

تمایز طیفهای رامان با روشهای آماری و الگوریتم ژنتیک

## استخراج ویژگیهای بهینه برای تمایز طیفهای رامان بدست آمده از نمونه های مختلف پوستی با استفاده از روشهای آماری و الگوریتم ژنتیک

زهرة دهقانی بیدگلی<sup>۱</sup>، محمد حسین میران بیگی<sup>۲\*</sup>، رسول ملک فر<sup>۳</sup>

۱- دانشجوی دکترای مهندسی پزشکی، دانشکده برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران

۲- دانشیار گروه مهندسی پزشکی، دانشکده برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران

۳- دانشیار گروه فیزیک، دانشکده علوم پایه، دانشگاه تربیت مدرس، تهران، ایران

تاریخ پذیرش مقاله: ۹۰/۴/۲۵

تاریخ دریافت: ۸۹/۱۲/۱۰

### چکیده

**مقدمه:** طیف نگاری رامان یکی از روشهای طیف نگاری مبتنی بر پراکندگی غیرالاستیک نور تک رنگ می باشد. این روش می تواند اطلاعات مفیدی درباره فرکانس ارتعاشات پیوندهای مولکولی نمونه بدست دهد. لذا با استفاده از این روش می توان به مطالعه تغییرات مولکولی نمونه پرداخت.

**مواد و روشها:** در این تحقیق ۱۵۳ طیف رامان از نمونه های پوستی نرمال و پوست خشک شده در نتیجه شستشو با محلول شوینده سدیم لوریل سولفات به صورت درون تنی اخذ گردید. سپس پیش پردازشهایی شامل حذف خط زمینه، حذف نویز و نرمالسازی روی سیگنالها انجام شد. در ادامه با استفاده از آنالیز آماری و الگوریتم ژنتیک به جستجوی ویژگیهای بهینه برای تمایز دو گروه نرمال و خشک پرداختیم. در انتخاب ویژگی با استفاده از آنالیز آماری، آمارگان T، فاصله باتاچاریا و آنتروپی در جداسازی دو گروه محاسبه شد و با توجه به این که آمارگان T در تمایز این دو گروه کارایی بهتری از خود نشان داد، لذا از این روش برای انتخاب ویژگیهای بهینه استفاده گردید. بار دیگر با استفاده از الگوریتم ژنتیک ویژگیهای بهینه انتخاب شده و در نهایت به منظور ارزیابی کارایی هر یک از این روشها در انتخاب ویژگیهای بهینه، با استفاده از طبقه بندهای LDA، SVM، KNN و شبکه عصبی به ارزیابی دسته ویژگیهای انتخاب شده توسط هر یک از این دو روش ( آماری و الگوریتم ژنتیک) پرداختیم.

**نتایج:** در مقایسه نتایج طبقه بندی با روش های مختلف انتخاب ویژگی و نوع طبقه بندی کننده، بهترین نتایج در ترکیب انتخاب ویژگی با استفاده از الگوریتم ژنتیک و طبقه بندی با SVM حاصل شد. این ترکیب تفکیک دوکلاس را با دقت ۹۰٪، حساسیت ۸۹٪ و قطعیت ۹۱٪ امکانپذیر می سازد.

**بحث و نتیجه گیری:** از بررسی های صورت گرفته و نتایج بدست آمده می توان چنین نتیجه گرفت که الگوریتم ژنتیک کارایی بیشتری در مقایسه با روشهای آماری برای انتخاب ویژگیهای متمایز کننده طیفهای رامان از خود نشان می دهد. همچنین نتایج حاصل از این تحقیق نشان دهنده پتانسیل طیف نگاری رامان در بررسی اثر مواد مختلف روی پوست و یا بیماریهای پوستی مرتبط با خشکی پوست می باشد. (مجله فیزیک پزشکی ایران، دوره ۸، شماره ۲، پیاپی (۳۱)، تابستان ۹۰: ۳۳-۲۷)

**واژگان کلیدی:** طیف نگاری رامان، الگوریتم ژنتیک، طبقه بندی

\* نویسنده مسؤل: دکتر محمدحسین میران بیگی

آدرس: گروه مهندسی پزشکی، دانشکده برق و کامپیوتر، دانشگاه

تربیت مدرس، تهران، ایران

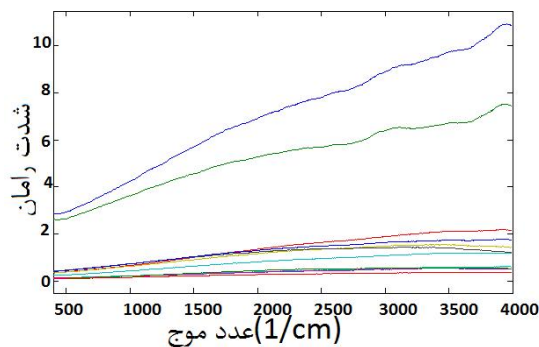
miranbmh@modares.ac.ir

تلفن: ۸۲۸۸۳۳۷۰ (۲۱) ۹۸+

## ۱- مقدمه

در این مقاله از گروه دوم تحت عنوان پوست خشک یاد می‌کنیم. از مجموع ۱۵۳ طیف ذکر شده، ۷۳ طیف، مربوط به پوست نرمال و ۸۰ طیف، مربوط به پوست خشک می‌باشد. شایان ذکر است که این آزمایشها بصورت *in vivo* روی پوست موش (رت) انجام گرفته است.

طیفها با استفاده از سیستم Thermo Nicolet Almaga شرکت Thermo Scientific اخذ شده است، لیزر این سیستم از نوع <sup>1</sup>Nd:YLF می‌باشد، که از هارمونیک دوم آن با طول موج ۵۳۲ نانومتر استفاده می‌شود. توان لیزر بکار رفته ۵۰ میلی وات و شعاع لیزر در فاصله کانونی ۱۰۰ میکرومتر می‌باشد. طیفها با ۳۲ بار اسکن در محدوده طیفی ۴۰۰-۴۰۰۰ بر سانتیمتر با رزولوشن ۴ بر سانتیمتر اخذ شده اند. هرطیف بطور جداگانه دارای ۱۰۰۰ نقطه داده است، که در محدوده عدد موجی بین ۴۰۰ تا ۴۰۰۰ بر سانتیمتر قرار دارد. هر طیف دارای دو بعد می‌باشد که در یک بعد عدد موج و در بعد دیگر شدت پراکندگی رامان در هر عدد موج قرار دارد. در شکل ۱ ده نمونه طیف رامان که بطور تصادفی از بین داده های تحت مطالعه انتخاب شده، نشان داده شده است.



شکل ۱- ده نمونه طیف رامان بدست آمده از پوست نرمال و خشک (قبل از پیش پردازش)

## ۲-۱- پیش پردازش

بیشتر شکستها و موفقیت های هر روش آنالیز داده ای به داشتن داده های مناسب در آغاز کار بستگی دارد. به همین دلیل همواره پیش پردازش داده قبل از انجام آنالیزهای بیش مورد نیاز است. خط زمینه در رامان اسپکتروسکوپی بسیار رایج است و در نتیجه فرآیند لومینسانس (فلورسانس و فسفرسانس)،

طیف نگاری رامان یکی از روشهای طیف نگاری بر اساس پراکندگی غیر الاستیک نور تک رنگ لیزر می‌باشد [۱]. هنگامیکه نور تک رنگ لیزر، به یک نمونه برخورد می‌کند، بخشی از نور پراکنده می‌شود. در نور پراکنده شده علاوه بر فرکانس نور فرودی (پراکندگی رایلی) تابشهایی در فرکانسهای متفاوت نیز مشاهده می‌گردد که پراکندگی رامان نامیده می‌شود. میزان اختلاف فرکانس نور تابشی و پراکنده شده به فرکانس نوسان پیوندهای اتمی موجود در ماده تحت مطالعه بستگی دارد. لذا می‌توان با ثبت میزان پراکندگی در هر فرکانس به ساختار مولکولی ماده تحت مطالعه پی برد، بدین ترتیب طیف رامان می‌تواند به عنوان اثر انگشت نمونه بکار گرفته شود و از این تکنیک در شناسایی مواد مختلف از جمله نمونه های بیولوژیک استفاده نمود. این ویژگی طیف سنجی رامان، آنرا به ابزاری قدرتمند در تشخیصهای بیولوژیک بدل ساخته است. بگونه ای که در چند دهه گذشته رامان اسپکتروسکوپی به عنوان یک روش زیست پزشکی و تحلیل زیستی در کاربردهای کلینیکی بکار گرفته شده است [۲-۹].

اما آنچه تاکنون مانع از توسعه این روش در کاربردهای کلینیکی بوده است، دشواری استخراج اطلاعات مفید و صحیح از طیفهای رامان نسبتا پیچیده و دارای قله ها و دره های متعدد است. از این رو در طی دو دهه گذشته، گروههای متعددی به بررسی طیفهای رامان بدست آمده از نمونه های بیولوژیک با استفاده از الگوریتمهای مختلف پردازش سیگنال پرداخته اند [۱۰-۱۵].

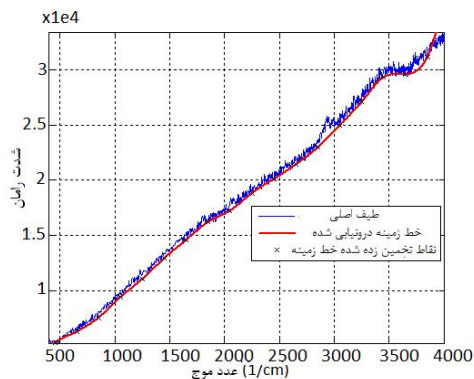
در این تحقیق سعی شده است تا ضمن اتخاذ یک روش پیش پردازش مناسب، بتوانیم به ویژگیهای بهینه در تمایز نمونه های مختلف پوستی دست یابیم. لذا در این مسیر به مقایسه روشهای آماری و الگوریتم ژنتیک در استخراج ویژگیهای بهینه پرداخته ایم.

## ۲- مواد و روشها

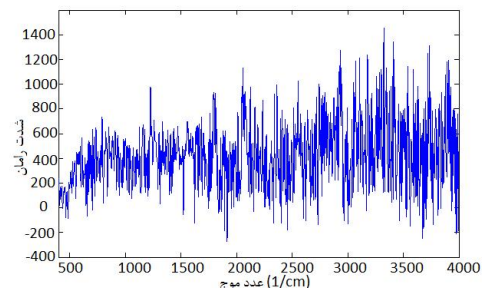
در این تحقیق ۱۵۳ طیف رامان از نمونه های پوستی نرمال و پوست شسته شده با شوینده سدیم لوریل سولفات اخذ شده است. این ماده شوینده موجب خشکی پوست می‌گردد لذا

<sup>1</sup> Neodymium doped Yttrium Lithium Fluoride

تمایز طیفهای رامان با روشهای آماری و الگوریتم ژنتیک

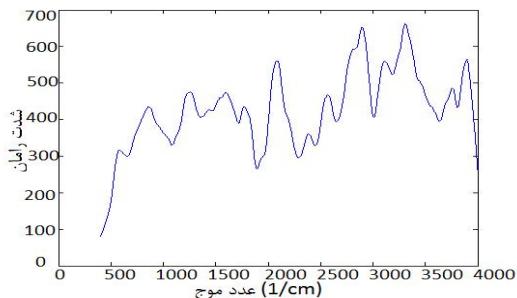


الف

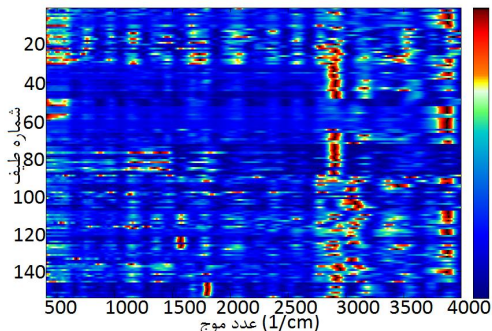


ب

شکل ۲- الف) خط زمینه تخمین زده شده برای یک طیف نمونه و  
ب) همان طیف پس از حذف خط زمینه



شکل ۳- طیف بدست آمده پس از حذف نویز



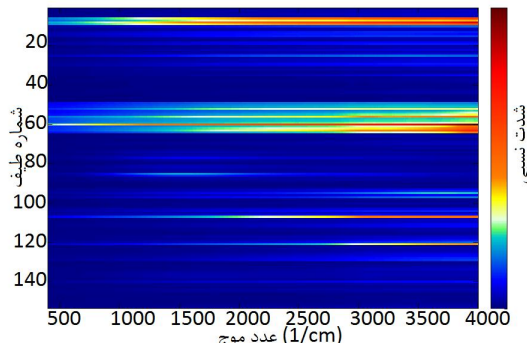
ب

فرآیند تابش غیر لیزری (نور اتاق، نور خورشید، تابش جسم سیاه، لومینسانس شیمیایی) و یا پراکندگی رامان از منابع دیگر غیر از آنالیت (بستر، حلال و ادوات نوری) می باشد [۱].

در این تحقیق برای حذف خط زمینه از پنجره های جابجا شده به طول ۱۰۰ و تقریب اسپیلاین<sup>۱</sup> استفاده شده است [۱۶]. در شکل ۲- الف یک نمونه از طیف به همراه خط زمینه تخمین زده آن و در شکل ۲- ب همان طیف پس از حذف خط زمینه در شکل ۲- ب نشان داده شده است. به منظور حذف نویز و هموارسازی طیف نیز از رگرسیون خطی و زنادار استفاده شده است [۱۷]. شکل موج بدست آمده برای همان طیف پس از حذف نویز در شکل ۳ نشان داده شده است. مجدداً خط زمینه باقی مانده به روش ذکر شده قبلی حذف شده، در نهایت نرمالسازی به ماکزیمم شدت طیف طبق رابطه زیر انجام گرفته است.

$$SN = \frac{s - \min(s)}{\max(s) - \min(s)} \quad (1)$$

در رابطه فوق SN مقدار شدت طیف نرمالیزه شده، S مقدار طیف قبل از نرمالیزاسیون و  $\min(S)$  و  $\max(S)$  بترتیب حداقل و حداکثر شدت سیگنال می باشد. تصویر حرارتی (نمایش دامنه سیگنال بصورت شدت رنگ) از تمامی طیفها قبل و بعد از پیش پردازش، بترتیب در شکلهای ۴- الف و ۴- ب نشان داده شده است. همانگونه که در شکل مشاهده می شود، پس از انجام پیش پردازشهای ذکر شده، طیفها معنی دار تر و دارای قله ها و دره های قابل تفسیری هستند. در حالیکه قبل از پیش پردازش طیفها متأثر از خط پس زمینه و نویز زیادی هستند که تمایز آنها را مشکل می سازد.



الف

شکل ۴- تصویر حرارتی از کل طیفها ۴- الف) قبل از پیش پردازش و ۴- ب) پس از پیش پردازش

<sup>1</sup> Spline interpolation

## ۲-۲- استخراج ویژگی

در رابطه فوق،  $\mu$ ،  $S$  و  $n$  به ترتیب بیانگر میانگین، واریانس و تعداد نمونه های جوامع آماری او ۲ می باشد. به عنوان معیار دیگری برای اندازه گیری میزان جدایی پذیری دو کلاس از فاصله باتاچاریا استفاده نموده ایم. با فرض نرمال بودن توزیع هر کلاس، فاصله باتاچاریا از رابطه ۳ قابل محاسبه است [۱۹].

$$D_B = \frac{1}{8}(m_1 - m_2)^T P^{-1}(m_1 - m_2) + \frac{1}{2} \ln \left( \frac{\det P}{\sqrt{\det P_1 \det P_2}} \right) \quad (۳)$$

در رابطه فوق  $m_i$  و  $P_i$  میانگین و کوواریانس توزیعها و  $P = (P_1 + P_2) / 2$  می باشد.

در نهایت نیز به بررسی میزان تفکیک پذیری دو کلاس در هر عدد موج با استفاده از آنتروپی، مطابق رابطه ۴ پرداختیم. در این رابطه  $i$  اندیس هر کلاس،  $p(S_i)$  احتمال متناظر با کلاس  $S_i$  و  $C$  مجموعه کلاسها می باشد [۲۰].

$$Entropy(S) = - \sum_{i \in C} p(S_i) \log(p(S_i)) \quad (۴)$$

نتایج بدست آمده برای فاصله دو کلاس با استفاده از سه معیار فوق در شکل ۵ نشان داده شده است. همانگونه که ملاحظه می شود، با استفاده از روش آمارگان  $T$  فاصله بیشتری بین دو کلاس وجود دارد. لذا از این روش برای انتخاب ۱۰۰ ویژگی که بیشترین فاصله را در دو کلاس دارند، استفاده شده است.

**۲-۲-۲- استخراج ویژگی با استفاده از الگوریتم ژنتیک**  
در این مرحله از تحقیق با استفاده از الگوریتم ژنتیک به جستجوی ۲۰ ویژگی که بهترین تفکیک را بین دو کلاس برقرار نماید اقدام نمودیم. در اجرای این الگوریتم تعداد نسلها را برابر ۱۰۰ و جمعیت هر نسل را برابر ۵۰ در نظر گرفتیم. تابع هدف استفاده شده جدایی پذیری دو کلاس را با استفاده از ترکیب خطی احتمال پسین و نرخ خطای تجربی طبقه بند خطی ماکزیمم می کند.

در مواردی که تعداد نقاط ویژگی از تعداد نمونه ها خیلی بیشتر باشد استفاده از روشهای داده کاوی<sup>۱</sup> به منظور استخراج و انتخاب ویژگی حائز اهمیت خواهد بود. زیرا استفاده از تمام ویژگیها غیر عملی بوده و موجب کاهش کارایی مدل می گردد. با فرض اینکه ماتریس داده ورودی،  $D_{N \times M}$  با  $N$  نمونه و  $M$  ویژگی وجود دارد، هدف از استخراج ویژگی پیدا کردن یک زیر فضا با  $m$  ویژگی، از فضای مشاهدات دارای  $M$  ویژگی، خواهد بود، به گونه ای که نقاط این زیر فضا از نظر آماری دارای اختلاف معنی دار باشند.

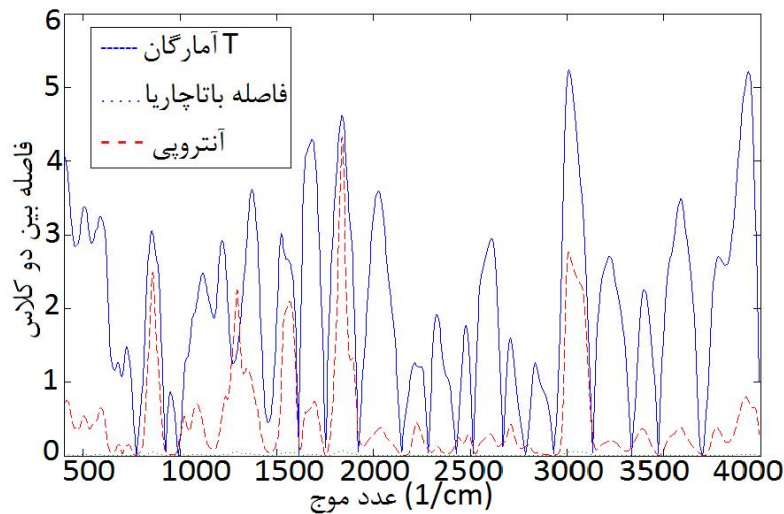
در این تحقیق با استفاده از دو روش آنالیز آماری و الگوریتم ژنتیک، به جستجوی ویژگیهای بهینه به منظور تفکیک نمونه های پوستی پرداخته ایم. سپس به عنوان معیاری برای ارزیابی این روشها در انتخاب ویژگیهای مناسب، به طبقه بندی نمونه های پوستی در دو کلاس نرمال و خشک با استفاده از ویژگیهای بدست آمده و طبقه بندی کننده های آنالیز خطی LDA<sup>۲</sup>، نزدیکترین همسایگی KNN<sup>۳</sup>، ماشین حامی بردار SVM<sup>۴</sup> و شبکه عصبی NN<sup>۵</sup> پرداخته ایم.

## ۲-۲-۱- استخراج ویژگی با استفاده از آنالیز آماری

ابتدا به منظور تعیین میزان تفکیک پذیری دو کلاس به بررسی فاصله دو کلاس با استفاده از معیارهای آمارگان  $T$ <sup>۶</sup>، آنتروپی<sup>۷</sup> و باتاچاریا<sup>۸</sup> پرداخته ایم. در این آزمونها مقدار  $p$ -value کمتر از ۰/۰۵ در نظر گرفته شده است. در آزمون آمارگان  $T$ ، با فرض توزیع نرمال نمونه ها در هر کلاس، از رابطه ۲ برای محاسبه آمارگان  $T$  استفاده نموده ایم [۱۸].

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (۲)$$

<sup>۱</sup> Data mining  
<sup>۲</sup> Linear Discriminant Analysis  
<sup>۳</sup> K Nearest Neighbors  
<sup>۴</sup> Support Vector Machine  
<sup>۵</sup> Neural Network  
<sup>۶</sup> T-test  
<sup>۷</sup> Entropy  
<sup>۸</sup> Bhattacharyya



شکل ۵- نمایش فواصل آمارگان T، باتاچاریا و آنتروپی بین دو کلاس در هر عدد موج

### ۳- نتایج

همانگونه که قبلاً اشاره شد به منظور مقایسه ویژگیهای انتخاب شده با استفاده از آمارگان T و الگوریتم ژنتیک از طبقه بندی کننده های LDA، KNN، SVM و شبکه عصبی برای جدا سازی دو گروه با استفاده از ویژگیهای انتخاب شده توسط این دو روش، بهره بردیم.

در طبقه بندی، از طبقه های ورودی یک مجموعه یادگیری و آزمون بطور تصادفی انتخاب شد. با توجه به تعداد کم نمونه ها، از روش ارزیابی چرخشی، جهت جلوگیری از ایجاد هرگونه بایاس و خطا هنگام انتخاب ویژگی و طبقه بندی استفاده شد. بدین ترتیب در هر چرخش ۵٪ از طبقه ها به منظور ارزیابی و مابقی برای آموزش استفاده گردید و این کار ۱۰۰ بار روی داده ها صورت گرفت و در نهایت روی نتایج میانگین گیری انجام شد.

در مورد هر طبقه بند پارامترهای تشخیص صحیح یا دقت<sup>۱</sup>، حساسیت<sup>۲</sup> و قطعیت<sup>۳</sup> به عنوان معیار کارایی طبقه بند استفاده استفاده شده است. در هر مورد دقت بصورت نسبت تشخیصهای صحیح به تعداد کل نمونه ها، حساسیت بصورت

نسبت نمونه های پوستی که به درستی خشک تشخیص داده شده اند به کل نمونه های پوست خشک و قطعیت بصورت نسبت تعداد نمونه هایی که به درستی نرمال تشخیص داده شده اند به کل نمونه های نرمال، محاسبه شده است.

در اولین قدم، مقدار آمارگان T برای مجموعه یادگیری محاسبه گردید. سپس تعداد ۱۰۰ ویژگی که آمارگان T بالاتری نسبت به بقیه داشتند، انتخاب شد و با استفاده از طبقه بندی کننده های ذکر شده به جدا سازی دو گروه پرداختیم که نتایج آن در جدول ۱ نشان داده شده است.

در گام دوم به منظور بهبود میزان تفکیک دو گروه، تعداد ویژگیها با استفاده از آنالیز مولفه های اصلی PCA<sup>۴</sup>، از ۱۰۰ ویژگی به ۲۰ ویژگی کاهش داده شد و بار دیگر طبقه بندی تکرار گردید. نتایج این بخش نیز در جدول ۱ در ستونهای با عنوان PCA نشان داده شده است. مشاهده می کنیم که دو طبقه بندی کننده LDA و SVM پس از استفاده از PCA تفکیک بهتری بین دو کلاس ایجاد نموده اند.

در مرحله بعدی با استفاده از الگوریتم ژنتیک به یافتن ۲۰ ویژگی بهینه برای جدا سازی دو گروه پرداختیم. نتایج طبقه بندی با استفاده از این ویژگیها در جدول ۲ نشان داده شده است. همانگونه که مشاهده می شود با استفاده از این روش

<sup>1</sup> Correct Rate

<sup>2</sup> Sensitivity

<sup>3</sup> Specificity

<sup>4</sup> Principle Component Analysis

داده شد که الگوریتم ژنتیک در استخراج ویژگیهای بهینه از طیفهای رامان کارآمدتر از روشهای آماری است. بگونه ای که با استفاده از این روش در استخراج ویژگی و استفاده از طبقه بند مناسبی همچون SVM توانستیم به دقت ۹۰٪، حساسیت ۸۹٪ و قطعیت ۹۱٪ در تمایز نمونه های پوستی نرمال از خشک دست یابیم.

درصد تشخیص درست برای تمام طبقه بندی کننده ها حتی نسبت به حالت استفاده از آمارگان T به همراه PCA افزایش یافته است.

#### ۴- بحث

از یافته های بدست آمده می توان دریافت که روش استخراج ویژگی نقش مهمی در کارایی طبقه بندی دارد و در اینجا نشان

جدول ۱- نتایج طبقه بندی پوست نرمال و خشک پس از استخراج ویژگی با استفاده از آزمون T، بدون PCA و با PCA

	NN		SVM		KNN		LDA	
	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
تشخیص صحیح	۰/۸۰۳۳	۰/۸۲۰۱	۰/۸۶	۰/۸۵۵۷	۰/۸۱۴۳	۰/۸۲	۰/۸۶۴۳	۰/۷۷۲۹
حساسیت	۰/۹۲۱۲	۰/۹۲۱۰	۰/۸۹۶۷	۰/۸۹۶۷	۰/۸۹	۰/۹۷۳۳	۰/۹۴	۰/۸۲۳۳
قطعیت	۰/۷۸۵۷	۰/۷۹۴۳	۰/۸۳۲۵	۰/۸۲۵۰	۰/۷۵۷۵	۰/۷۰۵۰	۰/۸۰۵۷	۰/۷۳۵۰

می شوند و نیز از سوی دیگر بسیاری از بیماریهای پوستی منجر به خشکی پوست می شود.

لذا نتایج برآمده از این تحقیق پتانسیل طیف سنجی رامان را به عنوان ابزاری قدرتمند برای تعیین کیفیت محصولات آرایشی و بهداشتی و یا تشخیص بیماریهای پوستی نشان می دهد.

امید است که نتایج این تحقیق بتواند در کمک به انتخاب روش پیش پردازش و داده کاوی مناسب در تحقیقات آینده به منظور تشخیص تغییرات و یا ناهنجاریهای پوستی با استفاده از طیف سنجی رامان موثر باشد.

#### ۵- تشکر و قدردانی

این مقاله برگرفته از رساله دکتری تحت حمایت مالی معاونت پژوهشی دانشگاه تربیت مدرس می باشد که بدینوسیله از این معاونت محترم قدردانی می گردد.

جدول ۲- نتایج طبقه بندی پوست نرمال و خشک پس از استخراج ویژگی با استفاده از الگوریتم ژنتیک

	NN	SVM	KNN	LDA
تشخیص صحیح	۰/۸۲۲۹	۰/۹۰۲۹	۰/۸۴۱۴	۰/۸۹۷۱
حساسیت	۰/۸۳۳۹	۰/۸۹۳۳	۰/۸۸۳۳	۰/۹۱۳۳
قطعیت	۰/۸۲	۰/۹۱	۰/۸۱	۰/۸۸۵۰

#### ۵- نتیجه گیری

نتایج حاصل از این تحقیق نشاندهنده پتانسیل طیف سنجی رامان در تشخیص تغییرات نمونه های بیولوژیکی است. همانگونه که در این تحقیق نشان داده شد الگوریتم ژنتیک کارایی بالایی در استخراج ویژگی از طیفهای رامان نسبت به روشهای آماری دارد. بگونه ای که با استفاده از این الگوریتم و طبقه بند مناسب توانستیم به درصد تمایز مطلوبی (بالغ بر ۹۰٪) بین نمونه های پوستی نرمال و خشک دست یابیم. با توجه به اینکه بسیاری از مواد آرایشی و بهداشتی بر محتوای آب پوست تاثیر می گذارند و موجب خشکی پوست

#### منابع

1. Lewis IR. Handbook of Raman Spectroscopy. 1st ed. Madison: Marcel Dekker; 2001.
2. Park A, Jung IW, Baek SJ, Kim JY, Shin D. An enhancement on the detection of skin cancer based on band approach of confocal raman spectra: Proceeding of International Symposium on Information Technology Convergence; 2007 Nov 23-24; Jeonju, Korea. IEEE Computer Society Washington, DC.; 2007.273-7.

3. Pappas D, Smith BW, Winefordner JD. Raman spectroscopy in bioanalysis. *Talanta*. 2000 24;51(1):131-44.
4. Lyng FM, Faoláin EO, Conroy J, Meade AD, Knief P, Duffy B, et al. Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool. *Exp Mol Pathol*. 2007; 82(2):121-9.
5. Zhao J, Lui H, McLean DI, Zeng H. Real-time Raman spectroscopy for non-invasive skin cancer detection - preliminary results. *Conf Proc IEEE Eng Med Biol Soc*. 2008;2008:3107-9.
6. Kudelski A. Analytical applications of raman spectroscopy. *Talanta*. 2008;76(1):1-8.
7. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, et al. Melanoma Diagnosis by Raman spectroscopy and neural networks: structure alterations in protein and lipids in intact cancer tissue. *J Invest Dermatol*. 2004;122(2):443-9.
8. Li X, Lin J, Jin H, Tang J, Qing S, Liu Q. Spectral analysis for diagnosis of esophagus dysplasia using fluorescence and Raman spectroscopy. *Conf Proc IEEE Eng Med Biol Soc*. 2004;1:141-4.
9. Sigurdsson S, Philipsen PA, Hansen LK, Larsen J, Gniadecka M, Wulf HC. Detection Of Skin Cancer By Classification Of Raman Spectra. *IEEE Trans Biomed Eng*. 2004;51(10):1784-93.
10. Gobinet C, Vrabie V, Tfayli A, Piot O, Huez R, Manfait M. Pre-processing and Source Separation Method for Raman spectra analysis of biomedical samples. *Conf Proc IEEE Eng Med Biol Soc*. 2007;6208-11.
11. Hyun Chung S, Suk Park K. A study on the factor number determination methods in the partial least squares model for the urinalysis using raman spectroscopy. *Conf Proc IEEE Eng Med Biol Soc*. 2004;1:490-3.
12. Wang W, Adalı T, Emge D. Target detection and identification using canonical correlation analysis and subspace partitioning: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP; 2008, Mar 30 - April 4; Caesars Palace, Las Vegas, Nevada, USA. IEEE; 2008. p. 2117-20.
13. Wang W, Adalı T. Constrained ICA and its Application to Raman Spectroscopy: Proceedings of the Antennas and Propagation Society International Symposium; 2005, Jul 3-8; IEEE;4B:109-12.
14. Wang W, Adalı T, Li H, Emge D. Detection using correlation bound and its application to raman spectroscopy: Proceeding of the IEEE International Workshop On Machine Learning For Signal Processing; 2005 Sep 28-30; Mystic, Connecticut, USA. IEEE;2005:259-62.
15. Ye Z, Ye Y, Mohamadian H, Bhattacharya P, Kang K. Fuzzy Filtering and Fuzzy K-Means Clustering on Biomedical Sample Characterization: Proceeding of the IEEE Conference on Control Applications; 2005, Aug 28-31; Toronto, Canada; IEEE;2005:90-5.
16. Andrade L, Manolakos ES. Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing. *The Journal of VLSI Signal Processing*. 2003;35(3):229-43.
17. Draper NR, Smith H. *Applied Regression Analysis*. 3th ed. New York: Wiley; 1998.
18. Hollander M, Wolfe D. *Nonparametric Statistical Methods*. 2th ed. New York: Wiley; 1999.
19. Kailath T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. on Comm. Technology*. 1967;15:52-60.
20. Giannopoulou EG. *Data Mining in Medical and Biological Research*. 1st ed. Vienna: In-The; 2008.