Fast Characteristic of Skin Lesions by Machine-Learning of Raman Spectrum

Hua Zhang¹, Danhua Wang¹, Limei Qu¹, Ying Xue³, Xinli Li³, Bei Li²,³,*, Bin Liu¹,*

¹The First Hospital of Jilin University, Changchun 130061, China
²The State Key Lab of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, CAS, 130033 Changchun, China
³HOOKE Instruments Ltd., 130033 Changchun, China

Correspondence

Prof. Bin Liu, The First Hospital of Jilin University, 71 Xinmin St, Changchun, Jilin, China. Tel: +8613756661290; E-mail: l_bin@jlu.edu.cn;

Prof. Bei Li, The State Key Lab of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, CAS, 130033 Changchun, Jilin, China. Tel: +8618343116622; E-mail: beili@ciomp.ac.cn.

Abstract

Background: The traditional diagnosis of skin lesions mainly relies on dermoscope and pathological biopsy, of which the former is non-objective and the latter is invasive and time-consuming. It is necessary to find an objective and non-invasive inspection method for the diagnosis of skin cancer which is the most common malignant tumor. Herein, we aimed to fast identify the skin cancers on ultrathin frozen fresh tissue sections by combining Raman spectroscopy detection and machine learning technology.

Methods and material: 22 fresh frozen tissue sections including 3 squamous cell carcinomas, 11...
basal cell carcinomas, 2 malignant melanomas, 3 seborrheic keratosis, and 3 melanocytic nevi, were included and performed Raman detection. To prevent the discrete Raman data distribution affecting the generalization ability of the learning model, a series of adaptive preprocessing algorithms were first applied to standardize the raw Raman data of five skin lesions. The processed Raman data were performed visualized cluster analysis by principal components analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). And, using K-nearest Neighbor (KNN) and support vector machine (SVM) classifiers, two predictive models for diagnose were established and evaluated in the training set and test set by the confusion matrixes and receiver operating characteristic (ROC) curves.

**Results:** The mean variance Raman spectrum graph of 5 skin lesion types were acquired after standardization procession and 4 peak positions with large differences were found. Through dimensionality reduction by PCA and t-SNE, the visual clustering results of Raman data showed heterogeneous intra-cluster homogeneity and inter-cluster dispersion. The test accuracies reached 94.56% and 98.94% in KNN and SVM classifiers respectively. The areas under the ROCs of the two classifiers, in the category dimension and the sample dimension, were all more than 0.99 which is close to the perfect classification effect.

**Conclusions:** Raman spectroscopy is a competitive candidate for the fast and accurate diagnosis of skin lesions and the molecular information provided may be used in the pathological classification, predicting immunotherapy responsiveness and stratifying prognostic risk. Furthermore, the combination of Raman spectroscopy and machine learning methods showed great diagnostic capabilities with high accuracy is a promising tool for the diagnosis of skin lesions.

**Keywords:** skin tumor, Raman spectroscopy, machine learning, molecular diagnosis
Introduction

Skin cancers, including basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma (MM), are the most common malignancies worldwide. [1] Among them, BCC and SCC collectively named nonmelanoma skin cancer (NMSC), account for about 40% of all malignant tumors. [1, 2] According to the World Health Organization (WHO) in 2014, 2 ~ 3 million and 13 kilo people suffered from NMSC and MM worldwide each year respectively. [3] In the United States, approximately 3 million patients of NMSC are expected to be treated each year while ~10 thousands new cases and almost 7 thousands death arise. [4, 5] Moreover, ~ 40% of patients will relapse within 2 years. [6]

The diagnosis of skin lesions mainly relies on dermoscope and pathological biopsy. Dermoscope is a non-invasive in situ diagnostic tool based on visual and morphological recognition. However, it is non-objective and highly dependents on the experience of doctors. Many studies showed the accuracy of melanoma diagnosis by dermatologists varies between 56% and 82.8%, while up to one third of melanomas were misdiagnosed as benign lesions. [7, 8] Although pathological biopsy is the gold standard for diagnosis, its invasion and time-consuming burden on patients and doctors, and increases a number of unnecessary biopsies. The number needed to treatment for the resection of one malignant skin lesion was reported 20 ~ 59. [9, 10] Therefore, it is necessary to find a non-invasive, objective and high-efficient screening and diagnosis method.

Raman spectroscopy (RS) has been widely used for molecular detection due to its ultra-fine capability of acquiring chemical composition, structure and spatial information. In 1997, Gniadecka et al. observed the changes in lipids (1301 - 1321 cm\(^{-1}\)) and proteins (2939 - 2948 cm\(^{-1}\)) of skin lesions through RS. [11] As more attention were attracted by Raman in medicine [12, 13], RS has
been widely used in tumor identification and efficiently improved the sensitivity and specificity of
tumor diagnosis.[14-16] However, there are still some challenges for RS for skin lesions. Firstly,
paraffin embedding, formalin fixation as well as dewaxing were widely used in tissue
preparation.[17-19] They were usually accompanied with changes of tissue composition and loss of
tissue fluid, resulting in amounts of untrue and incomplete information acquired.[20, 21] Therefore,
"zero processing" tissue specimens may be the best choice for Raman. Secondly, the Raman spectra
on tissue contains complex and huge information reflecting the intricate lesion progression and
microenvironment. [22, 23] The result is, too many peaks arise and become stupendous to
distinguish them and identify the underlying skin cancers.

Artificial intelligence (AI) based on machine learning (ML) has attracted huge attention in recent
years in medical research. A method of non-negative matrix factorization was used and revealed the
diversity of tumor mutation process.[24] A distance-based classifier with a 90%-above recognition
accuracy was established and used to develop more personalized cancer immunotherapy.[25]
Moreover, the accuracies of predicting cancer susceptibility, recurrence and mortality were
increased by 15% - 25% by ML algorithm.[26] Powerful data processing capabilities of AI may be
able to provide good assistance for RS diagnosis.

Here in this paper, we aimed to fast identify skin lesions using 22 ultrathin frozen fresh tissue
sections by combining RS detection and ML technology (Fig. 1). To prevent the affections of
discrete Raman data on the generalization of the learning models, a series of adaptive preprocessing
algorithms were first applied to standardize the Raman data before performing both supervised and
unsupervised identification. Linear principal components analysis (PCA) and non-linear t-
distributed stochastic neighbor embedding (t-SNE) dimensionality reduction were used to visualize
clustering and both showed heterogeneous intra-cluster homogeneity and inter-cluster dispersion. K-nearest Neighbor (KNN) and support vector machine (SVM) classifiers were used in category classification. The mean areas under the receiver operating characteristic (ROC) curves of KNN and SVM in the category and sample dimensions were all greater than 0.99, which is close to the perfect effect. Besides, the significant differences of tryptophan, phenylalanine and proline in 5 skin lesion types indicated by RS may be used in the pathological classification, predicting immunotherapy responsiveness and stratifying prognostic risk.

Methods

Sample preparation

During the period from December 2018 to September 2019, patients with skin lesions who passed the routine examinations and were recommended for surgical resections by more than two specialists, entered our screening process. According to our inclusion criteria (16 - 65 years old; no underlying diseases; no history of tumors; no history of drug use, smoking and drinking in the last month; non-pregnant status), a total of 40 patients were included. After surgical removal of the lesion, we dripped the embedding agent (OCT embedding agent, Servicebio, Cat: G6059) on the lesion tissue (25 × 25 × 2 mm) which was placed on the supporter. Then the tissue together with the supporter were put on the freezing table (Leica CM1950 incubator cryomicrotome) and clamped by the microtome holder. Two adjacent sections with thickness of 5 μm were obtained with one for Hematoxylin-eosin (HE) staining. While the other one was kept at 4 °C for Raman inspection on the quartz glass slide (Delmon Technology; model 72 mm × 25 mm × 1 mm, manufacturer Wuxi Zejia Analytical Instrument Co., Ltd). All the section processes were
performed by professional technicians from Pathology Department of the First Hospital of Jilin University.

**HE staining and pathological diagnosis**

The frozen sections attached to the glass slides were fixed with a mixture of 95% ethanol and 5 ml glacial acetic acid for 1 min and washed with tap water. Then we stained sections with hematoxylin for 1 - 2 min, and washed with tap water and dilute ammonia for 20 s and 30 s respectively. After washing for another 20 s with tap water, sections were eosin-stained for 20 s - 1 min, dehydrated and sealed with neutral gum. All the works were completed by professional technicians from Pathology Department of the First Hospital of Jilin University. The prepared HE sections were handed over to two or more pathologists for diagnosis.

**Raman spectroscopy detection**

To ensure homogenization, we selected 22 sections with a single pathological result, including 3 SCC cases, 11 BCC cases, 2 MM cases, 3 SK cases and 3 MN cases. After anchoring lesion areas of the HE sections, the corresponding quartz slides were placed on the stage of the Raman spectrometer (HOOKE D100, HOOKE Instruments Ltd., China) for detection and spectrum collection. The laser wavelength for Raman detection is 785 nm. The single spectrum acquisition condition is 25 mw for 10 s. 50 spectra for each sample and 1100 spectra in total were collected respectively.

**Raman spectrum processing**

To decrease the noise and interference, it is necessary to train the data before identifying skin lesions by Raman spectra. In order to reduce impact of the data scale and distribution on the learning model, we standardized the spectral data and map it to the same dimension. Then, preprocessing operations such as smoothing and baseline correction were performed. Additionally, to smooth the curve, the Savitzky-Golay (SG) filtering algorithm was applied to fit the low-frequency component (the signal
part) and remove the high-frequency component (the noise part) from the Raman spectrum. Finally, we used adaptive iteratively reweighted penalized least squares algorithm (airPLS) baseline correction algorithm to remove the background introduced by fluorescence, chip and the tissue slide itself. As an example, the Raman spectrum standardization results of BCC were shown in Figure 2, as well as the preprocessing schemes and corresponding algorithms in Table 1.

**Algorithms analysis**

**Unsupervised learning**

The Raman spectra of skin lesions has 701 dimensions, of which many are noise and redundant information that have no contribution to classifications. To compress spectral dimensions and reduce overfits, a feature dimension reduction method is required. Without knowing data feature contribution, unsupervised learning used in data compression can improve the usability of the algorithms and their performance in high dimensions, and is helpful for the visualization of the data.

1. **PCA (principal component analysis)** is a conversion technique used in unsupervised linear data, which is the most widely used data compression algorithm[27, 28]. It carries out orthogonal transformation (a kind of linear transformation, in which the inner product of two vector spaces remains unchanged during transformation) according to the data characteristics to eliminate the correlation between each component of the original vectors. The corresponding eigenvectors with decreasing eigenvalues are obtained by transformation. After orthogonal transformation, the high-dimensional space Raman spectrum can be expressed as a low-dimensional space.

2. **t-SNE (t-distributed stochastic neighbor embedding) algorithm** is a non-linear dimension reduction method. First, it converts the Euclidean distance between two high-dimensional space data points into similarity probability. Then, the joint probability of the high-dimensional space data point and the corresponding low-dimensional space analog data point is used to replace the conditional probability in the random neighborhood embedding algorithm. t-SNE makes the shorter
distance data points in high dimensional space have larger distances after mapping, so that the points in the same cluster were gathered more closely, and the points in different clusters farther apart, which effectively solves the data crowding problem in the low-dimension space.

**Supervised learning**

Compared with unsupervised learning, the training data had eigenvalues and label values. Through the study of training data, the learning model independently established the connection between eigenvalues and label values, and predicted label values based on data features. Following are the methods we used, KNN (K-nearest Neighbor) and SVM (support vector machine).

(1) KNN, also known as the nearest neighbor algorithm, is based on an analogous learning method by comparing a given test tuple with its similar training tuple. For each new data, the closest K data will be found in the given data tuple, then the K data and the new data will be initially set to the same category. In this paper, Euclidean distance was adopted. Suppose the Euclidean distance of two points or tuple sum $(X_1 = (x_{11}, x_{12}, \ldots, x_{1n}), X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$) is:

$$
\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}
$$

(1)

(2) SVM is a method of classifying linear and nonlinear data and has been widely used in many clinical predictions[29, 30]. It maps the training data to higher new dimensions and searches for the best classification plane (decision boundary) which can separate the data into different classes. Two data tuples can always be separated by the decision boundary as higher as the non-linear mapping dimension is. For RS data with a lot of eigenvectors, the calculation of the inner product in the high-dimensional space is too large to solve and remains the core of SVM function. In this article, SVM kernel function selects polynomial kernel: the kernel function of samples $x_i$ and $x_j$ is $\kappa(x_i, x_j)$, $d$ is the degree of the polynomial, when $d = 1$, polynomial kernel degenerates to linear kernel:

$$
\kappa(x_i, x_j) = (x_i^T x_j)^d
$$

(2)
According to our inclusion criteria and single pathology principle, another three lesion tissues (Table 5, 6) were obtained for LC-MS. After getting the same amounts of tissues from the central of lesions, chromatographic grade methanol (mass volume ratio 1 g : 2.5 mL) were immediately added in and vortexed for 1 min. Then tissues were homogenized for 3 min with 2 ~ 3 zirconium dioxide grinding beads. After grinding 3 min, the homogenates were centrifuged at 14,000 rpm for 10 min at 4 °C, and the upper aqueous layers were used in LC-MS analysis. The standards of phenylalanine and tryptophan were dissolved in pure methanol to obtain 2.00 mg/mL stock solutions. Before sample detection and analysis, the stock solutions were diluted with pure methanol and made into mixed standards (2000 ng/mL).

The high-resolution mass spectrometry (MS) (Q Exactive™, Thermo Fisher Scientific (China) Co., Ltd.) coupled with electrospray ionization (ESI) was performed in the positive and negative ion switching scan mode. Parallel reaction monitoring (PRM) was selected in the detection. The resolution of the equipment is 17500 and scan range 50.0~500.0 m/z. During detection, the spray voltage was set at 3.2 kV in positive ionization mode, capillary temperature at 300 °C and nitrogen at 40 Arb. Data collection time was 8.00 min. Analyte information were shown in Table 2. The Liquid chromatography (LC) (UltiMate 3000 RS, Thermo Fisher Scientific (China) Co., Ltd.) used T3 column (2.1 × 150 mm 3 μm, waters) with a flow rate of 0.30 ml/min which was maintained at 35 °C. The aqueous phase was 10 mM ammonium formate solution at PH 3.0 adjusted by formic acid and the organic phase was acetonitrile. The elution gradients were showed in Table 3. The injection volume was 5 μL for each sample.

The chromatogram acquisition and integration were processed by the software Xcalibur 3.0 (Thermo Fisher) and linear regression with 1/X² as weighting coefficient was performed to get the standard curves of phenylalanine and tryptophan (Table 5, 6).

Results

Raman characteristics and molecular information of five skin lesion types
In order to avoid the uneven distribution of categories, 50 Raman spectra were collected for each lesion sample and a total of 1100 Raman spectra were acquired. The 1100 spectra were sequentially standardized and batched. After processing, the data was trained and tested for the identification model. The following Figure 3 showed the mean variance map of the preprocessed Raman spectra of 5 skin lesions, where the solid lines were the mean spectra, and the shaded bars represented the standard deviations within groups.

From the mean variance graph of Raman spectra of 5 skin lesions, 4 peak positions (720 cm\(^{-1}\), 752 cm\(^{-1}\), 853 cm\(^{-1}\), 1002 cm\(^{-1}\)) with significant differences were noticed. Their physical origins and peak intensity disparities were summarized in Table 4 and Figure 4. The peak intensities and spans of SCC, MM and BCC at 720 cm\(^{-1}\) (nucleic acid band)[31] (Table 4) were higher than those of SK and MN (Fig. 4A). As nucleic acid is positively correlated with tumor malignancy[32], this result preliminarily verified the ability of RS in tumor detection. The peak intensities of 752 cm\(^{-1}\) (the symmetrical respiration of tryptophan)[33] and 1002 cm\(^{-1}\) (the vibration mode of the ring breathing caused by phenylalanine)[33, 34] (Table 4) in SCC and MM were significantly higher than other lesions (Fig. 4B, D). In order to verify the reliability of our test results, we used LC-MS to detect the contents of phenylalanine and tryptophan in SCC, SK and MN. The contents of phenylalanine (Table 5) and tryptophan (Table 6) in SCC were indeed higher than SK and MN. In addition, the peak intensity of 853 cm\(^{-1}\) (the stretching of collagen proline ring (C-C))[35] (Table 4) in MN was much higher than other lesions, followed by SK (Fig. 4C), indicating that proline decreases as the malignancy of skin lesions increases.

**Visualized clustering results in t-SNE and PCA**

Using the standardized Raman data of 5 skin lesions, two ways were carried out for cluster analysis. Figure 5 showed that nonlinear t-SNE dimensionality reduction and linear PCA were used to visualize the clustering results. Figure 5A showed t-SNE dimensionality reduction results which using two largest contribution dimensions of t-SNE 1 and t-SNE 2 achieved highly nonlinear distinguishable for these 5 types of skin lesions. Figure 5B showed the result of PCA three-dimensional visualization. The three largest principal components of PC1, PC2 and PC3 spectra
were used to achieve linear separable of 5 types of skin lesions. The above two unsupervised learning methods both showed heterogeneous intra-cluster homogeneity and inter-cluster dispersion.

**Confusion matrixes and validations of SVM and KNN models**

Next, RS data of five skin lesions were learned and analyzed by two common splitters of SVM and KNN in supervised learning. 20% RS data were tested and confusion matrixes of the recognition results were showed in Figure 6 (A, B). In the confusion matrixes, the horizontal direction represented the true category label, the vertical direction labeled the represented category label (the category label corresponding to the highest predicted probability), and the diagonal value indicated the recognition accuracy of the corresponding category test data. Calculating the mean value of the diagonal lines of the confusion matrixes, KNN and SVM test accuracies were 94.56% and 98.94% respectively. In KNN, 11.1% of SCCs were misjudged as BCCs, 5.6% of SKs were confused with MMs and 10.5% of MNs were misdiagnosed as SCCs (Fig. 6A). In SVM, 5.3% of MNs were misjudged as SKs (Fig. 6B).

With false positive rate (FPR) as the horizontal axis and true positive rate (TPR) as the vertical axis, ROC curves of five skin lesion categories were drawn in KNN and SVM, and area under curve (AUC) was used to measure the excellence of the prediction models (Fig. 6C, D). Macro-average ROC curves were drawn using the mean value of ROC curves of 5 categories indicating category dimension prediction. Micro-average ROC curves were drawn using the mean value of ROC curves of all test samples indicating sample dimension prediction.

After calculating, the AUCs of the five categories in KNN classifier were all greater than 0.97, and the mean AUCs in the category dimension and the sample dimension were both 0.99 (Fig.
The AUCs of all test samples and categories both were 1 in the SVM classifier (Fig. 6D). These data indicated that KNN and SVM were all close to perfect classifiers.

Discussions

Since the specimens in pathology department were all preserved with formalin fixation and paraffin embedding, many human tissue specimens were tested directly within the wax blocks[17, 18] and some underwent gentle dewaxing treatment[19] or digital dewaxing of RS signal[46]. Although some studies demonstrated that the detection of wax or formalin-fix blocks have no effects on the Raman spectra of specimens and paraffin tissues can be almost completely dewaxed[21, 47, 48], some studies revealed that the amount of paraffin remained varied in different tissues[21], and formalin fixation affected Raman information of tissues[20]. Huang Z et al compared the Raman spectra of fresh human bronchial tissues with formalin-fixed tissues and found that formalin fixation has a significant effect on the near-infrared Raman spectra of tissues and the diagnostic markers from the 980-1100 and 1500-1650 cm\(^{-1}\) regions of fixed tissues do not seem to be suitable for in vivo lung cancer detection[20]. To provide accurate Raman information for in vivo applications, it may be better to use fresh tissue specimens. Some experiments use frozen or fresh tissue blocks.[49, 50] However, although they were free from the interference of various treatments in vitro, the influences of residual substances on the surface of the skin cannot be avoided. Meanwhile, the detection site cannot be accurately guaranteed to be the same pathological type as the HE staining indicates because of the block thickness. The thickness of the fresh frozen tissue section we used is 5 \(\mu m\), about 1/2 of cell diameter, and it is adjacent to the HE section, which maybe avoid the above problems. Moreover, all of our tests were completed within 1 hour with a layer of water film dripped on the surface of the section during detection to minimize the qualitative change of the tissue after
resection. Although our RS detections were performed in vitro, the 785 nm laser we use can detect deep tissues at 200 μm, laying the foundation for the non-invasive detection of RS on human body. Concerning the vivo testing, different pathological types of the same lesion should be further detected in next study.

So far, most of the Raman identifications for skin lesions are mainly concentrated between the lesion and the normal tissue,[51-53] the tumor and the adjacent tissue,[54] or mainly applied in the rapid determination of the resection margin during the operation.[55-57] In our study, five types of lesions were analyzed simultaneously, including benign, malignancy, MM and NMSC. In order to get better classification effects, unsupervised learning (PCA, t-SNE) and supervised learning (KNN, SVM) were used in cluster and classification analysis of 5 types of skin lesions. PCA and t-SNE both show better clustering based on categories (Fig. 5). With its lower coupling feature, t-SNE showed its superiority in high-dimensional Raman spectral data dimensionality reduction and visualization (Fig. 5A). Two classifiers of KNN and SVM showed high test accuracies in the RS identification of 5 skin lesion types. In KNN, 11.1% of SCC was misjudged as BCC and 5.6% of SK was confused with MM (Fig. 6A). For SCC, BCC, SK and MM, further surgical resections and pathological biopsies are necessary, so the above misjudgments are acceptable in the clinic. 10.5% of MNs were misjudged as SCC (Fig. 6A), which may increase unnecessary biopsies. Compared to the lower-skilled physicians, KNN has more experience in diagnosis deserving the title of "senior physician". SVM, with only 5.3% of MNs misjudged as SK (Fig. 6B), did show an almost ideal classification effect. Moreover, in KNN AUCs of ROC curves in the category dimension and the sample dimension were both 0.99, and in SVM were both 1 (Fig. 6C, D), indicating that both classifiers are perfect classifiers of RS in the identification of 5 skin lesion types. All in all, the
application of machine learning in RS can better identify skin lesion types and provides a better bridging method for the application of RS in AI diagnostics.

How to achieve individualized treatment is still one of the ten challenges facing tumor immunotherapy. The essential amino acid tryptophan catabolism is recognized as an important microenvironmental factor that suppresses antitumor immune responses in cancer and regulates T cell proliferation, activation and anti-tumor effects. Phenylalanine is involved in regulating cell cycle progression, modulating invasion-related signaling/function proteins, and promoting tumor cell adhesion and spread. The lack of phenylalanine can induce focal adhesion kinase-dependent apoptosis and mitochondria-initiated apoptosis. Interestingly, the different contents of tryptophan and phenylalanine were detected by RS in the 5 skin lesion types we studied and were highly consistent with LC-MS results (Fig. 4B, D, Table 5, 6). In addition, as the main component of collagen, the decrease of proline indicates tumor metastasis and poor prognosis. Similarly, lower strengthen of proline signal was found in malignant lesions as compared to benign lesions in our study (Fig. 4C). These results indicate that RS provides reliable molecular information related with tumor therapy and progression. Furthermore, AI-aided RS may be a reliable screening method for immunotherapy responsiveness and individualized therapy. These conjectures will be examined in our following research.

Conclusion

In summary, RS is a competitive candidate for the fast and accurate diagnosis of skin lesions with ultrathin frozen fresh sections providing high-quality Raman spectra. And, the application of machine learning methods in Raman spectrum classification showed excellent diagnostic capabilities for 5 skin lesion types. KNN and SVM predictive models diagnosed 5 skin lesion types.
with almost perfect accuracy. The significant differences of tryptophan, phenylalanine and proline indicated by RS may imply different progression and treatment responsiveness of 5 skin lesion types. These results identify that ML-aided RS is a potential tool in clinic diagnosis and screening of tumor immunotherapy, progression and prognosis.

**Abbreviations**

BCC: Basal cell carcinoma; SCC: Squamous cell carcinoma; MM: Melanoma; NMSC: Nonmelanoma skin cancer; WHO: World Health Organization; RS: Raman spectroscopy; AI: Artificial intelligence; ML: Machine learning; PCA: Principal components analysis; t-SNE: t-distributed stochastic neighbor embedding; KNN: K-nearest Neighbor; SVM: Support vector machine; ROC: Receiver operating characteristic; HE: Hematoxylin-eosin; LC-MS: Liquid chromatography-mass spectrometry; MS: Mass spectrometry; ESI: Electrospray ionization; PRM: Parallel reaction monitoring; FPR: False positive rate; TPR: True positive rate; AUC: Area under curve.

**Acknowledgements**

Not applicable.

**Authors’ contributions**

H. Zhang, Y. Xue, B. Li and B. Liu conceived the study. H. Zhang and D. Wang collected samples. D. Wang and L. Qu performed tissue section preparation and HE staining. H. Zhang and Y. Xue performed Raman detections. H. Zhang, Y. Xue and X. Li analyzed the data and established the predictive models. H. Zhang, Y. Xue and X. Li wrote the paper, and B. Li and B. Liu revised the manuscript. All authors read and approved the final manuscript.

**Funding**
Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

The human samples were approved by the institutional review board at First Hospital of Jilin University. All the experiments were performed in accordance with relevant guidelines and regulations. Skin lesion tissues were obtained from patients recommended for surgical resections at First Hospital of Jilin University after informed consents were obtained from the patients. All subjects were provided written informed consents in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no potential conflict of interest.

References

1. Cakir BO, Adamson P, Cingi C. Epidemiology and economic burden of nonmelanoma skin cancer. Facial Plast Surg Clin North Am. 2012; 20: 419-422.
2. Brandt MG, Moore CC. Nonmelanoma Skin Cancer. Facial Plast Surg Clin North Am. 2019; 27: 1-13.
3. Radiation: Ultraviolet (UV) radiation and skin cancer [http://www.who.int/uv/faq/skincancer/en/index1.html]
4. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. JAMA Dermatol. 2015; 151: 1081-1086.
5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020; 70: 7-30.
6. Raasch BA, Buettner PG. Multiple nonmelanoma skin cancer in an exposed Australian population. International journal of Dermatology. 2002; 41: 652-658.
7. Morton C, Mackie R. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. The British journal of dermatology. 1998; 138: 283-287.
8. Elmore J, Barnhill R, Elder D, Longton G, Pepe M, Reisch L, Carney P, Titus L, Nelson H, Onega T et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ (Clinical research ed). 2017; 357: j2813.
9. Redondo P, Gimenez de Azcarate A, Núñez-Córdoba JM, Andreu EJ, Garcia-Guzman M, Aguado L, Prosper F. Efficacy of Autologous Melanocyte Transplantation on Amniotic Membrane in Patients With Stable Leukoderma. JAMA Dermatology. 2015; 119: 1339-1346.
10. English DR, Del Mar C, Burton RC. Factors influencing the number needed to excise: excision rates of pigmented lesions by general practitioners. Medical journal of Australia. 2004; 180: 16-19.
11. Gniadecka M, Wulf HC, Nielsen OF, Christensen DH, Hercogova J. Distinctive molecular abnormalities in benign and malignant skin lesions: studies by Raman spectroscopy. Photochemistry and photobiology. 1997; 66: 418-423.
12. Santos IP, Caspers PJ, Bakker Schut TC, van Doorn R, Noordhoek Hegt V, Koljenovic S, Puppels GJ. Raman Spectroscopic Characterization of Melanoma and Benign Melanocytic Lesions Suspected of Melanoma Using High-Wavenumber Raman Spectroscopy. Anal Chem. 2016; 88: 7683-7688.
13. Huser T, Chan J. Raman spectroscopy for physiological investigations of tissues and cells. Advanced drug delivery reviews. 2015; 89: 57-70.
14. Zhao J, Zeng H, Kalia S, Lui H. Wavenumber selection based analysis in Raman spectroscopy improves skin cancer diagnostic specificity. Analyst. 2016; 141: 1034-1043.
15. Zhang Y, Moy AJ, Feng X, Nguyen HTM, Sebastian KR, Reichenberg JS, Wilke CO, Marky MK, Tunnell JW. Assessment of Raman Spectroscopy for Reducing Unnecessary Biopsies for Melanoma Screening. Molecules. 2020; 25.
16. P. Santos I, van Doorn R, Caspers PJ, Bakker Schut TC, Barroso EM, Nijsten TEC, Noordhoek Hegt V, Koljenovic S, Puppels GJ. Improving clinical diagnosis of early-stage cutaneous melanoma based on Raman spectroscopy. British Journal of Cancer. 2018; 119: 1339-1346.
17. Grosset A, Dallaire F, Nguyen T, Birlea M, Wong J, Daoust F, Roy N, Kougiumoutzakis A, Azzi F, Aubertin K et al. Identification of intraductal carcinoma of the prostate on tissue specimens using Raman micro-spectroscopy: A diagnostic accuracy case-control study with multicohort validation. PLoS medicine. 2020; 17: e1003281.
18. Chen Y, Zheng X, Chen G, He C, Zhu W, Feng S, Xi G, Chen R, Lan F, Zeng H. Immunoassay for LMP1 in nasopharyngeal tissue based on surface-enhanced Raman scattering. International journal of nanomedicine. 2012; 7: 73-82.
19. Gajjar K, Heppenstall LD, Pang W, Ashton KM, Trevisan J, Patel, II, Llabjani V, Stringfellow HF, Martin-Hirsch PL, Dawson T et al. Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis. Anal Methods. 2012; 5: 89-102.
20. Huang Z, McWilliams A, Lam S, English J, McLean D, Lui H, Zeng H. Effect of formalin fixation on the near-infrared Raman spectroscopy of normal and cancerous human bronchial tissues. International Journal of Oncology. 2003.
21. Fullwood LM, Griffiths D, Ashton K, Dawson T, Lea RW, Davis C, Bonnier F, Byrne HJ, Baker MJ. Effect of substrate choice and tissue type on tissue preparation for spectral histopathology.
by Raman microspectroscopy. Analyst. 2014; 139: 446-454.

22. Auner GW, Koya SK, Huang C, Broadbent B, Trexler M, Auner Z, Elias A, Mehne KC, Brusatori MA. Applications of Raman spectroscopy in cancer diagnosis. Cancer Metastasis Rev. 2018; 37: 691-717.

23. Rau JV, Marini F, Fosca M, Cippitelli C, Rocchia M, Di Napoli A. Raman spectroscopy discriminates malignant follicular lymphoma from benign follicular hyperplasia and from tumour metastasis. Talanta. 2019; 194: 763-770.

24. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. Signatures of mutational processes in human cancer. Nature. 2013; 500: 415-421.

25. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017; 547: 89-93.

26. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007; 2: 59-77.

27. Nerviani A, Boutet M, Tan W, Goldmann K, Purkayastha N, Lajtos T, Hands R, Lewis M, Kelly S, Pitzalis C. IL-23 skin and joint profiling in psoriatic arthritis: novel perspectives in understanding clinical responses to IL-23 inhibitors. Annals of the rheumatic diseases. 2020.

28. Hahn V, Knutsdottir H, Luo X, Bedi K, Margulies K, Haldar S, Stolina M, Yin J, Khakoo A, Vaishnav J et al. Myocardial Gene Expression Signatures in Human Heart Failure With Preserved Ejection Fraction. Circulation. 2021; 143: 120-134.

29. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, Zhu S, Ye Z. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine. 2019; 17: 119.

30. Ma X, Wu Y, Zhang L, Yuan W, Yan L, Fan S, Lian Y, Zhu X, Gao J, Zhao J et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. Journal of translational medicine. 2020; 18: 146.

31. Puppels G, de Mul F, Otto C, Greve J, Robert-Nicoud M, Arndt-Jovin D, Jovin T. Studying single living cells and chromosomes by confocal Raman microspectroscopy. Nature. 1990; 347: 301-303.

32. Naber SP. Molecular pathology--detection of neoplasia. New England Journal of Medicine. 1994; 331: 1508-1510.

33. Jess PR, Smith DD, Mazilu M, Dholakia K, Riches AC, Herrington CS. Early detection of cervical neoplasia by Raman spectroscopy. Int J Cancer. 2007; 121: 2723-2728.

34. De Gelder J, Scheldeman P, Leus K, Heyndrickx M, Vandenabeele P, Moens L, De Vos P. Raman spectroscopic study of bacterial endospores. Analytical and bioanalytical chemistry. 2007; 389: 2143-2151.

35. Chen M, Lord R. Laser-excited Raman spectroscopy of biomolecules. VIII. Conformational study of bovine serum albumin. Journal of the American Chemical Society. 1976; 98: 990-992.

36. Munn DH, Shafizadeh E, Attwood JT, Bondarev I, Pashine A, Mellor AL. Inhibition of T cell proliferation by macrophage tryptophan catabolism. Journal of Experimental Medicine. 1999; 189: 1363-1372.

37. Curti A, Pandolfi S, Valzasina B, Aluigi M, Isidori A, Ferri E, Salvestrini V, Bonanno G, Rutella S, Durelli I et al. Modulation of tryptophan catabolism by human leukemic cells results in the
conversion of CD25- into CD25+ T regulatory cells. Blood. 2007; 109: 2871-2877.

38. Botticelli A, Mezi S, Pomati G, Cerbelli B, Cerbelli E, Roberto M, Giusti R, Cortellini A, Lionetto L, Scagnoli S et al. Tryptophan Catabolism as Immune Mechanism of Primary Resistance to Anti-PD-1. Front Immunol. 2020; 11: 1243.

39. Loayza-Puch F, Rooijers K, Buil LC, Zijlstra J, Oude Vrielink JF, Lopes R, Ugalde AP, van Breugel P, Hofland I, Wesseling J et al. Tryptophan Catabolism as Immune Mechanism of Primary Resistance to Anti-PD-1. Front Immunol. 2020; 11: 1243.

40. Elia I, Broekaert D, Christen S, Boon R, Radaelli E, Orth MF, Verfaillie C, Grunewald TGP, Fendt SM. Proline metabolism supports metastasis formation and could be inhibited to selectively target metastasizing cancer cells. Nat Commun. 2017; 8: 15267.

41. Tang L, Zeng J, Geng P, Fang C, Wang Y, Sun M, Wang C, Wang J, Yin P, Hu C et al. Global Metabolic Profiling Identifies a Pivotal Role of Proline and Hydroxyproline Metabolism in Supporting Hypoxic Response in Hepatocellular Carcinoma. Clin Cancer Res. 2018; 24: 474-485.

42. Fu YM, Yu ZX, Ferrans VJ, Meadows GG. Tyrosine and phenylalanine restriction induces G0/G1 cell cycle arrest in murine melanoma in vitro and in vivo. Nutr Cancer. 1997; 29: 104-113.

43. Fu Y-M, Meadows GG. Specific amino acid dependency regulates the cellular behavior of melanoma. The Journal of nutrition. 2007; 137: 1591S-1596S.

44. Fu Y-M, Zhang H, Ding M, Li Y-Q, Fu X, Yu Z-X, Meadows GG. Specific amino acid restriction inhibits attachment and spreading of human melanoma via modulation of the integrin/focal adhesion kinase pathway and actin cytoskeleton remodeling. Clinical & experimental metastasis. 2005; 21: 587-598.

45. Fu Y-M, Yu Z-X, Pelayo BA, Ferrans VJ, Meadows GG. Focal adhesion kinase-dependent apoptosis of melanoma induced by tyrosine and phenylalanine deficiency. Cancer research. 1999; 59: 758-765.

46. Tfaïli A, Gobinet C, Vrabie V, Huez R, Manfait M, Piot O. Digital dewaxing of Raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies. Applied spectroscopy. 2009; 63: 564-570.

47. O Faolain E. Raman Spectroscopic Evaluation of Efficacy of Current Paraffin Wax Section Dewing Agents. Journal of Histochemistry and Cytochemistry. 2005; 53: 121-129.

48. Short MA, Lui H, McLean D, Zeng H, Alajlan A, Chen XK. Changes in nuclei and peritumoral collagen within nodular basal cell carcinomas via confocal micro-Raman spectroscopy. J Biomed Opt. 2006; 11: 34004.

49. Cicchi R, Cosci A, Rossari S, Kapsokalyvas D, Baria E, Maio V, Massi D, De Giorgi V, Pimpinelli N, Saverio Pavone F. Combined fluorescence-Raman spectroscopic setup for the diagnosis of melanocytic lesions. J Biophotonics. 2014; 7: 86-95.

50. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, Hergcogova J, Rossen K, Thomsen HK, Gniadecki R et al. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. J Invest Dermatol. 2004; 122: 443-449.

51. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. Cancer Res. 2012; 72: 2491-2500.

52. Sigurdsson S, Philipsen PA, Hansen LK, Larsen J, Gniadecka M, Wulf HC. Detection of Skin
Cancer by Classification of Raman Spectra. IEEE Transactions on Biomedical Engineering. 2004; 51: 1784-1793.

Schleusener J, Gluszczynska P, Reble C, Gersonde I, Helfmann J, Fluhr JW, Lademann J, Rowert-Huber J, Patzelt A, Meinke MC. In vivo study for the discrimination of cancerous and normal skin using fibre probe-based Raman spectroscopy. Exp Dermatol. 2015; 24: 767-772.

Schut TCB, Caspers PJ, Puppels GJ, Nijssen A, Heule F, Neumann MH, Hayes DP. Discriminating basal cell carcinoma from its surrounding tissue by Raman spectroscopy. Journal of investigative dermatology. 2002; 119: 64-69.

Barroso E, Ten Hove I, Bakker Schut T, Mast H, van Lanschot C, Smits R, Caspers P, Verdijk R, Noordhoek Hegt V, Baatenburg de Jong R et al. Raman spectroscopy for assessment of bone resection margins in mandibulectomy for oral cavity squamous cell carcinoma. European journal of cancer (Oxford, England : 1990). 2018; 92: 77-87.

Matousek P, Stone N. Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring. Chemical Society reviews. 2016; 45: 1794-1802.

Haka A, Volynskaya Z, Gardecki J, Nazemi J, Lyons J, Hicks D, Fitzmaurice M, Dasari R, Crowe J, Feld M. In vivo margin assessment during partial mastectomy breast surgery using raman spectroscopy. Cancer research. 2006; 66: 3317-3322.

Hegade PS, Chen DS. Top 10 Challenges in Cancer Immunotherapy. Immunity. 2020; 52: 17-35.

**Figure 1.** Schematic of machine-learning of skin lesion Raman spectra for fast diagnosis. Ultrathin frozen fresh sections of skin lesions were acquired and prepared for Raman detection with a 785nm laser. Raman spectra were processed gradually and analyzed by machine learning methods. Diagnose information were put out at last.

**Figure 2.** BCC Raman single spectrum standardization results. RI, Raman Intensity.

**Figure 3.** The mean variance graph of Raman spectra of 5 skin lesions after preprocessing.

**Figure 4.** Normalized Raman intensity for bands at 720 cm\(^{-1}\), 752 cm\(^{-1}\), 853 cm\(^{-1}\), and 1002 cm\(^{-1}\) of 5 skin lesions. RI, Raman Intensity. Data were presented as mean ± SD. * P < 0.05; ** P < 0.01; *** P < 0.001. (Student’s t-test)

**Figure 5.** Visualized clustering results of 5 skin lesion types after dimensionality reduction in t-SNE (A) and PCA (B).

**Figure 6.** Confusion matrix and ROC curves of test results. (A, B) Confusion matrixes of 20% RS
data test results in KNN (A) and SVM (B). (C, D) The ROC curves for the recognition probabilities of 5 skin lesion categories in KNN (C) and SVM (D).