Machine Learning of iKnife Data to Determine Tissue-Specific Biomarkers for Rapid Diagnosis of Cancer

Teerapon Sahwangarrom[1], James S. McKenzie[1,2], Yuchen Xiang[1,2], Zoltan Takats[1,2*]

[1]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, W12 0NN London, United Kingdom
[2]Department of Metabolism, Digestion, and Reproduction, Faculty of Medicine, Imperial College London, SW7 2AZ London, United Kingdom
[*]Principle supervisor

ABSTRACT

Background: A tumour-free surgical resection margin is a significant prognostic factor for many neoplasms. Scientists have developed the intelligent knife (iKnife) to assist surgeons by improving the accuracy of malignant tissue clearance. The iKnife uses mass spectrometry to measure the tissue's metabolomic composition, i.e. information regarding the biochemical metabolic profile that facilitates tissue type differentiation between cancer and normal tissues types. Previous studies have examined smaller cohorts of samples within single tissue type. The single-tissue approach can be effectively used for distinguishing between normal and cancer tissue types. However, these metabolite changes may not be relevant in other tissue types. It was therefore hypothesised that a small subset of molecular features that were recognised across more than one type of tissue may be powerful diagnostic biomarkers and can be used to improve tumour identification. ***Materials and Methods:*** Mass spectrometry data from the iKnife was used to test the hypothesis. Initially, mass spectrometry data were normalised using Probabilistic Quotient Normalisation (PQN) and then analysed using a wide range of machine learning techniques in order to find the optimal algorithm using repeated stratified K-fold cross-validation for 10 iterations. The final predictive model was utilised to identify a group of informative features using recursive feature elimination for 100 iterations. Then, the Kruskal-Willis test was employed to emphasise whether differences in the median intensities of the selected features were statistically significant between cancer and normal tissues. ***Results:*** The top 80 molecular features (m/z) from 931 features, which were selected by RFE, improved the classification performance between normal and cancer tissues from 90.7% to 95.1%. Also, the eight statistically significant features were identified based on the criteria of FDR < 0.01 by the Kruskal-Willis test and selection rate by RFE more than 30%. ***Discussion:*** Since the tissue types were imbalanced, the samples needed to be carefully separated into training and testing datasets. Stratified K-Fold cross-validation was selected as it maintains the balance or ratio between dataset labels. Our results of the subset of molecular features have been documented in many previous publications. These informative molecular features could be potential diagnostic markers that could distinguish between normal and cancer tissues across a range of cancer types.

BACKGROUND

Cancer is a disease that occurs as a result of cellular changes that create an uncontrolled growth and division of cells. In the United Kingdom, over 400,000 new cancer cases are diagnosed annually, while there are roughly 165,000 cancer deaths (*1*). There are several kinds of cancer treatments, including surgery, radiation therapy, and chemotherapy. Surgical excisions are the standard treatments for most solid tissue tumours. From an oncological perspective, a curative surgical intervention occurs when tumour tissue is completely removed, and any associated border of microscopically normal tissue (i.e. the tumour margin with clearances) (*2*). A previous study suggests that approximately 20% of breast cancer patients who are treated with breast-conserving surgery (lumpectomy) subsequently require surgery to remove all positive margins; therefore, onco-surgical techniques are generally not adequate cancer treatments (*3*). This highlights a major concern because the surgical resection margin is a significant prognostic factor for many cancers (*4, 5*). Moreover, if a surgeon is unable to completely excise all the malignant tissue, this indicates that residual tumour remains with the patient, and there is a risk of reoccurrence of the disease and increased patient mortality (*6*). Therefore, technological innovations should be developed that can aid surgeons in improving the accuracy of cancerous tissue clearance.

Biochemical differences need to be identified, to aid in developing a device that can identify cancer and normal tissues. The different histological characteristics of cancer and normal tissues suggest that their respective lipidomic and metabolic profiles might be dissimilar (*7*). Currently, magnetic resonance imaging is performed to aid in the delineation and localisation of the areas that need surgical excisions (*8*). It also guides surgeons' interpretation and pathology confirmation during oncological surgical planning. However, magnetic resonance often requires approximately 30 minutes, and it is costly because the

removed tissue has to be sent to the pathology laboratory for intraoperative histological examination. Moreover, it only allows a limited number of sampling points, and its histology could be interpreted subjectively, particularly in cases of non-oriented specimens or specimens that are sub-optimally prepared (*8*).

To resolve this issue and create a device that can be used during surgery, electrosurgical devices have been developed for improved accuracy of dissection, and they are currently used in surgical procedures such as gynaecologists' surgeries, cancer surgery, and others (*9*). These devices apply electric current to thermally destroy the targeted tissues. Also, electrosurgery devices are employed during surgical procedures to cut, coagulate, desiccate, and fulgurate targeted tissues (*10*). The device coupling with mass spectrometry has been developed in an attempt to create electrosurgery devices for real-time identification of tumour tissues, because mass spectrometry is a powerful technique that can be used to investigate the metabolic characteristics of the histological sections of multiple tissue types, including cancers (*11*).

Professor Zoltan Takats and his colleagues at the Imperial College London have developed the intelligent knife (iKnife), which is an electrosurgical mass spectrometry process that offers multivariate analysis for cancer diagnosis (*12*). A major by-product of electrocautery's use is that it creates smoke from evaporating tissues as they are being resected. It uses mass spectrometry to measure the tissue's metabolomic composition, and this offers biological information about the tissues (*13*). The concept of tumour classification with iKnife is that electricity heats the tip of the iKnife. Subsequently, the hot blade results in an explosion in the cells within the tissue, thereby releasing molecules in the smoke. However, surgical smoke is considered to be toxic; therefore, the smoke is sucked into a tube and fed into a mass spectrometer (*2*). The mass spectrometer analyses the molecules and creates a fingerprint. The fingerprint offers information about the type of tissue that is currently being cut to scientists. The results of the mass spectrometric profiles are specific to the type of tissue that is analysed, which enables tissue identification and characterisation that is similar to histopathological analysis (*14*). Also, this analytical coupling creates new biochemical information sets that describe the tissue and its associated pathology.

In typical iKnife experiments, a wide range of mass spectrum is scanned to generate the required fingerprint. However, the majority of the collected data is redundant and can be considered as, for example, noise (either chemical or instrumental), or molecular features that do not differentiate between tissue types (*15*). To remove the redundant data for mass spectrometry, the mass spectrometry data is subsequently analysed using machine learning techniques that can identify a smaller panel of features that have enhanced diagnostic potential (*16*).

Moreover, previous studies have examined smaller cohorts of samples (typically < 300 samples) within a single tissue type (*2*). The single-tissue approach can be effectively used for differentiating between cancer and normal tissue types. However, these metabolite changes may not be relevant in other tissue types. Nonetheless, it could be possible to identify diagnostic markers that are found across more than one type of tissues when there are more samples across a range of tissues (*17*). Furthermore, the analysis of cancer metabolites within a range of tissue types may help to develop a greater mechanistic insight into cancer (*18*).

Accordingly, it was hypothesised that a small subset of features that are identified across more than one type of tissues may be powerful diagnostic biomarkers (universal biomarkers), which can improve the classification of tumours and normal tissues. The overall aim of this project was to identify a small panel of molecular features that could effectively differentiate normal from cancer tissues, over a range of tissue types (breast, ovarian, and colorectal cancers). In an attempt to find the best algorithm that could classify normal and cancer tissues accurately, the iKnife data were subjected to a wide range of machine learning techniques. Subsequently, the final predictive model was combined with a powerful feature-selection algorithm and used to identify informative molecular features.

MATERIALS AND METHODS

Data acquisition and data normalisation

The mass spectrometric profiles of specimens were retrieved from the oracle database at Professor Takats' laboratory. The mass spectrometry data included 931 columns (from mass-to-charge ratio 150 to 997 range) and 1,755 observations. The mass-to-charge ratio (m/z) of the molecular ion is equal to the molecular weight of the compound, i.e. lipids or metabolites. The dataset comprised of 1,755 samples consisting of a range of cancer types, including breast cancer (n = 1367: cancer, n = 561; normal, n = 806), colorectal cancer (n = 304: cancer, n = 106; normal, n = 198), and ovarian cancer (n = 84: cancer, n = 30; normal, n = 54). After the acquisition of the mass spectrometry, the raw data files were imported into Python programming to perform machine learning using scikit-learn package (*19*). Initially, duplicates were removed from the dataset before the stage of data normalisation began.

Subsequently, Probabilistic Quotient Normalisation (PQN), which is a normalisation method, was used to transform the mass spectrometric profiles. This normalisation technique

has been reported to be significantly robust and accurate, in comparison to integral and the vector-length normalisations (*20*). The PQN assumes that the changes in the biochemical interesting concentration impact only parts of the mass spectrum, while dilution will impact the metabolite signals. This method began with the calculation of the reference spectrum, based on the median spectrum. Subsequently, the quotient of a given test spectrum and reference spectrum was calculated for each variable of interest, and the median of all quotients was estimated. All the variables of the test spectrum were also divided by the median quotient (*21*). Then, the log transformation was used to convert skewed mass spectrometry data to symmetry data, with the use of the median of the non-zero intensities via a non-linear transformation (*22*). Generally, this was used to adjust heteroscedasticity and handle the zero-value of the mass spectrometry data.

The study of mass spectrometry data using unsupervised machine learning technique

Unsupervised learning refers to the training of machines with information that is neither classified nor labelled, and also enabling the algorithm to act on that information without guidance. Principal Component Analysis (PCA) is an unsupervised machine learning technique that attempts to derive a set of low-dimensional features from a bigger set while preserving the maximum variance (*23*). PCA was applied to map multidimensional mass spectrometry data into an uncorrelated set of components, thereby capturing most of the variations within the dataset. The PCA plot of the first two components was used to examine the similarities and differences in the molecular ion composition of tissue specimens. Considering that the principal components are ordered based on their variability percentage, the first principal component (PC1) is always the axis that explains the maximum variance in the data. Therefore, the PCA loading plot was constructed to determine which features were influential in separating cancer from normal tissue along the PC1 direction (*24*).

Generation of predictive models using a supervised machine learning technique

Supervised machine learning refers to the training of machines using a full set of labelled data, while concurrently training an algorithm to create predictive models. To construct the predictive model and find the optimal machine learning algorithm for normal and cancer classifications and, the normalised data were used to evaluate varying machine learning models, including Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Neighbours Classifier (KNN), Decision Tree Classifier (CART), Gaussian Naïve Bayes (NB), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). All the classification models were validated using repeated stratified K-Fold cross-validation for 10 iterations. They were configured with the same random seeds to ensure that the same splits were performed for the training data, and each algorithm was evaluated using a similar method. The repeated stratified K-Fold cross-validation was used because stratification is the process of rearranging the data, to ensure each fold is a good representative of the whole data (*25*). The finalised predictive model was the algorithm that showed the highest accuracy of the classification performance.

Feature selection using recursive feature elimination (RFE)

Feature selection is a method that is intended to reduce the number of input variables to only the most useful ones for a model, to effectively predict the target variable (*26*). The two benefits of performing feature selection are: (i) it can help to improve the accuracy of classification performance, and (ii) it helps to reduce the training time because irrelevant features are removed, and less data indicate that the algorithms train faster (*27*).

Recursive feature elimination (RFE) is a wrapper-type feature-selection algorithm that uses filter-based feature selection internally (*28*). The idea of RFE works by searching a group of molecular features by: (i) starting with all the features in the training data, (ii) ranking the features by importance (model coefficients), (iii) removing the least important features, and (iv) refitting a machine learning model (*29*). This process is repeatedly performed until a specified number of molecular features remain. The RFE method was accessed via the RFE class in scikit-learn. The RFE was performed for 100 iterations and repeated stratified K-Fold was used to split the dataset into 80% for training data and 20% for testing data for each iteration. The 30 most useful molecular features for each iteration was selected using the optimal model.

Subsequently, the selected features were summarised the number of selections for all iterations and then calculated the percentage of selection rates. After that, the selected features were arranged from the highest selection rate to the lowest selection rate and then they were used to examine the classification performance for any improvements. The learning curve was also constructed to determine how classification improves when there is an increase in the number of selected molecular features. Next, a receiver operating characteristic curve (ROC) was used to compare and visualise the predictive performance before and after the feature selection. The ROC curve offers information about the model's capability to distinguish between normal and cancer tissues. The ROC curve was plotted with true positive rates against the false-positive rates. The true positive rates are on the y-axis (sensitivity)

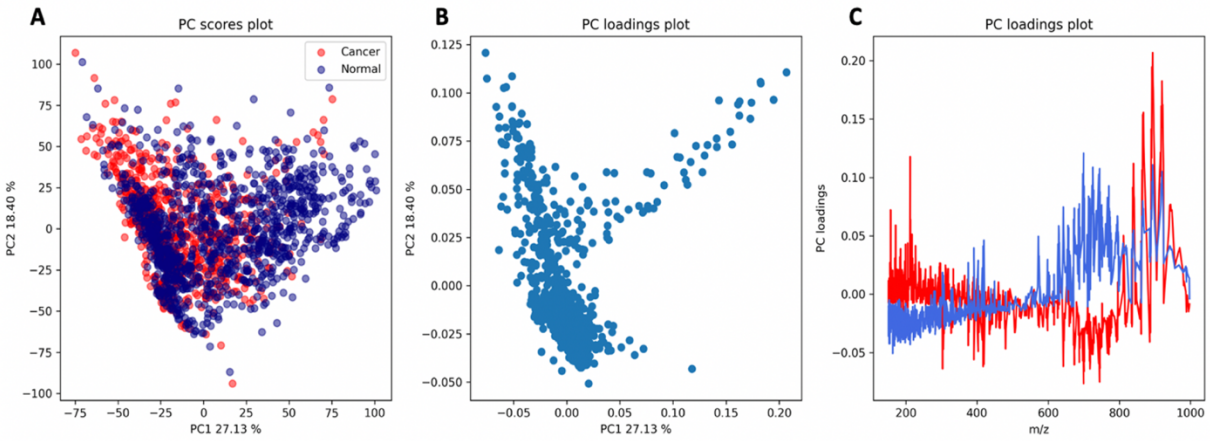

Figure 1. Principal component analysis plot of iKnife data. (A) A PCA score plot of cancer and normal tissues: cancer tissues (red) and normal tissues (blue). The x-axis represents PC1, while the y-axis portrays PC2. (B) A PC loading plot of PC1 and PC2 for mass spectra (m/z 150-997). The x-axis represents PC1, while the y-axis represents PC2. (C) A PC loading plot of PC1 and PC2 for the mass spectra: PC1 (red) and PC2 (blue). The x-axis represents the mass spectra range from 150 to 997, while the y-axis shows the PC loading values.

and false-positive rates (1 – specificity) are on the x-axis. The false-positive rates offer information about the number of normal tissues that were incorrectly classified (*30*). Also, a learning curve was plotted to illustrate the fit time of the final predictive model for training data. This aided with comparing the training time (fit time) between before and after feature selection.

Univariate statistical analysis

The univariate statistical analysis was conducted with the Kruskal-Wallis test to make the selected features, which were selected by RFE, more reliable. It was also used to determine whether the difference in the median values of the individual peak intensities of the selected features across the tissue types (*31*). The Bonferroni adjustment was used to account for multiple comparisons (*32*).

RESULTS

Principal component analysis (PCA analysis)

This section examines the clustering between observations and highlights certain useful features that can be valuable for separating cancer tissues from normal tissues. After the normalisation of the mass spectrometry data, the PCA plot was constructed to visualise the clustering between normal and cancer tissue samples (Figure 1). The principal components are ordered by the percentage of variability that they denote. As previously mentioned, generally, the first principal component (PC1) is the axis that explains the maximum variance direction in the data.

For this study, the PC1 showed 27.13% of the total variation between the principal components, while the second principal component (PC2) accounted for 18.40% (Figure 1A). According to the plot, both PC1 and PC2 showed the majority of the variation. Since PC3 represented 9.27% of the total variation, this indicated that a two-dimensional

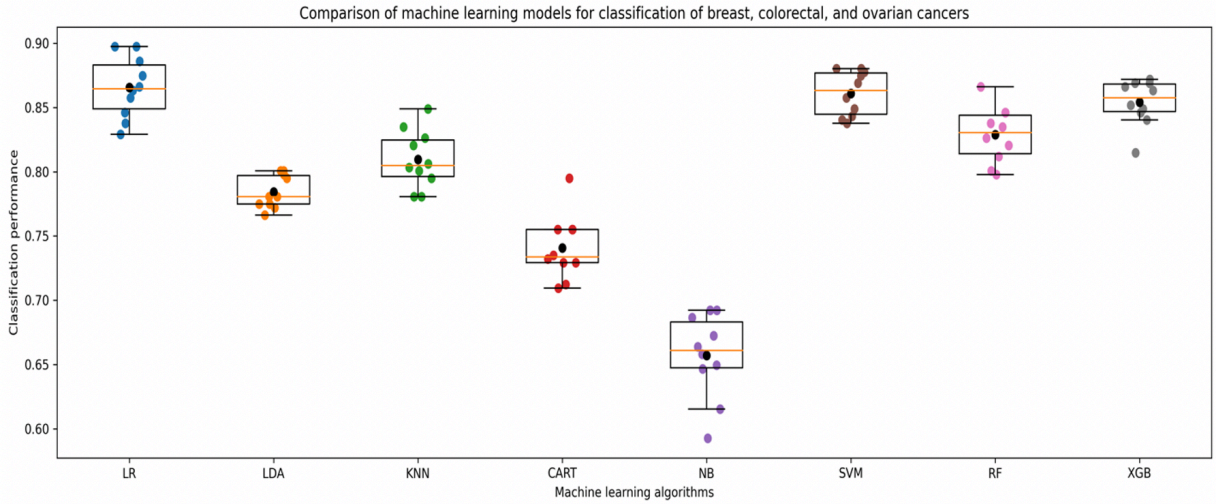

Figure 2. Boxplot of evaluation of machine learning algorithms using repeated stratified K-Fold cross-validation for 10 iterations. The x-axis shows eight machine learning models, while the y-axis represents the classification performance. The black dot explains the mean of accuracy scores for each algorithm. The top and bottom of the coloured band portray, the 75th and the 25th percentiles of the group respectively with the median denoted by the yellow line. The jitter represents the classification performance for each iteration.

graph that uses PC1 and PC2 would offer a good approximation of the mass spectrometry data, considering that it would account for 45.53% of the variation in the data (Figure 1A).

Since the PC1 denotes the maximum variance direction in the data, the PCA loading plot was constructed to identify which features impact the separation between the two clusters (cancer and normal) along the x-axis, and how these features are correlated (Figure 1B). The variables that offer similar information were grouped because they were correlated. When the mass spectrometry value for one variable increases or decreases, the mass spectrometry value for the other variable also tends to change in a similar direction. Furthermore, the distance to the origin conveys certain information. For instance, the farther a variable lies from the plot's origin, the stronger the impact of such a variable on effectively separating cancerous tissues from normal tissues. Therefore, the peaks that contributed the most to the first two principal components function as tissue-specific indicators. The intensity distribution of the mass-to-charge ratio (m/z) - 893.737, 768.553, 171.138, and 687.499 provided a relatively higher result (loading coefficients > 0.05) (Figure 1C). However, according to the PCA plot, the clusters of cancerous and normal tissues overlapped (Figure 1A). Thus, the results of the PCA could not be used to highlight any useful features that could be used to separate cancerous tissues from normal tissues.

The logistic regression model was the final predictive model for the tumour classification

To determine the optimal machine learning algorithm that was suitable for the mass spectrometry data, eight machine learning algorithms were compared with their classification performance using repeated stratified K-Fold cross-validation. The top machine learning algorithm that provided the highest accuracy score of classification was the LR model (0.874 ± 0.022), which was the mean of accuracy (87.4%) with standard deviation (2.2%) (Figure 2). These findings suggested that it would be valuable to conduct further studies on the analysis of the LR concerning tumour classification.

The top 80 selected molecular features improved LR performance for the classification

Next, this study attempted to identify a subset of the most informative molecular features, which a group of variables that could effectively differentiate between tumour and

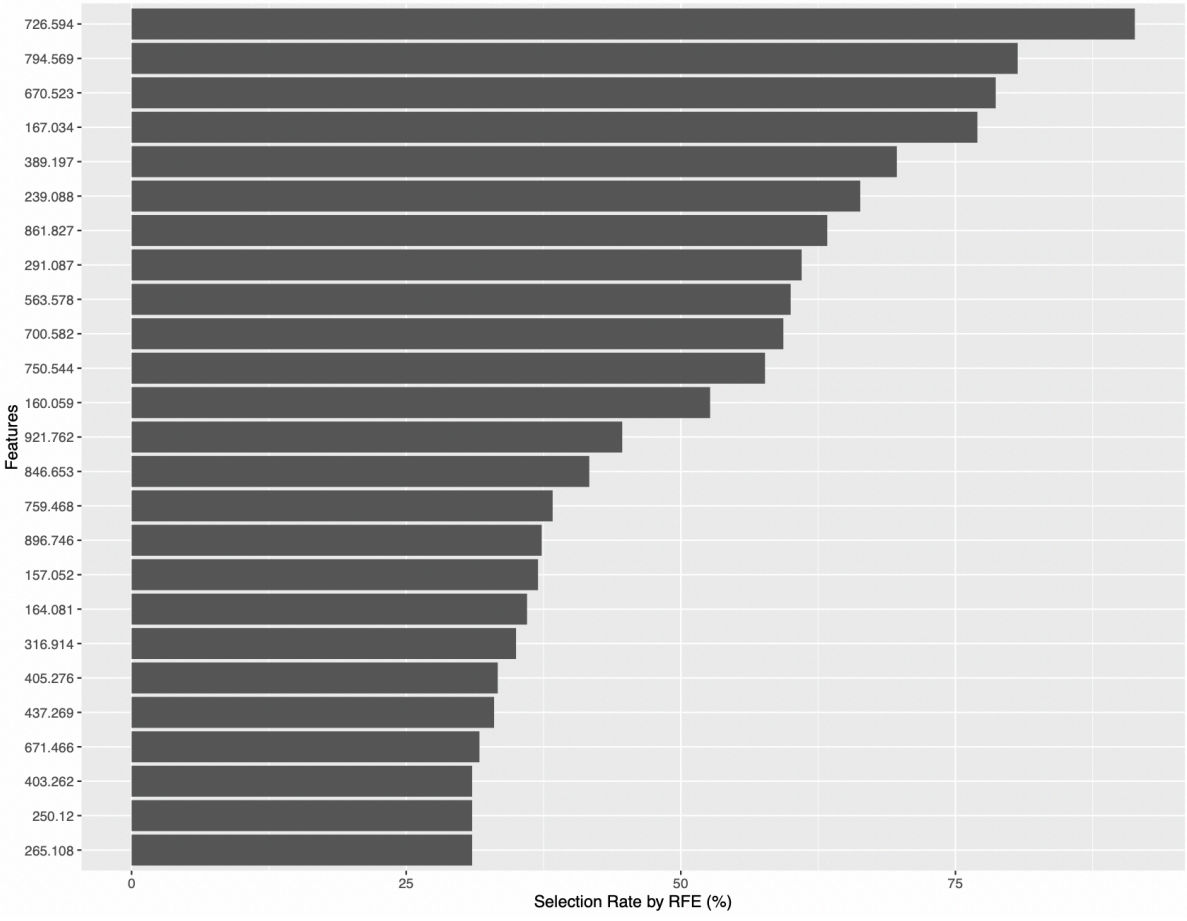

Figure 3. Selection rate of the top 25 features selected by RFE. The x-axis represents the selection rate (%) by RFE, which was performed 100 iterations. The y-axis portrays the top 25 features, which were selected by RFE more than 30%.

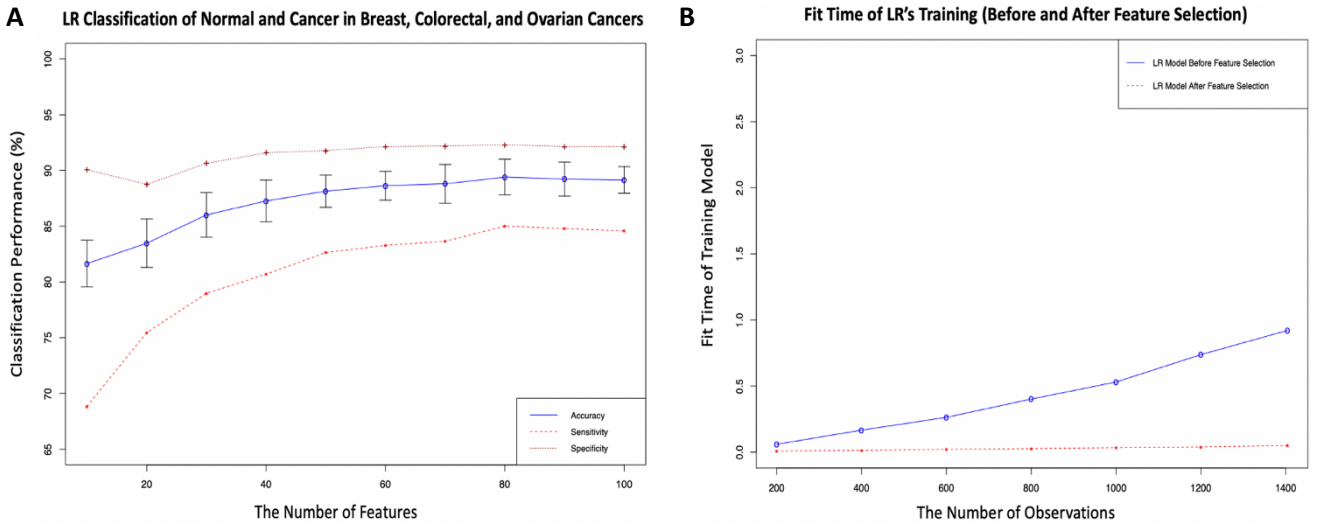

Figure 4. The performance of LR model after the feature selection. (A) The learning curve of predictive performance of the LR model responding to the top 100 selected features by RFE. The x-axis represents the number of selected features (from top 10 to top 100), while the y-axis shows the classification performance. (B) The learning curve depicting the fit time for training the LR model before and after the feature selection: the LR performance before the feature selection (blue) and the LR performance after the feature selection (red). The x-axis represents the number of samples that denote training data, accounting for 80% of the entire samples in the data, while the y-axis shows the fit time for training the LR model.

normal tissues within a range of cancer types. Recursive feature elimination (RFE) was used to identify the useful molecular features by eliminating features from the training dataset. Subsequently, to the LR, RFE was conducted using repeated stratified K-Fold and splitting the data (80% of the entire dataset for training data, 20% of the entire dataset for testing data) for 100 iterations. The m/z - 726.594, 794.569, and 670.523 - illustrated the most selected molecular features, which account for 91%, 81%, and 79% of the total number of iterations respectively (Figure 3).

After the feature selection, 362 selected features had to be investigated to determine their impact on the improvement of predictive performance and see how predictive performance improved. The learning curve was constructed to determine how the predictive performance of the LR model was improved when a subset of the selected features was changed. The results showed that the top 80 selected molecular features improved the predictive performance, with the highest accuracy given as 89.40% ± 1.60% (Figure 4A). Furthermore, before and after the feature selection, the ROC curve was used to compare and visualise the performance of the classification. The ROC curve offers information about the model's capability to distinguish between normal and cancer tissues. The ROC curve was plotted with the true positive rates against the false positive rates, whereby the true positive rates are plotted on the y-axis (sensitivity) and the false-positive rates (1 – specificity) are on the x-axis. Sensitivity offers information about the percentage of patients with cancer who were correctly identified. The false-positive rate offers information about the number of normal tissues that were incorrectly classified. This result was evaluated against the classification performance before the feature selection was conducted. The predictive performance of the LR algorithm with the top 80 features improved its accuracy score from 90.7% to 95.1% (Figure 5). Furthermore, after the feature selection, the training time for the LR model was reduced in comparison to the training time before the feature selection (Figure 4B). This finding suggests that the top 80 features would have the most informative molecular features that could improve the accuracy of classification between normal and tumour tissues.

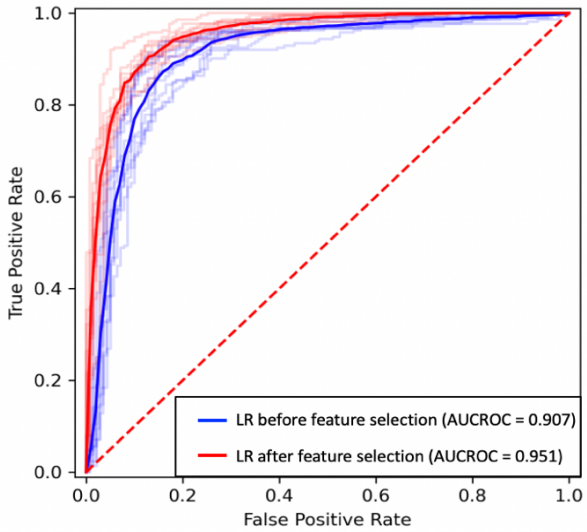

Figure 5. The ROC curve of the LR's performance after feature selection. ROC curve of the LR model: LR performance before feature selection (blue), LR performance after feature selection (red). The x-axis portrays false positive rate, while the y-axis shows true positive rate. The light bule and red colours show the ROC curve for 10 iterations. However, the dark blue and red colours represent the mean of true positive rate. The diagonal line (red dashed line) depicts perfect chance of random classification.

The Kruskal-Wallis test investigated the top 80 selected features

The Kruskal-Wallis test was performed to confirm whether the differences in the median intensities of the top 80 selected features were statistically significant between normal and cancer tissues. There were eight statistically significant features, which were selected by an RFE that was more than 30%. The most significant m/z that was detected in cancer tissues in comparison to normal tissues, was 389.197 (FDR = 6.74E-49), which was selected by the RFE for 69.7% (Figure 6C). However, the m/z value of 896.746 (FDR = 1.75E-19), which was selected by the RFE for 33%, was the most significant m/z that was likely to be found in normal tissues in comparison to cancer tissues (Figure 6C). The mass spectrometry experiments were performed to identify the lipidomic and metabolite species. In comparison to healthy tissue, cancer tissue had the top four ion species that were significantly identified: 794.569, 389.197, 671.466, and 700.582 (Figure 6A and 6C). However, the normal tissue had the top four ion species that were significantly identified: 157.052, 896.746, 921.762, 164.081 (Figure 6B and 6C).

DISCUSSION

The purpose of this project is to identify a small panel of molecular features that can effectively distinguish between cancer and normal tissues over a range of cancer types including breast, colorectal, and ovarian cancers. An analysis was performed using iKnife mass spectrometry data. Initially, a large number of tissue samples, i.e. > 10,000, were present. However, many of them had mislabelled information with respect to malignant or normal tissues. Since the tissue types were imbalanced, the samples needed to be separated carefully into training and testing datasets. If the tissue types were large enough, the technique of leave tissue type out was selected using group shuffle split cross-validation. Cross-validation is a resampling method which is used to evaluate machine learning models on a limited data sample. The group shuffle split iterator generates a sequence of randomised partitions in which a subset of groups, i.e. tissue types are held out for each split. However, if the sizes of tissue types were not large enough, a suitable technique of cross-validation iterators was required to handle this issue. There are three types of cross-validation iterator techniques that were candidates to be used in this study, i.e. K-fold, stratified K-fold, and shuffle split (Figure 7). The K-fold procedure is popular and easy to understand and frequently results in a less biased model compared to other methods (*33*). It ensures that every observation from the original dataset has the opportunity to appear in the training and testing dataset; this represents one of the best approaches if there is limited input data (Figure 7A). The stratified K-fold technique is a variation of the K-fold method. In contrast to the K-fold

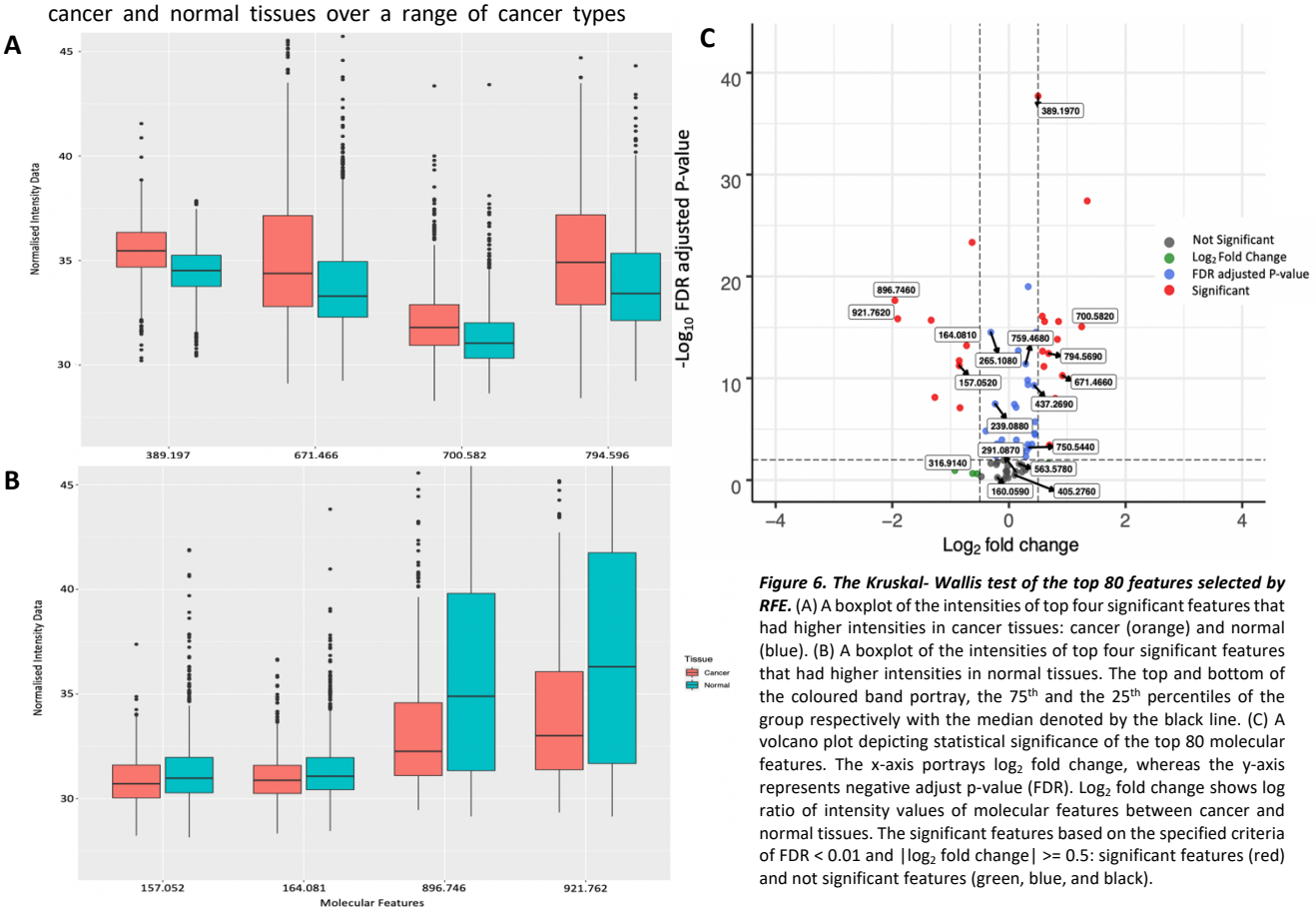

Figure 6. The Kruskal- Wallis test of the top 80 features selected by RFE. (A) A boxplot of the intensities of top four significant features that had higher intensities in cancer tissues: cancer (orange) and normal (blue). (B) A boxplot of the intensities of top four significant features that had higher intensities in normal tissues. The top and bottom of the coloured band portray, the 75th and the 25th percentiles of the group respectively with the median denoted by the black line. (C) A volcano plot depicting statistical significance of the top 80 molecular features. The x-axis portrays $\log_2$ fold change, whereas the y-axis represents negative adjust p-value (FDR). Log_2 fold change shows log ratio of intensity values of molecular features between cancer and normal tissues. The significant features based on the specified criteria of FDR < 0.01 and $|\log_2$ fold change$|$ >= 0.5: significant features (red) and not significant features (green, blue, and black).

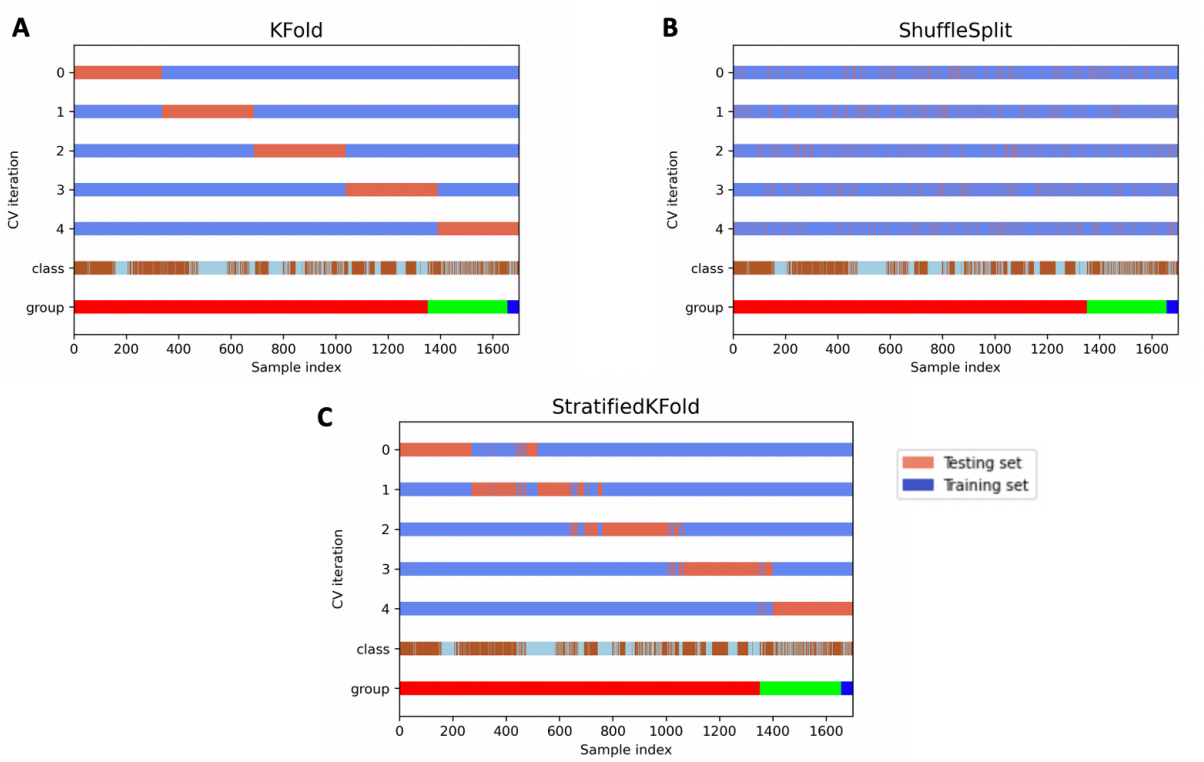

Figure 7. The visualization of three cross-validation behaviours for five iterations. (A) K Fold cross-validation. (B) Shuffle split cross-validation. (C) Stratified K Fold cross-validation. The x-axis represents the index of patient samples, whereas the y-axis portrays cross-validation iterations that the data are separated into testing set (orange) and training set (blue). The class shows cancer (light blue) and healthy tissues (brown). The group represents three cancer types: breast cancer (red), colorectal cancer (green), and ovarian cancer (dark blue).

technique, stratification works by maintaining the balance or ratio between dataset labels (Figure 7C). If the entire dataset has two labels, e.g. cancer and normal tissues and these have a 40:60 ratio, each stratified subsample should keep the same ratio (*33*). The shuffle split procedure randomly separates the entire data during each iteration to generate training and testing sets (Figure 7B). When compared with K-fold cross-validation, the shuffle split method is more efficient, but it is more likely to be prone to sampling bias (*33*). Furthermore, since machine learning model performance is very sensitive to sample balancing, using the stratified K-fold strategy often makes the model more stable for subsamples.

The recursive feature elimination (RFE) is a wrapper feature selection algorithm. It is one of the best and most effective feature selection procedures based on the concept of repeatedly constructing the logistic regression model (the final predictive model) and selecting either the best performing features based on the coefficient or feature importance (*34*). In this study, the stability and robustness of the RFE method was increased by performing RFE for 100 iterations in order to select a small group of features that were strongly informative and effective. The informative molecular features were chosen based on the number of counts where a feature was deemed to be important. According to the ROC curve, the top 80 molecular features selected by RFE could improve the LR performance from 90.7% (before the feature selection) to 95.1% (after the feature selection).

The Kruskal-Wallis test was used to investigate the difference in the median values of the selected features, i.e. peak intensities across cancer and normal tissues. We found that some selected features' false discovery rate (FDR adjusted p-value), which were highly selected by RFE, were not significant on the Kruskal-Willis testing, i.e. the m/z -726.594 (FDR adjusted p-value = 0.48, 91% for the selection rate). A potential reason for this was that the isotope species highly correlated and those which would have a low false discovery rate (FDR adjusted p-value), were removed because they are weaker features according to feature importance. The isotope species are chemical elements with a similar atomic number with practically identical chemical behaviours, but they differ in atomic masses and physical properties. For example, feature A and B are correlated; they may not have high importance. However, when feature B is removed, the importance of feature A increases, thus making feature A an important feature. Moreover, in comparison to the Kruskal-Wallis test, we performed the feature selection using RFE, which was selected variables based on the coefficients of the logistic regression model (the predictive model), which was assigned the weights to each feature.

In this study, eight statistically significant features were identified, which were selected by an RFE of more than 30%. The intensities of the four significant features (m/z) that had higher intensities in cancer tissue compared to healthy tissue were: 794.569, 389.197, 700.582, and 671.466. The intensities of the four significant features (m/z) that had higher intensities in normal tissue as compared to cancer tissue were: 921.762, 896.746, 157.052, and 164.081. A study by St John *et al* (2017) reported that the m/z 921.762 was detected as a statistically significant feature (FDR adjusted p-value < 0.01) with high intensities in normal tissue to differentiate between malignant and healthy tissues in breast cancer (*9*). Furthermore, a previous publication by Phelps *et al* (2018) determined that the m/z 700.582 and 671.466 were detected as statistically significant variables (FDR adjusted p-value < 0.01) in normal tissue to discriminate between cancer and healthy tissues in ovarian cancers (*14*). The study by Tzafetas *et al* (2020) reported that the m/z 794.569 was also detected as statistically significant molecular features (FDR adjusted p-value < 0.01) for the classification of cancer and normal tissues in cervical cancers (*16*). In conclusion, our findings of a small subset of informative molecular features could be potential diagnostic markers that can distinguish differences between normal and malignant tissue over a range of cancer types.

Since this data included over 10,000 tissue samples many of which had been mislabelled as normal or cancer tissues, this research needs more histological annotation samples to balance the tissue groups and make it less biased with breast cancer samples. However, the disorder of lipid metabolism plays a critical role in carcinogenesis and cancer development because they generally cause abnormal expression of a large number of genes, proteins and also signaling pathway (*35*). In order to study the related lipidomic and metabolomic pathways in cancers, further avenues of this research are required to identify lipid metabolite for the eight informative features depending on their mass spectrum and exact mass with the use of LIPID Metabolite and Pathways Strategy (LIPID MAPS) Lipidomic Gateway and METLIN fragmentation (*16*). After the lipid metabolites are identified, their related biological pathway in cancers will be detected using metabolic network analysis (*36*). Finally, the small group of molecular features will be examined their diagnostic power using a triple quadrupole instrument, which can enhance the sensitivity for the performance of the tumour classification. The results then will need to be compared to previously published results regarding the processing time and diagnostic accuracy for a range of tissues.

CODE AVAILABILITY

The code for data analysis in this study are available at https://github.com/Teerapon789/Machine_Learning_of_iKnife_Data_Project2

ACKNOWLEDGEMENTS

I would like to thank the head of intelligent knife (iKnife) and mass spectrometry laboratory, Professor Zoltan Takats, for the precious opportunity to conduct research under his guidance. We also wish to express our thanks to all of the laboratory members for their helpful feedback, which has been invaluable to developing and expanding our knowledge. A special thanks is provided to James McKenzie as a day-to-day supervisor and Yuchen Xiang for all their guidance and supports during the project.

REFERENCES:

1. P. E. Marik, G. P. Zaloga, Immunonutrition in high-risk surgical patients: a systematic review and analysis of the literature. *JPEN J Parenter Enteral Nutr* **34**, 378-386 (2010).
2. J. Balog *et al.*, Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci Transl Med* **5**, 194ra193 (2013).
3. R. Jeevan *et al.*, Reoperation rates after breast conserving surgery for breast cancer among women in England: retrospective study of hospital episode statistics. *BMJ* **345**, e4505 (2012).
4. A. Westgaard *et al.*, Resectable adenocarcinomas in the pancreatic head: the retroperitoneal resection margin is an independent prognostic factor. *BMC Cancer* **8**, 5 (2008).
5. A. Wibe *et al.*, Prognostic significance of the circumferential resection margin following total mesorectal excision for rectal cancer. *Br J Surg* **89**, 327-334 (2002).
6. O. Yossepowitch *et al.*, Positive surgical margins in radical prostatectomy: outlining the problem and its long-term consequences. *Eur Urol* **55**, 87-99 (2009).
7. P. Inglese, G. Correia, P. Pruski, R. C. Glen, Z. Takats, Colocalization Features for Classification of Tumors Using Desorption Electrospray Ionization Mass Spectrometry Imaging. *Anal Chem* **91**, 6530-6540 (2019).
8. S. Eeckhaoudt, L. Van Vaeck, R. Gijbels, R. E. Van Grieken, Laser microprobe mass spectrometry in biology and biomedicine. *Scanning Microsc Suppl* **8**, 335-358 (1994).
9. E. R. St John *et al.*, Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. *Breast Cancer Res* **19**, 59 (2017).
10. K. C. Schafer *et al.*, In vivo, in situ tissue analysis using rapid evaporative ionization mass spectrometry. *Angew Chem Int Ed Engl* **48**, 8240-8242 (2009).
11. A. Brunelle, O. Laprevote, Lipid imaging with cluster time-of-flight secondary ion mass

spectrometry. *Anal Bioanal Chem* **393**, 31-35 (2009).
12. J. Balog *et al.*, Identification of biological tissues by rapid evaporative ionization mass spectrometry. *Anal Chem* **82**, 7343-7350 (2010).
13. K. C. Sachfer *et al.*, In situ, real-time identification of biological tissues by ultraviolet and infrared laser desorption ionization mass spectrometry. *Anal Chem* **83**, 1632-1640 (2011).
14. D. L. Phelps *et al.*, The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). *Br J Cancer* **118**, 1349-1358 (2018).
15. S. J. Cameron *et al.*, Rapid Evaporative Ionisation Mass Spectrometry (REIMS) Provides Accurate Direct from Culture Species Identification within the Genus Candida. *Sci Rep* **6**, 36788 (2016).
16. M. Tzafetas *et al.*, The intelligent knife (iKnife) and its intraoperative diagnostic advantage for the treatment of cervical disease. *Proc Natl Acad Sci U S A* **117**, 7338-7346 (2020).
17. L. Hanel, M. Kwiatkowski, L. Heikaus, H. Schluter, Mass spectrometry-based intraoperative tumor diagnostics. *Future Sci OA* **5**, FSO373 (2019).
18. N. Goossens, S. Nakagawa, X. Sun, Y. Hoshida, Cancer biomarker discovery and validation. *Transl Cancer Res* **4**, 256-269 (2015).
19. M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, S. K. Dhillon, Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak* **19**, 48 (2019).
20. B. Li *et al.*, Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis. *Sci Rep* **6**, 38881 (2016).
21. F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **78**, 4281-4290 (2006).
22. W. E. Wolski *et al.*, Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process. *BMC Bioinformatics* **6**, 285 (2005).
23. M. Pooladi *et al.*, Cluster and Principal Component Analysis of Human Glioblastoma Multiforme (GBM) Tumor Proteome. *Iran J Cancer Prev* **7**, 87-95 (2014).
24. Z. Zhang, A. Castello, Principal components analysis in clinical studies. *Ann Transl Med* **5**, 351 (2017).
25. I. Tsamardinos, E. Greasidou, G. Borboudakis, Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach Learn* **107**, 1895-1922 (2018).
26. Z. M. Hira, D. F. Gillies, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinformatics* **2015**, 198363 (2015).
27. L. Goh, N. Kasabov, An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data. *J Bioinform Comput Biol* **3**, 1107-1136 (2005).
28. N. S. Escanilla *et al.*, Recursive Feature Elimination by Sensitivity Testing. *Proc Int Conf Mach Learn Appl* **2018**, 40-47 (2018).
29. X. Lin *et al.*, A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J Chromatogr B Analyt Technol Biomed Life Sci* **910**, 149-155 (2012).
30. C. M. Florkowski, Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* **29 Suppl 1**, S83-87 (2008).
31. F. S. Nahm, Nonparametric statistical tests for the continuous data: the basic concept and the practical use. *Korean J Anesthesiol* **69**, 8-14 (2016).
32. S. Y. Chen, Z. Feng, X. Yi, A general introduction to adjustment for multiple comparisons. *J Thorac Dis* **9**, 1725-1729 (2017).
33. M. A. Little *et al.*, Using and understanding cross-validation strategies. Perspectives on Saeb et al. *Gigascience* **6**, 1-6 (2017).
34. X. Lin *et al.*, Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. *Molecules* **23**, (2017).
35. J. Long *et al.*, Lipid metabolism and carcinogenesis, cancer development. *Am J Cancer Res* **8**, 778-791 (2018).
36. V. Palombo *et al.*, PANEV: an R package for a pathway-based network visualization. *BMC Bioinformatics* **21**, 46 (2020).

ABBREVIATIONS

iKnife, Intelligent knife; PQN, Probabilistic Quotient Normalisation; m/z, Mass-to-charge ratios; RFE, Recursive feature elimination; FDR, False discovery rate; PCA, Principal component analysis; PC1, First principal component; PC2, Second principal component; PC3, Third principal component; PC loading, Principal component loading; LR, Logistic Regression; LDA, Linear Discriminant Analysis; KNN, K-Neighbours Classifier; CART, Decision Tree Classifier; SVM, Support Vector Machine; NB, Gaussian Naïve Bayes; XGBoost, Extreme Gradient Boosting; ROC, Receiver operating characteristic; LIPID MAPS, LIPID Metabolite and Pathways Strategy.