevant spectral features. Given the high computational cost of evaluating all possible feature combinations, we limited the selection process to the 50 most influential features, balancing performance gains with computational efficiency.
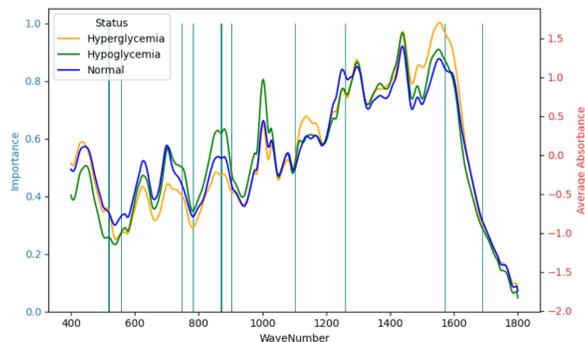

Fig. 4: Average spectra by status via OSC+PRF+DR processing

4 Evaluation and Discussion

4.1 Experimental Setup

To evaluate the performance of our models, we used Group 10-fold cross-validation, as Guo et al. [12] emphasized the risks of using standard k-fold cross-validation with Raman spectra. They recommended splitting data at the highest

hierarchical level, such as by biological or technical replicates, to ensure an unbiased estimate of the generalization error. Group k-fold cross-validation prevents data leakage by ensuring that samples from the same group are not included in both the training and validation sets, thereby providing a more reliable estimate of the model's generalization performance.

Finally, average accuracy, precision, recall, and F1-score across all folds were calculated to evaluate and compare the overall performance of the models. These metrics provide a comprehensive assessment, with accuracy reflecting the proportion of correct predictions, precision measuring the ability to avoid false positives, recall assessing the ability to identify true positives, and the F1-score offering a balanced evaluation by combining precision and recall.

4.2 Results

Table 1 presents the classification performance of various approaches for Raman spectra analysis using group 10-fold cross-validation. Three methods, OSC, OSC+PRF, and OSC+PRF+DR, were evaluated using different models, namely SVM, Random Forest, and Extra Trees. The results indicate that the Extra Trees classifier performed best for OSC+PRF and OSC+PRF+DR, while Random Forest was the optimal choice for OSC. Among the three approaches, OSC+PRF+DR demonstrated the highest performance, achieving an accuracy of 0.760, precision of 0.675, recall of 0.708, and an F1-score of 0.675, outperforming both OSC+PRF and OSC.

Table 1: Classification Performance for Raman Spectra Analysis

Processing Method	Best Model	Accuracy	Precision	Recall	F1-Score
OSC	Random Forest	0.635	0.641	0.648	0.621
OSC+PRF	Extra Trees	0.696	0.609	0.644	0.606
OSC+PRF+DR	Extra Trees	0.760	0.675	0.708	0.675

Table 2 presents the classification performance results for Raman surface analysis using group 10-fold cross-validation. As with the spectra analysis, best-performing models were Random Forest for OSC and Extra Trees for OSC+PRF and OSC+PRF+DR. A similar trend was observed, with OSC+PRF+DR outperforming OSC+PRF and OSC. However, all approaches demonstrated higher performance compared to spectra analysis. OSC+PRF+DR achieved the highest performance across all metrics, with an accuracy of 0.857, precision of 0.873, recall of 0.864, and an F1-score of 0.857.

Table 2: Classification performance of approaches for Raman surface analysis

Processing Method	Best Model	Accuracy	Precision	Recall	F1-Score
OSC	Random Forest	0.683	0.700	0.685	0.687
OSC+PRF	Extra Trees	0.825	0.838	0.831	0.824
OSC+PRF+DR	Extra Trees	0.857	0.873	0.864	0.857

The OSC+PRF+DR approach performed significantly better on Raman surface data compared to Raman spectra data across all classes on group 10-fold cross-validation, as shown in Figure 5. For the Normal class, precision improved from 0.640 to 1.0, and the F1-score increased from 0.591 to 0.842. The Hyperglycemic class showed enhanced precision (0.674 to 0.792) and perfect recall (1.0), leading to an improved F1-score (0.710 to 0.884). Similarly, the Hypoglycemic class showed improved precision (0.712 to 0.826) and recall (0.774 to 0.864), resulting in a higher F1-score (0.725 to 0.844).
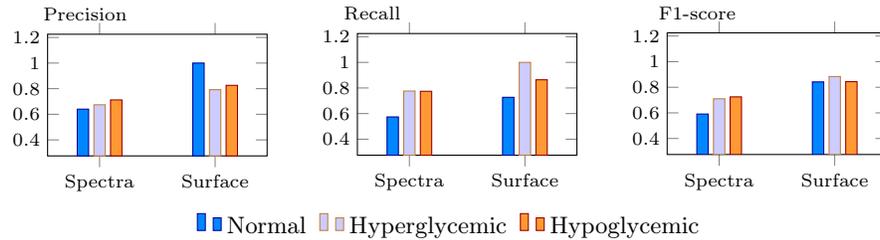

Fig. 5: Average performance per class of OSC+PRF+DR approach using Extra Tress model

Overall, the results indicate that Raman surface analysis delivers superior classification performance compared to Raman spectra analysis across all approaches. This further emphasizes the effectiveness of surface-based Raman data for improved classification performance in medical applications. In addition, OSC+PRF+DR consistently achieved the best performance, making it the most effective method in both scenarios.

5   Conclusion and Future Work

In this study, we explored the use of machine learning for classifying exosomes derived from normal, hyperglycemic, and hypoglycemic cells based on their Raman spectra or surface. To improve classification accuracy, we applied several preprocessing and feature selection techniques, including OSC, OSC+PRF and OSC+PRF+DR. The results showed that OSC+PRF+DR, with Extra Trees as the classifier, consistently outperformed other methods, achieving the highest performance in both spectra and surface analyses. Specifically, OSC+PRF+DR resulted in an accuracy of 0.760, precision of 0.675, recall of 0.708, and an F1-score of 0.675 for spectra data, and an accuracy of 0.857, precision of 0.873, and an F1-score of 0.857 for surface data in a 10-fold cross-validation.

The study also found that Raman surface analysis provided superior classification results compared to Raman spectra analysis, highlighting its potential for medical applications. By integrating OSC, OSC+PRF, and OSC+PRF+DR, the approach improved spectral cleaning, noise reduction, and feature selection, thereby enhancing classification accuracy. These findings emphasize the promise

of Raman spectroscopy and graph-based techniques for biomarker identification, suggesting that future work could refine feature selection and investigate outlier detection methods to further enhance performance.

Future work should focus on addressing potential overfitting to specific laboratory or spectrometer conditions, as all samples were sourced from the same machine. Acquiring additional samples under varied conditions would help assess model performance on independent data. Additionally, using nested cross-validation would further validate the OSC+PRF and OSC+PRF+DR, ensuring a more reliable evaluation of the models' generalization ability.

Acknowledgment

References

1. Alqaissi, E., Alotaibi, F., Ramzan, M.S.: Graph data science and machine learning for the detection of covid-19 infection from symptoms. PeerJ Computer Science **9**, e1333 (2023)
2. Amjad, A., Ullah, R., Khan, S., Bilal, M., Khan, A.: Raman spectroscopy based analysis of milk using random forest classification. Vibrational Spectroscopy **99**, 124–129 (2018)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks **56**(18), 3825–3833 (2012)
4. Chen, H., Sultan, S.F., Tian, Y., Chen, M., Skiena, S.: Fast and accurate network embeddings via very sparse random projection. In: CIKM'19. p. 399–408 (2019)
5. Chen, X., et al.: Raman spectroscopy combined with a support vector machine algorithm as a diagnostic technique for primary sjögren's syndrome. Scientific Reports **13**(1) (2023)
6. Cuong, D.V., Ngo, V.M., Cappellari, P., Roantree, M.: Analyzing shared bike usage through graph-based spatio-temporal modeling. IEEE Open Journal of Intelligent Transportation Systems **5**, 115–131 (2024)
7. Das, R.S., Agrawal, Y.K.: Raman spectroscopy: Recent advancements, techniques and applications. Vibrational Spectroscopy **57**(2), 163–176 (2011)
8. Eilers, P., Boelens, H.: Baseline correction with asymmetric least squares smoothing (2005)
9. Gallagher, N.: Savitzky-golay smoothing and differentiation filter (2020)
10. Gautam, R., Vanga, S., Ariese, F., Umapathy, S.: Review of multidimensional data processing approaches for raman and infrared spectroscopy. EPJ Techniques and Instrumentation **2**(1) (2015)
11. Gleich, D.F.: Pagerank beyond the web. SIAM Review **57**(3), 321–363 (2015)
12. Guo, S., Bocklitz, T., Neugebauer, U., Popp, J.: Common mistakes in cross-validating classification models. Analytical Methods **9**(30) (2017)
13. Hu, S., et al.: Raman spectroscopy combined with machine learning algorithms to detect adulterated suichang native honey. Scientific Reports **12**, 3456 (2022)
14. Kalluri, R.: The biology and function of exosomes in cancer. Journal of Clinical Investigation **126**(4), 1208–1215 (2016)

15. Kalluri, R., LeBleu, V.S.: The biology, function, and biomedical applications of exosomes. Science **367**(6478) (2020)
16. Kopeć, M., Beton, K., Jarczewska, K., Abramczyk, H.: Hyperglycemia and cancer in human lung carcinoma by means of raman spectroscopy and imaging. Scientific Reports **12**(1), 18561 (2022)
17. Li, J., et al.: Exosome detection via surface-enhanced raman spectroscopy for cancer diagnosis. Acta Biomaterialia **144**, 1–14 (2022)
18. Martyna, A., 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 **202**, 104029 (2020)
19. Neo4j Inc.: Neo4j graph platform – the leader in graph databases (2023), `https://neo4j.com`
20. O'Sullivan, J.: Development of Bioplasmonic Platforms for Extracellular Vesicle Capture and Analysis. Master's thesis, Dublin City University, Ireland (2022)
21. Qin, J., Xu, Q.: Functions and application of exosomes (2014)
22. Ruhnau, B.: Eigenvector-centrality — a node-centrality? Social Networks **22**(4), 357–365 (2000)
23. Schumacher, W., Stöckel, S., Rösch, P., Popp, J.: Improving chemometric results by optimizing the dimension reduction for raman spectral data sets. Journal of Raman Spectroscopy **45**(10), 930–940 (2014)
24. Shin, H., et al.: Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. ACS Nano **14**(5) (2020)
25. Sierra, J., Marrugo-Ramírez, J., Rodriguez-Trujillo, R., Mir, M., Samitier, J.: Sensor-integrated microfluidic approaches for liquid biopsies applications in early detection of cancer. Sensors **20**(5), 1317 (2020)
26. Sun, Y., Tao, Q., Wu, X., Zhang, L., Liu, Q., Wang, L.: The utility of exosomes in diagnosis and therapy of diabetes mellitus and associated complications. Frontiers in Endocrinology **12** (2021)
27. de la Torre Gomez, C., Goreham, R.V., Serra, J.J.B., Nann, T., Kussmann, M.: 'exosomics'—a review of biophysics, biology and biochemistry of exosomes with a focus on human breast milk. Frontiers in Genetics **9** (2018)
28. Wang, Z., Chen, W., Zhou, W., Zhang, R., Song, R., Yang, D.: 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), 1892–1900 (2021)
29. Whitaker, D.A., Hayes, K.: A simple algorithm for despiking raman spectra. Chemometrics and Intelligent Laboratory Systems **179**, 82–84 (2018)
30. Xie, Y., Su, X., Wen, Y., Zheng, C., Li, M.: Artificial intelligent label-free sers profiling of serum exosomes for breast cancer diagnosis and postoperative assessment. Nano Letters **22**(19), 7910–7918 (2022)
31. Zhang, J., Luo, Y.: Degree centrality, betweenness centrality, and closeness centrality in social network. In: the 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017) (2017)
32. Zhang, L., et al.: Raman spectroscopy and machine learning for the classification of breast cancers. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **264**, 120300 (2022)
33. Zhou, J., Liu, L., Wei, W., Fan, J.: Network representation learning: From preprocessing, feature extraction to node embedding. ACM Computing Surveys **55**(2), 1–35 (2022)