

## تشخیص سرطان پروستات و پیش‌بینی احتمال ابتلا به این بیماری در شاغلین

فاطمه سرباز عقدائی<sup>۱</sup>، ولی درهمی<sup>۲\*</sup>، محمد جواد زارع سخویدی<sup>۳</sup>، مهرداد مستغاثی<sup>۴</sup>  
رضیه سلطانی گردفرامرزی<sup>۵</sup>، مریم موسوی<sup>۶</sup>

### چکیده

مقدمه: تشخیص بیماری‌های مختلف در علم پزشکی یکی از زمینه‌های پرکاربرد داده‌کاوی محسوب می‌شود که در سال‌های اخیر تحقیقات و مطالعات فراوانی پیرامون آن انجام شده است. در این مقاله به مساله تشخیص سرطان پروستات با استفاده از سیستم فازی پرداخته شده است. هدف، تشخیص سرطان پروستات و هم‌چنین پیش‌بینی میزان احتمال ابتلا به این بیماری است. روش بررسی: در روش پیشنهادی ابتدا بر روی مجموعه داده در دسترس، پیش‌پردازش انجام شده و در مرحله بعد عملیات خوشه‌بندی انجام می‌گیرد. سپس یک سیستم فازی مرتبه صفر سوگنو برای پیش‌بینی طراحی می‌شود. هر خوشه به عنوان مقدم یک قاعده فازی و خروجی هر قاعده درصد احتمال بیماری در آن خوشه در نظر گرفته می‌شود. برای هر نمونه جدید، درجه تعلق به هر خوشه محاسبه و سپس با ترکیب خروجی‌های حاصل از هر قاعده، میزان احتمال ابتلای آن نمونه به سرطان پروستات، پیش‌بینی می‌شود. در نهایت با در دست داشتن این احتمال و حد آستانه در نظر گرفته شده برای این احتمال، ابتلا یا عدم ابتلای نمونه موردنظر به سرطان پروستات تشخیص داده می‌شود.

نتایج: نتایج آزمایشات حاکی از دقت خوب سیستم ارائه شده در پیش‌بینی احتمال بیماری می‌باشد. نتیجه‌گیری: از نتایج حاصل از این تحقیق می‌توان به منظور پیش‌بینی احتمال ابتلای شاغلین جوان به سرطان پروستات با توجه به مشاغل مختلفی که به آن‌ها اشتغال دارند و میزان مواجهه با عوامل خطرزا در هر یک از مشاغل استفاده کرد که در صورت بالا بودن این احتمال، به عنوان فرد در معرض خطر شناخته شده و در مواردی حتی شاید نیاز به تغییر شغل وجود داشته باشد.

واژه‌های کلیدی: پیش‌بینی، سرطان پروستات، سیستم‌های فازی

۱- کارشناس ارشد مهندسی نرم افزار، دانشگاه علم و هنر یزد، یزد، ایران

۲- دانشیار، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

۳- استادیار، گروه بهداشت حرفه‌ای، دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران

۴- متخصص طب کار، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران

۵- کارشناس ارشد بهداشت حرفه‌ای، عضو مرکز تحقیقات بیماری‌های ناشی از صنعت، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید صدوقی یزد، یزد، ایران

۶- کارشناس ارشد آموزش پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

\* (نویسنده مسئول): تلفن: ۳۱۲۳۲۳۶۵، پست الکترونیکی: vderhami@yazd.ac.ir

تاریخ پذیرش: ۱۳۹۴/۰۸/۱۷

تاریخ دریافت: ۱۳۹۳/۱۱/۲۱

## مقدمه

سرطان پروستات بعد از سرطان ریه، دومین سرطان شایع در بین مردان محسوب می‌شود که در کشورهای پیشرفته در رتبه اول قرار دارد (۱،۲). این سرطان یکی از مهم ترین علل مرگ و میر در بین مردان در کشورهای اروپایی می‌باشد (۳). در ایران هم از گذشته تاکنون شیوع این بیماری افزایش چشمگیری داشته به طوری که در سال ۱۳۶۵، بر طبق گزارش اداره کل مبارزه با بیماری‌های غیرواگیر مرکز مبارزه با سرطان، جزء سرطان‌های غیرشایع و با ۱/۶٪ کل سرطانهای کشور در رده ۱۳ بوده و در سال ۸۴ با شامل شدن ۸/۶٪ کل سرطان‌های کشور به رده چهارم رسیده است (۴). احتمال ابتلا به سرطان پروستات برای مردان در طول زندگی ۱ در ۶ نفر است و ۱ به ۳۲ نفر به علت این نوع سرطان فوت می‌کنند (۵). اثرات طولانی مدت درمان مانند بی اختیاری ادراری، ناتوانی جنسی و التهاب رکتوم ناشی از پرتو‌تابی، تاثیرات بسیار نامطلوبی روی کیفیت زندگی بیماران دارد (۶). متوسط طول مدت اقامت در بیمارستان برای بیمار مبتلا به سرطان پروستات بین ۵ تا ۱۰ روز است که هزینه زیادی را برای وی به وجود می‌آورد و همچنین باری بر سیستم درمانی است (۷). سرطان شغلی ۴ تا ۲۰ درصد از سرطان‌های شایع را تشکیل می‌دهد. این نوع سرطان گروه‌های خاصی از جامعه را بیشتر درگیر نموده و بر خلاف بسیاری از مخاطرات ناشی از روش زندگی، مخاطرات شغلی سرطان معمولاً غیرارادی می‌باشند. سرطان شغلی کاملاً قابل پیش‌گیری بوده و مداخلات در سطح محیط‌های کاری می‌توانند جان میلیون‌ها نفر را در سال حفظ نمایند. علت سرطان شغلی مواجهه با عوامل سرطان زا در محیط کاری می‌باشد. بهترین و آسان‌ترین روش برای مبارزه با مشکلات و عوارض ناشی از یک بیماری، پیشگیری از آن بیماری است. یکی از راه‌های پیشگیری آن است که افراد در معرض خطر ابتلا و افرادی که احتمال ابتلای آن‌ها به این بیماری بالاتر است شناسایی شوند و با کاهش عوامل موثر در ابتلای افراد به این بیماری، از بروز بیماری در این افراد جلوگیری شود. با توجه به آن که سرطان پروستات، سرطانی است که اغلب در سنین بالا

رخ می‌دهد، پیش‌بینی آن در افراد جوان باعث می‌شود فرصت کافی برای افراد در معرض ابتلا برای ایجاد تغییر در شیوه زندگی به منظور پیشگیری از بیماری وجود داشته باشد. سرطان پروستات، مثل بسیاری از سرطان‌ها علاوه بر آن که به عوامل ژنتیکی و سابقه فامیلی بستگی دارد، وابستگی زیادی نیز به سبک زندگی، نوع تغذیه، و همچنین شغل افراد دارد. شغل و تماس با عوامل زیان‌آور محیط کار مثل عوامل شیمیایی، فلزات سنگین و غیره، از جمله عوامل مهمی است که تاثیر زیادی در ابتلای مردان به سرطان پروستات دارد. به همین دلیل چنان چه بتوان با دریافت برخی اطلاعات از مردان شاغل، مانند نوع تغذیه، سبک زندگی، میزان تماس با عوامل زیان‌آور محیط کار و غیره، پیش‌بینی کرد که هر فرد به چه میزان احتمال ابتلا به این بیماری را در آینده دارد، سبب خواهد شد افرادی که احتمال ابتلای بالایی دارند شناسایی شوند و با تغییر در سبک زندگی و یا حتی تغییر شغل خود از ابتلا به این بیماری در آینده جلوگیری کنند. سیستم‌های فازی، سیستم‌های پیش‌بینی کننده ای هستند که در زمینه تشخیص پزشکی به منظور برطرف کردن مشکلات عدم قطعیت مورد توجه قرار گرفته‌اند. با آموزش بانک اطلاعاتی شامل تشخیص‌های قبلی به سیستم فازی، این سیستم قادر خواهد بود به صورت خودکار، مساله تشخیص بیماری را انجام دهد.

## داده‌های آزمایش

برای رسیدن به هدف این مقاله، مجموعه داده‌ای تهیه شده است که حاوی برخی اطلاعات شخصی و دموگرافیک از تعدادی از افراد در دو گروه است. گروه اول، گروه مورد است که شامل افراد مبتلا به سرطان پروستات می‌شود که بومی استان فارس بوده‌اند و نام آن‌ها در طی سال‌های ۱۳۸۹ و ۱۳۹۰ در نرم افزار ثبت سرطان واحد مبارزه با بیماری‌های غیر واگیر دانشگاه علوم پزشکی شیراز ثبت شده است. گروه دیگر، گروه شاهد است که تعداد آن دو برابر گروه مورد است و شامل افراد سالمی هستند که با گروه مورد، از نظر فاکتور سن همسان‌سازی شده است. این مجموعه داده، شامل اطلاعات مربوط به ۶۳۶ نفر

پروستات مبتلا هست یا نه. به علاوه، این سیستم می‌بایست قادر باشد میزان احتمال ابتلای افراد به این بیماری را نیز پیش‌بینی کند. برای رسیدن به این منظور ابتدا عملیات پیش‌پردازش و خوشه‌بندی بر روی مجموعه داده انجام می‌شود. سپس یک سیستم فازی مرتبه صفر سوگنو برای پیش‌بینی طراحی می‌شود. هر خوشه به عنوان مقدم یک قاعده فازی و خروجی هر قاعده، درصد احتمال بیماری در آن خوشه در نظر گرفته می‌شود. برای هر نمونه جدید، درجه تعلق به هر خوشه محاسبه و سپس با ترکیب خروجی‌های حاصل از هر قاعده، میزان احتمال ابتلای آن نمونه به سرطان پروستات، پیش‌بینی می‌شود. در نهایت با در دست داشتن این احتمال و حد آستانه در نظر گرفته شده برای این احتمال، ابتلا یا عدم ابتلای نمونه مورد نظر به سرطان پروستات تشخیص داده می‌شود. با توجه به درصد احتمال بیماری در هر خوشه و درجه تعلق نرمالایز شده فرد به هر خوشه می‌توان با استفاده از فرمول زیر خروجی این سیستم برای هر نمونه‌ای که به آن وارد می‌شود را محاسبه کرد. این خروجی، میزان احتمال ابتلای نمونه جدید به سرطان پروستات است.

$$\sum_{i=1}^n P(i) * Mi(x) \quad (1)$$

$P(i)$ : احتمال ابتلا به بیماری در خوشه  $i$ ،  $Mi(x)$ : درجه تعلق نرمالایز شده نمونه  $x$  به خوشه  $i$ ،  $n$ : تعداد خوشه‌ها  
به این صورت میزان احتمال ابتلای فرد ارائه شده به سیستم به سرطان پروستات محاسبه می‌شود. با استفاده از این احتمال و با معرفی یک حد آستانه به سیستم، سیستم فازی هم چنین می‌تواند ابتلا یا عدم ابتلای فرد به سرطان پروستات را تشخیص دهد. این امر به این صورت انجام خواهد گرفت که اگر این احتمال برای فرد جدید بیشتر از حد آستانه بود، آن فرد به عنوان فرد بیمار و در غیر این صورت به عنوان فرد سالم شناخته خواهد شد.

پیش‌پردازش روی داده‌ها

ابتدا می‌بایست بر روی مجموعه داده‌ای پیش‌پردازش‌هایی انجام شود تا داده برای انجام مراحل بعدی آماده شود. این

است. پارامترهایی که در مورد این افراد مورد استفاده قرار می‌گیرد عبارتند از: سن، وزن، قد، قومیت که دارای مقادیر فارس و لر و ترک و عرب است، میزان تحصیلات که دارای مقادیر بی‌سواد، دبستان و راهنمایی، دبیرستان و دانشگاهی است، داشتن یا نداشتن مصرف سیگار، سابقه مصرف سیگار (چند سال)، تعداد نخ سیگار مصرفی در روز، داشتن یا نداشتن مصرف قلیان، تعداد دفعات مصرف قلیان در روز، داشتن یا نداشتن مصرف مواد مخدر، داشتن یا نداشتن سابقه خانوادگی سرطان، داشتن یا نداشتن سابقه بیماری، میزان مصرف گوجه‌فرنگی و یا رب و سس گوجه‌فرنگی که شامل مقادیر کم، متوسط و زیاد است، میزان مصرف سبزیجاتی مثل کاهو و کلم که این مورد هم شامل مقادیر کم، متوسط و زیاد است، داشتن یا نداشتن تولد در خانواده روستایی، داشتن یا نداشتن زندگی در روستا، تعداد سال‌های زندگی در روستا، داشتن یا نداشتن زندگی در شهر، تعداد سال‌های زندگی در شهر، داشتن یا نداشتن زندگی در بیابان، تعداد سال‌های زندگی در بیابان، شغل کنونی، نوع فعالیت اوقات فراغت، نام تمام مشاغلی که فرد در طول زندگی آن‌ها را انجام داده است به همراه تعداد سال‌های اشتغال به آن مشاغل، ذکر انجام کار مربوطه به صورت "هر روز" یا "گاهی" و انجام کار مربوطه بیشتر از دو ساعت در روز یا کم تر از آن، و در نهایت بررسی میزان مواجهه مشاغل فرد با عواملی مانند کار فیزیکی، حشره کش ها، ارتعاش، نشستن طولانی مدت، دود دیزل، آلودگی تراشکاری، دود فلزی و جوشکاری، اشعه ماوراء بنفش و حلال‌ها. پارامتری که به عنوان خروجی مجموعه داده در نظر گرفته شده است، ابتلا یا عدم ابتلا به سرطان پروستات است که این پارامتر نیز برای افراد شرکت‌کننده در مطالعه، مشخص شده است.

روش پیشنهادی

در این مطالعه، هدف آن است که این مجموعه داده به سیستم فازی آموزش داده شود و سیستم فازی مذکور به عنوان یک پیش‌بینی‌کننده بتواند برای موارد ناشناخته‌ای که به آن ارائه می‌شود، تشخیص دهد فرد مورد آزمایش به سرطان

داده آموزشی، ۴۰۲ نمونه داریم که ۲۷۸ مورد از آن ها سالم و ۱۲۴ مورد از آن ها مبتلا هستند. داده آزمایشی نیز دارای ۱۰۰ نمونه است که ۷۸ مورد سالم و ۲۲ مورد مبتلا در آن وجود دارد.

۲- حذف پارامترهایی که با یکدیگر هم بستگی دارند که در مورد دو داده آموزشی و آزمایشی انجام خواهد گرفت. در این مرحله دو پارامتر میزان مواجهه با دود دیزل و میزان مواجهه با کارفیزیکی حذف می‌شود.

۳- انجام نرمال سازی که در مورد دو داده آموزشی و آزمایشی انجام خواهد گرفت.

بعد از انجام عملیات فوق، در نهایت به مجموعه داده ای با ۵۰۲ رکورد و ۱۷ پارامتر ورودی خواهیم رسید. البته پارامتر دیگری نیز برای هر رکورد داده وجود دارد که این پارامتر، برچسب سالم یا بیمار بودن برای هر نمونه است که به عنوان پارامتر خروجی شناخته می‌شود. در واقع توسط سیستم فازی و با استفاده از ۱۷ پارامتر ورودی، این پارامتر خروجی پیش بینی می‌شود.

این ۱۷ پارامتر عبارتند از:

سن، تحصیلات، BMI، میزان مصرف سبزیجاتی مانند کلم و کاهو، میزان مصرف گوجه فرنگی، رب و سس آن، سابقه خانوادگی سرطان، سابقه خانوادگی سرطان پروستات، سابقه مصرف سیگار (سال) \* تعداد نخ سیگار مصرفی در روز، تعداد دفعات مصرف قلیان در روز، مصرف یا عدم مصرف مواد مخدر، میزان مواجهه با عامل اشعه ماوراء بنفش، میزان مواجهه با عامل حشره‌کش‌ها، میزان مواجهه با عامل ارتعاش، میزان مواجهه با عامل نشستن طولانی مدت، میزان مواجهه با عامل دود فلزی و جوشکاری، میزان مواجهه با عامل آلودگی تراشکاری، میزان مواجهه با عامل حلال‌ها.

خوشه‌بندی

انجام عملیات خوشه‌بندی بر روی داده آموزشی که توسط نرم‌افزار ریپید ماینر انجام می‌گیرد. خوشه‌بندی به هر یک از روش‌های متفاوت k-means، k-medoids و x-means، انجام می‌گیرد.

پیش‌پردازش‌ها توسط نرم افزار ریپیدماینر انجام شده است و عبارتند از:

۱- اضافه کردن ستون id برای هر رکورد از داده؛

۲- حذف رکوردهایی که دارای برخی ستون‌های خالی هستند؛

۳- تبدیل تمام مقادیر اسمی به مقادیر عددی به صورت زیر:

تبدیل مقادیر "بله" به عدد ۱ و تبدیل مقادیر "خیر" به عدد صفر، برای پارامتر تحصیلات مقادیر بی‌سواد، دبستان و راهنمایی، دبیرستان و دانشگاه وجود دارد که این مقادیر به ترتیب به اعداد ۱ تا ۴ تغییر پیدا می‌کند، برای پارامتر داشتن یا نداشتن مصرف سیگار، مقادیر "هرگز" و "دارد" وجود دارد که به ترتیب به مقادیر صفر و ۱ تغییر پیدا می‌کند، برای پارامتر میزان مصرف گوجه‌فرنگی و پارامتر میزان مصرف سبزیجات مقادیر کم، متوسط و زیاد وجود دارد که به ترتیب به مقادیر ۰/۲۵، ۰/۵ و ۰/۷۵ تغییر پیدا می‌کند.

۱- حذف رکورد مربوط به افرادی که قومیتی به غیر از فارس دارند تا در بانک اطلاعاتی، فقط اطلاعات افراد با قومیت فارس وجود داشته باشد.

۲- حذف برخی از پارامترهای کم اهمیت یا بدون استفاده از داده

۳- ادغام برخی از پارامترها در یکدیگر و محاسبه پارامترهای جدید و حذف پارامترهای اولیه به منظور کاهش ابعاد داده

- محاسبه پارامتر BMI با استفاده از دو پارامتر وزن و قد
- ضرب «سابقه مصرف سیگار» در «تعداد نخ سیگار مصرفی

در روز» و ایجاد فاکتوری جدید

واضح است که پارامترهای وزن، قد، سابقه مصرف سیگار و تعداد نخ مصرفی در روز حذف خواهد شد.

۱- تقسیم تصادفی داده‌ها به دو بخش داده آموزشی و داده آزمایشی انجام می‌گیرد، به این صورت که ۸۰٪ از داده‌ها به صورت تصادفی برای آموزش و ۲۰٪ از داده‌ها نیز به صورت تصادفی برای آزمایش کنار گذاشته خواهد شد. البته این تقسیم تصادفی به این صورت است که نرخ افراد بیمار به سالم که در داده اولیه وجود دارد، در داده آموزشی و آزمایشی نیز رعایت خواهد شد. با این تقسیم بندی، برای

سپس عملیات زیر توسط همین نرم افزار بر روی خوشه‌های ایجاد شده انجام خواهد گرفت:

۱- محاسبه میانگین و انحراف معیار هر پارامتر ورودی در هر خوشه

در مورد هر یک از خوشه‌ها می‌بایست میانگین و انحراف معیار هر یک از ۱۷ پارامتر ورودی گرفته شده را به دست آورد. به این صورت که اگر در عملیات خوشه‌بندی، تعداد خوشه‌ها را به طور مثال ۲۰ در نظر بگیریم، ماتریسی با ۲۰ سطر و ۱۷ ستون برای هر یک از موارد میانگین و انحراف معیار داریم. به این صورت که هر یک از سطرها نشان‌دهنده هر یک از خوشه‌ها و هر یک از ستون‌ها نشان‌دهنده هر یک از ۱۷ پارامتر ورودی است و مقادیر موجود در هر یک از فیله‌ها نشان‌دهنده میانگین و یا انحراف معیار در خوشه و پارامتر متناظر است.

۲- محاسبه احتمال بیماری در هر خوشه

از بین نمونه‌هایی که در هر خوشه وجود دارد، تعدادی دارای برچسب سالم و تعدادی دارای برچسب بیمار هستند. برای محاسبه احتمال بیماری، نسبت بیماران در یک خوشه به کل نمونه‌های موجود در آن خوشه محاسبه شده است.

محاسبه درجه تعلق نمونه  $x$  به خوشه  $i$  و همچنین درجه تعلق نرمالایز شده:

ابتدا برای نمونه جدیدی که به سیستم وارد می‌شود، به ازای هر یک از ابعادش، درجه تطابق آن بعد با مجموعه فازی متناظر در خوشه مورد نظر محاسبه می‌شود. با توجه به آن که ورودی ۱۷ بعدی هست لذا می‌بایست ۱۷ درجه تطابق به دست آید. به منظور محاسبه درجه تعلق نمونه به خوشه مورد نظر، اپراتور اشتراک بر روی درجه تطابق‌ها اعمال می‌شود. با انجام عملیات گفته شده، درجه تعلق نمونه به هر یک از خوشه‌ها محاسبه می‌شود.

مراحل عملیات گفته شده در ادامه توضیح داده می‌شود:

۱- رکورد مربوط به نمونه  $x$  از مجموعه داده واکنشی و مورد استفاده قرار می‌گیرد. این رکورد دارای ۱۷ ویژگی می‌باشد که قبلاً به طور کامل توضیح داده شد.

۲- در مورد هر یک از ویژگی‌های رکورد مربوط به نمونه  $x$ ، درجه تطابق آن ویژگی با بعد متناظر در خوشه  $i$  محاسبه می‌شود.

در واقع این مرحله به محاسبه درجه تطابق ویژگی  $b$  ام نمونه  $x$  با خوشه  $i$  می‌پردازد. در ادامه توضیح بیشتری در مورد این مرحله داده می‌شود:

۷ همان‌طور که قبلاً گفته شد برای هر ویژگی مجموعه داده، میانگین و انحراف معیار آن ویژگی در هر خوشه محاسبه شده است.

۷ با استفاده از دو مقدار میانگین و انحراف معیار در مورد ویژگی  $b$  ام در خوشه  $i$  ام، تابع عضویت گوسی برای آن ویژگی در آن خوشه رسم می‌شود و سپس درجه تطابق ویژگی  $b$  ام از نمونه  $x$  با این تابع عضویت گوسی محاسبه می‌شود.

در مورد هر یک از ویژگی‌های موجود در رکورد نمونه  $x$  درجه تطابق با تابع عضویت گوسی متناظر محاسبه می‌شود. به دلیل وجود ۱۷ ویژگی در مورد هر رکورد، در این جا ۱۷ درجه تطابق به دست خواهیم آورد.

۳- برای محاسبه درجه تعلق نمونه  $x$  به خوشه  $i$  می‌بایست بر روی این درجه تطابق‌ها اپراتور اشتراک اعمال شود.

در این مطالعه دو عملگر ضرب و مینیمم به عنوان اپراتور اشتراک مورد استفاده قرار گرفته است. در نهایت به تعداد خوشه‌ها درجه تعلق برای نمونه مورد نظر ایجاد می‌شود.

در مرحله بعدی می‌بایست درجه تعلق نرمالایز نمونه به هر خوشه محاسبه شود که از طریق فرمول زیر این امر انجام می‌شود:

$$M_i(x) = \frac{m_i(x)}{\sum_{b=1}^n m_b(x)} \quad (2)$$

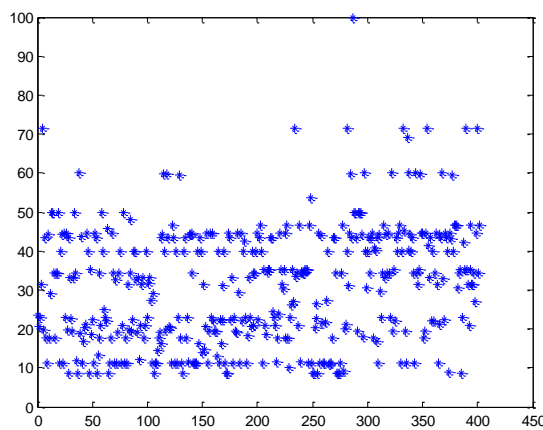
$M_i(x)$ : درجه تعلق نرمالایز شده نمونه  $x$  به خوشه  $i$ ،  $m_i(x)$ : درجه تعلق نمونه  $x$  به خوشه  $i$ ،  $m_b(x)$ : درجه تعلق نمونه  $x$  به خوشه  $b$ ،  $n$ : تعداد خوشه‌ها

ارزیابی روش پیشنهادی

همان‌طور که قبلاً گفته شد می‌بایست بر روی دیتای آموزش، خوشه‌بندی انجام گیرد. به منظور آزمایش روش پیشنهادی، از چندین نوع خوشه‌بندی استفاده شده است. از

مینیمم و ضرب در نظر گرفته شده است و نتایج اعمال هر یک از این توابع بر روی روش پیشنهادی مورد بررسی قرار گرفته است. در حالتی که از روش خوشه‌بندی k-medoids با تعداد ۲۰ خوشه و حد‌آستانه ۵۰ و اپراتور اشتراک مینیمم استفاده شده است، روش پیشنهادی برای داده آموزشی به بیشترین مقدار معیار صحت یعنی ۷۳٪ رسیده است. همان‌طور که قبلاً شرح داده شد، برای هر یک از نمونه‌های موجود در داده آموزشی احتمال ابتلا به بیماری سرطان پروستات محاسبه می‌شود.

شکل ۱ این احتمال را برای هر یک از ۴۰۲ نمونه مورد بررسی در داده آموزشی نشان می‌دهد.

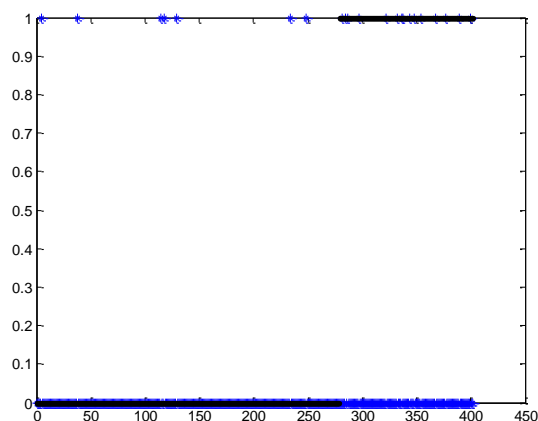


شکل ۱: درصد احتمال بیماری درمورد ۴۰۲ نمونه موجود در داده آموزشی

بیماری در این حالت همان‌طور که در شکل ۱ نشان داده شده است و همچنین حد‌آستانه ۵۰، تشخیص ابتلا یا عدم ابتلا ۴۰۲ نمونه موجود در داده آموزشی به بیماری مشخص شده است که در شکل ۲ این امر با نماد \* (ستاره) نشان داده می‌شود. البته در همین شکل خروجی واقعی نیز با نماد . (نقطه) نشان داده شده است.

خوشه‌بندی k-means با تعداد ۱۰، ۱۵، ۲۰، ۲۵ خوشه به منظور ارزیابی روش استفاده می‌شود. همچنین از خوشه‌بندی k-medoids با تعداد ۱۰ و ۲۰ خوشه و خوشه‌بندی x-means با تعداد ۴ و ۱۴ خوشه استفاده می‌شود. نتایج عملکرد سیستم برای هر یک از انواع خوشه‌بندی و هر تعداد خوشه بررسی می‌شود. فاکتور دیگری که مقدار آن بر روی آزمایش‌ها تاثیر زیادی دارد، حد‌آستانه است. همان‌طور که قبلاً صحبت شد، بعد از به دست آوردن میزان احتمال ابتلای یک نمونه به سرطان پروستات، با استفاده از مقدار حد‌آستانه، سیستم می‌تواند ابتلا یا عدم ابتلای نمونه به سرطان پروستات را تشخیص دهد. در آزمایش‌هایی که بر روی عملکرد روش پیشنهادی ارائه شده است، برای اپراتور اشتراک، دو تابع

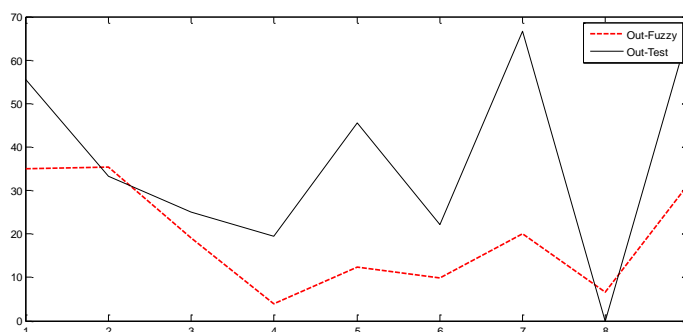
با توجه به مباحث قبلی، بعد از محاسبه احتمال بیماری در مورد هر یک از نمونه‌های موجود، با استفاده از مقدار حد‌آستانه تعیین شده، روش پیشنهادی می‌تواند ابتلا یا عدم ابتلای هر نمونه به سرطان پروستات را تشخیص دهد. همان‌طور که گفته شد در این حالت که به عنوان بهترین حالت برای روش پیشنهادی، مشخص شده است، برای حد‌آستانه مقدار ۵۰ در نظر گرفته شده است. با داشتن احتمال ابتلا به



شکل ۲: تشخیص ابتلا به بیماری در مورد ۴۰۲ نمونه موجود در داده آموزشی

می‌گیرد. به این صورت که احتمال بیماری در هر یک از ۱۰ خوشه محاسبه می‌شود. سپس سیستم فازی با استفاده از داده آموزشی، آموزش داده می‌شود و سپس در مورد هر یک از ۱۰ خوشه ایجاد شده از داده آزمایشی، نمونه‌های موجود در آن خوشه‌ها، جهت پیش‌بینی احتمال بیماری به سیستم طراحی شده داده می‌شود. میانگین مقادیر پیش‌بینی شده برای هر یک از نمونه‌ها در هر خوشه در این مرحله محاسبه می‌شود و این مقدار میانگین احتمال در هر خوشه با احتمال واقعی بیماری در آن خوشه محاسبه می‌شود. نتایج این بررسی در شکل ۳ ارائه شده است که این نتایج نشان‌دهنده آن است که روند تغییرات ایجاد شده روی میانگین احتمال‌های پیش‌بینی شده در خوشه‌ها و روند تغییرات روی میانگین واقعی خوشه‌ها یکسان است.

همان‌طور که گفته شد در شکل ۲ تشخیص ابتلا یا عدم ابتلا به بیماری با توجه به احتمال بیماری و حدآستانه ۵۰ و همچنین خروجی واقعی داده، نشان داده شده است. در این شکل، خروجی صفر نشان‌دهنده تشخیص سالم بودن برای نمونه و خروجی ۱ به معنی تشخیص مبتلا بودن نمونه به سرطان پروستات است. در گام بعدی آزمایش، به منظور بررسی بیشتر عملکرد روش پیشنهادی، ارزیابی دیگری را روی یکی از حالت‌های خوب آزمایش، انجام می‌دهیم. این حالت، حالتی است که از روش خوشه‌بندی k-means با تعداد ۱۰ خوشه و حد آستانه ۵۰ و اپراتور اشتراک مینیمم استفاده کرده است. هدف این است که با استفاده از روش خوشه‌بندی k-means و تعداد ۱۰ خوشه، داده آموزشی و داده آزمایشی را خوشه‌بندی کنیم. در گام بعدی هر یک از خوشه‌های ایجاد شده بر روی داده آزمایشی، مورد بررسی قرار



شکل ۳: مقایسه احتمال بیماری بین خروجی سیستم فازی با احتمال به دست آمده از داده تست

این صورت که اگر این احتمال بیشتر از حد آستانه باشد، نمونه به عنوان مبتلا به سرطان پروستات و در غیر این صورت به عنوان فرد سالم تشخیص داده می‌شود. در ارزیابی روش پیشنهادی از روش‌های متفاوت خوشه‌بندی، تعداد خوشه‌های متفاوت، حد آستانه دلخواه و اپراتورهای اشتراک متفاوت استفاده شده است. در حالتی که از روش خوشه‌بندی  $k$ -medoids با تعداد ۲۰ خوشه و حد آستانه ۵۰ و اپراتور اشتراک مینیمم استفاده شده است، روش پیشنهادی به بهترین عملکرد خود می‌رسد. به این صورت که در این حالت، معیار صحت برای داده آموزشی به مقدار ۷۳٪ رسیده است.

### نتیجه‌گیری

در این مقاله به مساله پیش بینی احتمال ابتلا به بیماری سرطان پروستات و تشخیص این بیماری در شاغلین پرداخته می‌شود. از نتایج حاصل از این تحقیق می‌توان به منظور پیش‌بینی احتمال ابتلای شاغلین جوان به سرطان پروستات با توجه به مشاغل مختلفی که به آن‌ها اشتغال دارند و میزان مواجهه با عوامل خطرزا در هر یک از مشاغل، مصرف یا عدم مصرف سیگار و قلیان، نوع تغذیه و همچنین سابقه خانوادگی سرطان و سرطان پروستات استفاده کرد که در صورت بالا بودن این احتمال، به عنوان فرد در معرض خطر شناخته شده و در مواردی حتی شاید نیاز به تغییر شغل وجود داشته باشد.

می‌توان روش ارائه شده در این مطالعه را بر روی مجموعه داده‌های بیماری‌های دیگر نیز پیاده‌سازی کرد و به منظور تشخیص بیماری و همچنین محاسبه احتمال بیماری مورد استفاده قرار داد. همچنین بررسی پارامترهای دیگر تاثیرگذار بر سرطان پروستات مثل بررسی عوامل خطرزای دیگری در محیط کاری، می‌تواند به بالابردن دقت سیستم کمک کند. پیشنهاد دیگری که برای کارهای آینده وجود دارد این است که مجموعه داده مورد استفاده دارای نمونه‌های خیلی زیادتری باشد تا بتوان پیش‌بینی دقیق‌تری را بر اساس آن انجام داد.

باید به این نکته نیز توجه داشت که به دلیل این که در این مطالعه به مساله پیش‌بینی احتمال پرداخته می‌شود، لازم است مجموعه داده مورد استفاده برای این مطالعه، دارای نمونه‌های زیادی باشد تا پیش‌بینی احتمال به صورت دقیق‌تر انجام گیرد. در صورتی که در مطالعه حاضر، مجموعه داده مورد استفاده دارای تعداد محدودی نمونه بود که می‌توان تاثیر این امر را بر روی دقت روش پیشنهادی ملاحظه نمود.

### بحث

در این مطالعه، مجموعه داده‌ای حاوی سوابق شغلی، میزان مواجهه با عوامل خطرزای محیط کار، نوع تغذیه و همچنین اطلاعات دیگری از شاغلینی که برخی از آن‌ها سالم و برخی مبتلا به سرطان پروستات بوده‌اند مورد بررسی قرار گرفته است. ابتدا بر روی این مجموعه داده، پیش‌پردازش انجام گرفته و سپس به دو بخش آموزشی و آزمایشی تقسیم شده است. به این صورت که ۸۰٪ از داده‌ها به عنوان داده آموزشی و ۲۰٪ از آن به عنوان داده آزمایشی در نظر گرفته شده است. عملیات خوشه‌بندی بر روی داده آموزشی انجام گرفته است. سپس یک سیستم فازی با استفاده از خوشه‌ها ساخته شده است به این صورت که تعداد قواعد این سیستم برابر با تعداد خوشه‌ها می‌باشد. در واقع هر خوشه قسمت مقدم این قواعد را تشکیل می‌دهد. درصد بیماران در هر خوشه محاسبه شده و به عنوان احتمال بیماری در آن خوشه در نظر گرفته شده است. در مرحله بعدی برای هر فرد جدیدی که به سیستم وارد می‌شود، درجه تعلق آن فرد به هر یک از خوشه‌ها محاسبه شده است. با توجه به احتمال بیماری در هر خوشه و درجه تعلق نرمالایز شده فرد به هر خوشه، خروجی این سیستم برای هر نمونه‌ای که به آن وارد می‌شود محاسبه شده است. این خروجی، میزان احتمال ابتلای نمونه جدید به سرطان پروستات است. به منظور تشخیص سالم یا مبتلا بودن هر نمونه جدید، حد آستانه‌ای برای مقدار پیش‌بینی شده احتمال در نظر گرفته شده است. به



**References:**

- 1- Larrañaga N, Galceran J, Ardanaz E, Franch P, Navarro C, Sánchez MJ, et al. *Prostate cancer incidence trends in Spain before and during the prostate-specific antigen era: impact on mortality*. Ann Oncology. 2010; 21(3): 83-9.
- 2- Peter DB, Danny RY, Lauren JK. *International epidemiology of prostate cancer: Geographical distribution and secular trends*. Molecular Nutrit Food Res 2009; 53(2): 171-84.
- 3- Abrahamsson PA. *Prostate Cancer: To Treat or Not to Treat?* Euro Urology Supplements 2009; 8(5): 418-23.
- 4- Azizi F, Janghorbani M, Hatami H. *Epidemiology and control of common diseases in Iran*. 3<sup>rd</sup> ed, Tehran: Khosravi, 2011; p.277-9. [Persian]
- 5- Hosseini M, Jahani Y, Mahmoodi M, Eshraghian MR, Yahyapoor Y, Koshtkar A. *Assessment of risk factors for prostate cancer in Mazandaran Province*. J Gorgan Med Sci 2008; 10(3): 58-64. [Persian]
- 6- Jabel Ameli P, Bahadori M. *Early detection of prostatic cancers*. J Medical Council of Islamic Republic of Iran 1999; 3(17): 231-8.
- 7- Turini M, Redaelli A, Gramegna P, Radice D. *Quality of life and economic considerations in the management of prostate cancer*. Pharmacoeconomics 2003; 21(8): 527-41.

## *Diagnosis of prostate cancer and predicting the probability of suffering the disease in workers*

*Fatemeh Sarbazaqdaee(Msc)<sup>1</sup>, Vali Derhami(Phd)<sup>\*2</sup>, Mohamad Javad Zaresakhvidi (Phd)<sup>3</sup>  
Mehrdad Mostaghaci(Phd)<sup>4</sup>, Raziye Soltani GerdFaramarzi (Msc)<sup>5</sup>, Maryam Musavi(Msc)*

<sup>1</sup> Department of computer engineering, Science and Arts University, Yazd, Iran

<sup>2</sup> Department of computer engineering, Yazd University, Yazd, Iran

<sup>3</sup> Department of Occupational Health, Faculty of Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

<sup>4</sup> Occupational & Environmental Medicine Specialist, Isfahan University of Medical Sciences, Isfahan, Iran

<sup>5</sup> Industrial Diseases Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

<sup>6</sup> Department of Medical Education, Iran University of Medical Sciences, Tehran, Iran

**Received:** 10 Feb 2015

**Accepted:** 08 Nov 2015

### **Abstract**

**Introduction:** Diagnosis of various diseases in medicine is one of the area's most widely used data mining in recent years and many researches have been done about it. In this study, the diagnosis of prostate cancer using fuzzy system was assessed. The goal was to diagnose the prostate cancer and to predict the possibility of suffering from the disease.

**Methods:** In the proposed method, at first, based on available dataset, pre-processing and clustering operations were carried out. Then a zero-order Sugeno fuzzy system was designed for prediction. Each cluster, as the first item of a fuzzy rule, was considered and out of a rule, percentage of disease possibility in each cluster was considered. For each new sample, the membership degree to the each cluster was computed and then by combining outputs from each rule, possibility of disease in the sample was predicted. Finally, by having possibility and threshold for possibility, having or not having the disease for desired sample was diagnosed.

**Results:** The results showed that the system has good accuracy in predicting the possibility of disease.

**Conclusion:** The results of this study can be used to predict the risk of prostate cancer in young workers according to different jobs they are employed and the amount of exposure to risk factors in each job. If this possibility is high, they are known as the person at risk and in some cases there may need to change higher jobs.

**Keywords:** Prediction; Prostate cancer; Fuzzy systems

**This paper should be cited as:**

Sarbazaqdaee F, Derhami V, Zaresakhvidi MJ, Mostaghaci M, Soltani GerdFaramarzi R. *Diagnosis of prostate cancer and predicting the probability of suffering the disease in workers*. Occupational Medicine Quarterly Journal 2016; 8(2): 62-71.

**\* Corresponding Author: Tel: +98 3531232365, Email: vderhami@yazd.ac.ir**