

بررسی اثر برخی موتاسیون ژنی در بیماری اندومتريوز با روش‌های داده کاوی

لیلا بهرامی^۱، دکتر بهرام صادقی بی‌غم^۲، دکتر کوروش کمالی^۳

نویسنده‌ی مسوول: گروه کامپیوتر، دانشکده‌ی کامپیوتر، دانشگاه تحصیلات تکمیلی علوم پایه زنجان، زنجان b_sadeghi_b@iasbs.ac.ir

دریافت: ۹۴/۵/۴ پذیرش: ۹۴/۱۰/۲۱

چکیده

زمینه و هدف: اندومتريوز یکی از بیماری‌های شایع در زنان ایرانی است که منجر به ناباروری و کم باروری در آن‌ها می‌شود، بنابراین شناسایی علل ژنتیکی بیماری در زنان می‌تواند به درمان آن کمک کند. با توجه به اینکه توارث ژنتیکی به‌عنوان یکی از عوامل مهم خطر ساز بیماری اندومتريوز می‌باشد لزوم تشخیص درست و زودهنگام این بیماری با توجه به عوارضی که در پی دارد دو چندان می‌شود. از این رو مطالعه‌ی حاضر جهت تعیین اثر برخی از موتاسیون‌های ژنی در بیماری اندومتريوز با روش‌های داده‌کاوی انجام گردید.

روش بررسی: در این پژوهش اثر ۹ ژن دخیل در بیماری اندومتريوز بر روی افراد مبتلا به اندومتريوز و افراد سالم (جمعا ۲۶۰ نمونه) بررسی شد. داده‌های مورد استفاده در این پژوهش، از مرکز درمان فوق تخصصی ناباروری پژوهشگاه فن آوری‌های نوین علوم زیستی جهاد دانشگاهی ابن سینا دریافت شد. مراحل اجرای پژوهش مبتنی بر فرایند استاندارد کریسپ در داده‌کاوی بود. مدل‌سازی به کمک چهار الگوریتم داده کاوی و در نرم افزار $\sqrt{v}$ Weka (New Zealand, University of Waikato) پیاده سازی شد.

یافته‌ها: با مقایسه‌ی چهار الگوریتم، مشاهده شد که مدل $K\text{-Star}$ بالاترین میزان دقت را دارا می‌باشد و فیلتر اعمال شده بر روی ویژگی‌ها سبب کاهش درصد مدل‌ها شد ولی بر روی مدل MLP تاثیر مثبتی داشت. همچنین پایین‌ترین درصد مربوط به روش $Naive\ Bayes$ بود. **نتیجه‌گیری:** مدل $K\text{-Star}$ بدون اعمال هیچگونه فیلتر با نتیجه‌ای برابر ۹۴/۶۱ بالاترین میزان دقت را در تشخیص زود هنگام بیماری اندومتريوز دارا می‌باشد.

واژگان کلیدی: داده‌کاوی، اندومتريوز، طبقه‌بندی، $Weka$ ، MLP ، $K\text{-star}$ ، $Naive\ bayes$ ، $J48$

مقدمه

به‌تحریکات هورمونی تخمدان پاسخ دهد، تعریف می‌شود (۱). این بافت ممکن است در لگن یا خارج از لگن وجود داشته باشد. این بیماری به‌طور شایع در لگن، تخمدان‌ها، کلدوساک خلفی، لیگامان رحمی، کلدوساک لگنی و تیغه‌ی رکتوواژینال مشاهده شده و در ۸ تا ۱۵ درصد زنان در سنین باروری یافت می‌شود (۲).

اندومتريوز بیماری است که طی آن بافت داخلی رحم که اندومتر خوانده می‌شود به اشتباه در جایی خارج از رحم قرار می‌گیرد. اندومتريوز یک بیماری خوش خیم سیستم ژنیتال زنان است و به‌صورت وجود بافتی شبیه بافت اندومتر بیرون از حفره رحم تعریف می‌گردد. به عبارت دیگر اندومتريوز به‌صورت رشد بافت اندومتر خارج از حفره رحمی که می‌تواند

۱- دانشجوی کارشناسی ارشد کامپیوتر، دانشگاه آزاد اسلامی واحد زنجان، زنجان

۲- دکترای علوم کامپیوتر، استادیار دانشکده‌ی کامپیوتر، دانشگاه تحصیلات تکمیلی علوم پایه زنجان، زنجان

۳- دکترای اپیدمیولوژی، استادیار گروه اپیدمیولوژی، مرکز تحقیقات بیماری‌های متابولیک، دانشگاه علوم پزشکی زنجان، زنجان

در اندومتريوز بافت اندومتر تغییر جا داده، اما همچنان به فعالیت خود ادامه می‌دهد و در هر دوره‌ی قاعدگی، این بافت ضخیم می‌شود و خونریزی می‌کند. در این بیماری، بافت‌های اطراف ممکن است تحریک شوند و در نهایت موجب زخمی شدن بافت و چسبندگی گردد. اندومتريوز می‌تواند باعث درد شدید، مخصوصاً در دوران عادت ماهانه گردد. مشکلات بارداری نیز ممکن است به وجود آورد (۳). بافت اندومتر ایجاد شده باعث علائمی چون درد، التهاب، چسبندگی و انسداد در این مناطق می‌شود که پس از گذشت چند سال باعث نازایی در خانم‌ها می‌شود. ۳۰ تا ۵۰ درصد زنانی که دچار اندومتريوز می‌باشند احتمال وقوع نازایی در آن‌ها وجود دارد. چسبندگی در رحم و لوله‌های تخمدان، التهاب ناشی از حضور نابجای بافت اندومتر، افزایش فاکتورهای التهابی مانند پروستاگلندین‌ها، اینترلوکین و تومور نکروز فاکتورها در فضای اطراف تخمک و اسپرم، همچنین تغییرات هورمونی ناشی از ستنز بالای استرادیول در بافت محیطی و کاهش اثر پروژسترون در فاز لوتئال سیکل ماهیانه، همگی از دلایل ایجاد نازایی در این بیماری می‌باشد (۴).

در سال ۱۹۹۶ انجمن پزشکی تولید مثل آمریکایی براساس مکان، اندازه و عمق بافت اندومتريوز این بیماری را به چهار مرحله تقسیم‌بندی کردند که در آن بیماری از حالت خفیف تا شدید تعریف گردید (۵). در حالت شدید (مرحله ۴) انسداد و چسبندگی در لوله‌های تخمدان و محوطه لگنی باعث ایجاد نازایی شدید در این بیماران شده است. به‌علت سیر خاموش و پیش‌رونده بیماری اندومتريوز، این بیماری در اغلب موارد در مراحل اولیه قابل تشخیص نبوده و عموماً زمانی تشخیص داده می‌شود که در حدود ۶ الی ۸ سال از شروع آن گذشته باشد (۶). متأسفانه روش‌های تشخیصی اولتراسونیک مانند سونوگرافی نیز در تشخیص این بیماری زیاد کارآمد نیست. در ضمن هیچ تست آزمایشگاهی که به‌طور دقیق وجود این بیماری را مشخص کند، نیز گزارش

نشده است. تنها راه تشخیص صحیح اندومتريوز در حال حاضر عمل جراحی لاپاراسکوپ می‌باشد این عمل نیاز به بی‌هوشی کامل دارد و پس از آن بیمار می‌بایست دوره نقاهت دردناکی را سپری کند. همچنین انجام عمل لاپاراسکوپ برای تشخیص و درمان اندومتريوز از نظر مالی برای بیمار مقرون به صرفه نیست (۷). نکته‌ی قابل ملاحظه در این بیماری این است که معمولاً کنترل و درمان این بیماری آسان است. فقط مشکل تشخیص به‌موقع این بیماری است و معمولاً زمانی تشخیص داده می‌شود که اثرات نامطلوب خود را همچون نازایی در شخص ایجاد کرده باشد. سال‌هاست که محققان به دنبال روش‌های غیرتهاجمی برای تشخیص این بیماری می‌باشند تا بتوانند بیماری را با توجه به فاکتورهای درگیر در آن، در مراحل اولیه شناسایی نمایند (۸). مطالعات همراهی در کل ژنوم (Genome-wide association studies (GWASs) در شناسایی اثرات متوسط واریاسیون‌های ژنی برای بیماری‌های پیچیده به اثبات موفقی رسیده‌اند و تا به امروز، هشتمین مطالعات مرتبط با ژنوم گسترده و تکرار مطالعات بر روی اندومتريوز از چند جمعیت منتشر شده است (۹). داده‌کاوی، به معنای استخراج اطلاعات نهفته، بازشناخت روابط و الگوهای پنهان و به‌طور کلی کشف دانش سودمند از داده با حجم بالا است. استخراج قواعد طبقه‌بندی، نوعی داده‌کاوی است که در آن دانشی به شکل چندین قانون ساده و فهم‌پذیر از داده کشف شده و در آینده برای تصمیم‌گیری و پیش‌گویی به‌کار برده می‌شود، از مهم‌ترین زمینه‌های کاربردی، استخراج قواعد طبقه‌بندی در حیطه پزشکی می‌باشد (۱۰). استخراج دانش از میان حجم انبوه داده‌های مرتبط با سوابق بیماری و پرونده‌های پزشکی افراد، با استفاده از فرآیند داده‌کاوی می‌تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسریع بیماری‌ها شده و اطلاعات ارزشمندی را به‌منظور شناسایی علل رخداد بیماری‌ها، تشخیص، پیش‌بینی و درمان بیماری‌ها با توجه به عوامل محیطی حاکم در اختیار

پیشنهادی مجموع ۱۳ ویژگی از مجموعه داده‌ها استخراج شده و آزمایش‌های گوناگون بر روی برای یک مجموعه داده ژن ۲۶۰ تایی از افراد مختلف در زمینه اندومتریوز، تست و اجرا شده است.

شناخت داده: در این بخش به تبیین مجموعه داده پرداخته می‌شود. شناخت داده شامل موارد زیر است: اولاً باید تک تک متغیرها مورد تعریف و توصیف واقع شوند. مجموعه داده بیماری اندومتریوز دارای تعدادی متغیر یا ویژگی و تعداد زیادی تراکنش است. هر یک از این ویژگی‌ها اشاره به یک موضوع در مورد بیماری اندومتریوز بیماران دارد. هم چنین تراکنش‌ها اشاره به بیماران دارد. در این بخش باید به توصیف مقادیر هر یک از این متغیرها نیز پرداخته شود. همچنین باید مشخص شود که هر کدام از این ویژگی‌ها از نوع کیفی هستند یا کمی. البته باید تعداد و نوع مقادیر هر یک از ویژگی‌ها نیز توصیف شود. این فاز می‌تواند به عنوان مبنایی برای فاز بعدی (یعنی شناخت داده و آماده‌سازی داده‌ها) جمعاً حدود ۶۰ درصد فعالیت‌های داده‌کاوی را در بردارد.

آماده‌سازی داده‌ها: این بخش، یکی از فازهای مهم در داده‌کاوی است. آماده‌سازی داده‌ها باید به روش مناسبی انجام شود. یک تعادلی برای آماده‌سازی و تمیز کردن داده‌ها باید وجود داشته باشد. به عبارت دیگر کوتاهی در این بخش و یا آماده‌سازی بیش از حد داده‌ها می‌تواند به مدل ساخته شده توسط داده کاوی صدمه بزند. زیرا در صورتی که داده‌ها تمیز نباشند دقت مدل کاهش می‌یابد و در صورت تمیز کردن بیش از حد داده‌ها، مدل ساخته شده با دنیای واقعیت تناسب ندارد. برخی از فعالیت‌هایی که در این بخش صورت می‌گیرد شامل موارد زیر است:

یکپارچه سازی داده: در این مرحله می‌بایست داده‌ها جمع‌آوری شده و انتخاب و یکپارچه شوند. **تمیز کردن داده‌ها:** در این مرحله به از بین بردن مقادیر احتمالی

متخصصان و دست‌اندرکاران حوزه سلامت قرار دهد. نتیجه این مساله، افزایش عمر و ایجاد آرامش برای افراد جامعه است (۱۱).

در این پژوهش به کمک داده‌های برگرفته از ژن‌های افراد گوناگون، با بهره‌گیری از الگوریتم‌ها و نرم‌افزارهای داده‌کاوی به پیش‌بینی بیماری اندومتریوز پرداخته می‌شود و با توجه به این‌که اندومتریوز یک بیماری ژنتیکی است اثر ۹ ژن را نشان خواهیم داد.

روش بررسی

این پژوهش اثر برخی از موتاسیون‌های ژنی دخیل در بیماری اندومتریوز را بر روی دو دسته شناخته شده سالم و بیمار مورد بررسی قرار داده است. داده‌های مورد استفاده در این پژوهش، مربوط به داده‌های بیماران اندومتریوز است. که از مرکز درمان فوق تخصصی ناباروری پژوهشگاه فناوری‌های نوین علوم زیستی جهاد دانشگاه ابن سینا دریافت شده است. در واقع اثر ۹ ژن در این پژوهش در ارتباط با بیماری اندومتریوز در قالب ۲۶۰ نمونه بررسی می‌شوند. مساله اصلی در این پژوهش، تشخیص درست و زودهنگام اندومتریوز برای درمان با کمک الگوریتم‌های داده‌کاوی می‌باشد. الگوریتم‌های داده‌کاوی به کمک نرم‌افزارهای متداول در داده‌کاوی مثل نرم افزار Weka پیاده‌سازی می‌شوند. مراحل اجرای پژوهش مبتنی بر فرایند استاندارد کریسپ در داده‌کاوی است. این مراحل شامل موارد زیر است: شناخت کسب و کار و گردآوری داده، شناخت داده، آماده‌سازی داده، مدل‌سازی، ارزیابی و بهینه‌سازی

شناخت کسب کار و گردآوری داده: همان‌طور که از این پژوهش مشخص است، با استفاده از داده‌کاوی می‌توان به شناسایی الگوهای توصیفی و پیش‌گویی کننده در مجموعه داده بیماران اندومتریوز پرداخت. در ابتدا جهت انجام پژوهش به گردآوری و توصیف داده‌ها پرداخته می‌شود. در این روش

مفقود در مجموعه داده پرداخته می شود. همچنین در صورتی که در مجموعه داده پارازیت وجود داشته باشد، بایستی برطرف گردد. همچنین در صورتی که در مجموعه داده ناسازگاری وجود داشته باشد، این ناسازگاری باید از بین برود.

تبدیل داده: در صورتی که نیاز به نرمال سازی باشد، باید عملیات نرمال سازی بر روی مجموعه داده صورت پذیرد. همچنین می توان از عملیات گسسته سازی بر روی مجموعه داده در صورت لزوم استفاده نمود. همچنین می توان به ساخت یک مجموعه ویژگی جدید نیز در مجموعه داده پرداخت.

کاهش داده: در این بخش می توان به کاهش تعداد متغیرها و تعداد تراکنش ها در صورت نیاز پرداخت. همچنین می توان به بالانس کردن ویژگی هدف در مجموعه داده پرداخت.

در این پژوهش با توجه به اینکه ویژگی های مربوط به هر ژن در مجموعه داده مختلفی قرار داشت، در ابتدا ویژگی های هر ژن مربوط به هر فرد بر اساس ژنوتیپ و گروه برچسب گذاری گردید. یکسری عملیات پیش پردازش اولیه از بین بردن مقادیر مفقود و غیره انجام شد و پس از آن اطلاعات ژنی هر فرد از مجموعه داده های متفاوت بر اساس گروه (سالم یا بیمار بودن فرد)، سن، وزن، قد و شاخص توده بدنی هر فرد شناسایی و استخراج گردید و در یک مجموعه داده کلی با ۱۵ ویژگی (کیفی، کمی) تجمیع شد. سپس عملیات پیش پردازش بر روی این مجموعه داده کلی نیز انجام گرفت و تعداد نمونه ها به ۲۶۰ نفر کاهش یافت همچنین سه ویژگی

(متغیر) به دلیل اینکه تاثیری در نتیجه ندارند حذف شدند. در نهایت تعداد ۱۳ متغیر در این پژوهش مورد بررسی قرار گرفت که شامل اطلاعات ۹ ژن، وزن، قد، سن و گروه (سالم، بیمار) می باشد. پس از آن این اطلاعات به فرمت مورد قبول نرم افزار Weka (ARFF) تبدیل گردید. و وارد نرم افزار گردید تا عملیات مدل سازی انجام گیرد.

مدل سازی: در این بخش مدل پیشنهادی پژوهش ارایه می شود. این مدل بر مبنای تکنیک های داده کاوی است. در این خصوص از سه تکنیک اصلی داده کاوی استفاده شده است که عبارتند از: ۱. انتخاب ویژگی ها. ۲. کشف آنومالی. ۳. طبقه بندی.

انتخاب ویژگی ها: در ادامه به استفاده از ابزار انتخاب ویژگی ها برای تعیین ویژگی های مهم در مجموعه داده پرداخته می شود. در تعیین ویژگی های مهم، تعداد ۱۳ ویژگی به دست آمد که در جدول (۱) مشاهده می شود.

کشف آنومالی: آنومالی به مجموعه اعمالی گفته می شود که با اجرای آن ها در پایگاه داده، باعث ایجاد مشکل یا افزون کاری می شود. سپس در این پژوهش نقاط آنومالی حذف گردید. به عنوان مثال ۱۴ رکورد که در یک مجموعه داده به عنوان فرد سالم و در مجموعه دیگر بیمار بودند حذف گردیدند.

طبقه بندی: در این مرحله ویژگی های مورد نظر طبقه بندی می شوند. چهار مدل طبقه بندی را بر روی همه ویژگی های ژنی در مجموعه کلی بدون اعمال فیلتر و با فیلتر به کار برده می شود و نتایج حاصل را مقایسه نموده مدل طبقه بندی که درصد بالاتری دارد انتخاب می شود.

جدول ۱: اطلاعات متغیرهای مورد بررسی در پژوهش

نام ویژگی	شرح	نوع	برچسب ژنوتیپ
TNFa	ترکیب ژنتیکی فرد	اسمی	TT=0 TC=1 CC=2
VEGF	ترکیب ژنتیکی فرد	اسمی	GG=0 GC=1 CC=2
XRCC4-Condon247	ترکیب ژنتیکی فرد	اسمی	CC=0 CA=1 AA=2
XRCC4-intron3	ترکیب ژنتیکی فرد	اسمی	II=0 ID=1 DD=2
XRCC4-Promoter	ترکیب ژنتیکی فرد	اسمی	TT=0 TG=1 GG=2
TNF 238	ترکیب ژنتیکی فرد	اسمی	GG=0 GA=1 AA=2
TNF 308	ترکیب ژنتیکی فرد	اسمی	GG=0 GA=1 AA=2
TNF 857	ترکیب ژنتیکی فرد	اسمی	CC=0 CT=1 TT=2
TNF 863	ترکیب ژنتیکی فرد	اسمی	CC=0 CA=1 AA=2
Age	سن	عددی	-
Group	بیمار یا سالم	عددی	False=0 True=1
Weight	وزن	عددی	-
Height	قد	عددی	-

یافته‌ها

در این قسمت از مدل طبقه‌بندی بیز الگوریتم Naïve Bayes، شبکه‌ی عصبی مصنوعی الگوریتم پرسپترون چند لایه (MLP)، مدل درختی الگوریتم J48 و الگوریتم K-Star از مدل طبقه‌بندی Lazy و بدون اعمال هیچگونه فیلتر در نرم افزار Weka استفاده شد. سپس فیلتر Discretize بر روی مقادیر ویژگی وزن و سن و قد اعمال

گردید که نتایج نیز نشان داده شد. فیلتر Discretize مقادیر یک صفت پیوسته را به تعداد دلخواه بازه گسسته تبدیل می‌کند. در این بررسی با اعمال فیلتر مقادیر سه ویژگی به پنج بازه تقسیم شدند.

مدل‌سازی با استفاده از الگوریتم‌های طبقه‌بندی؛ در ابتدا از روش طبقه‌بندی با استفاده از الگوریتم Naïve Bayes استفاده شد. مشاهده گردید، نتیجه برابر با ۶۳/۴۶ درصد است

و پس از اعمال فیلتر نیز مشخص شد که نتیجه برابر ۶۲/۶۹ درصد است. پس از آن از روش طبقه‌بندی با استفاده از الگوریتم MLP استفاده شد که نتیجه برابر با ۹۵ درصد بود و پس از اعمال فیلتر نتیجه برابر ۹۸/۸۴ درصد شد. سپس از روش طبقه‌بندی با استفاده از الگوریتم J48 استفاده گردید. نتیجه برابر با ۷۸/۴۶ درصد بود که پس از

اعمال فیلتر برابر ۷۵/۷۶ درصد شد و در نهایت در این قسمت از روش طبقه‌بندی با استفاده از الگوریتم K-Star استفاده گردید. همان گونه که در شکل (۲) مشاهده می‌شود، نتیجه برابر با ۹۹/۶۱۵۴ درصد است و در شکل (۳) نیز پس از اعمال فیلتر مشاهده می‌شود نتیجه برابر ۹۸/۰۷۶۹ درصد است.

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	259	99.6154 %
Incorrectly Classified Instances	1	0.3846 %

شکل ۲: نتایج حاصل از اجرای الگوریتم K-Star بر روی نمونه‌ها و بدون اعمال فیلتر و در نرم‌افزار Weka

=== Evaluation on training set ===

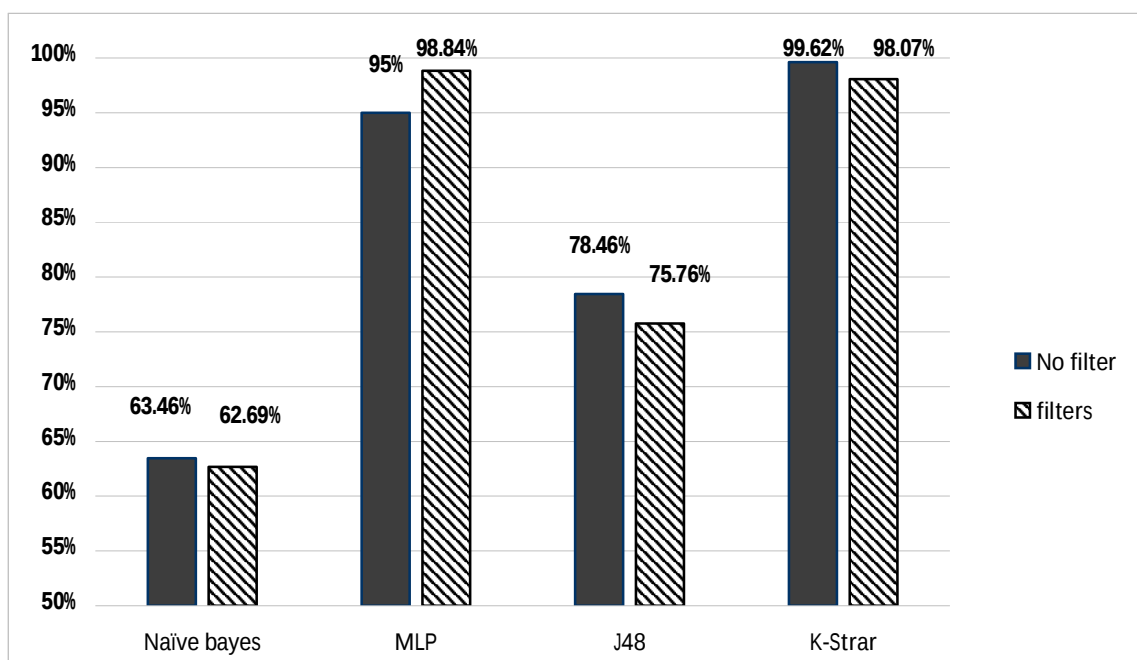
=== Summary ===

Correctly Classified Instances	255	98.0769 %
Incorrectly Classified Instances	5	1.9231 %

شکل ۳: نتایج حاصل از اجرای الگوریتم K-Star بر روی نمونه‌ها با فیلتر Discretize و در نرم‌افزار Weka

با مقایسه‌ی چهار الگوریتم اعمال شده بر روی ویژگی‌ها، و در نرم‌افزار Weka مشاهده شد که مدل K-Star بالاترین میزان دقت را دارا است. و فیلتر اعمال شده بر روی ویژگی‌ها سبب کاهش درصد مدل‌ها

شده است ولی بر روی مدل‌ها MLP تاثیر مثبتی داشته است. همچنین پایین‌ترین درصد مربوط به روش Naïve Bayes می‌باشد. مقایسه‌ی این چهار مدل در نمودار (۱) مشاهده می‌شود.



نمودار ۱: مقایسه چهار مدل طبقه‌بندی بر روی ویژگی‌های اندومتریوز

بحث

اندومتریوز بیماری پیچیده‌ای است که در نتیجه‌ی تأثیر متقابل ژنتیک و فاکتورهای محیطی به وجود می‌آید. تحقیقات متعدد نشان می‌دهند که احتمال ابتلا به این بیماری در زنانی که سابقه‌ی اندومتریوز در خانواده‌ی آن‌ها وجود داشته، بیشتر است (۱۲ و ۱۳) متأسفانه روش‌های تشخیصی اولتراسونیک مانند سونوگرافی نیز در تشخیص این بیماری زیاد کارآمد نیست. در ضمن هیچ فاکتور آزمایشگاهی که به‌طور دقیق وجود این را مشخص کند نیز گزارش نشده است. در حال حاضر تنها راه تشخیص صحیح اندومتریوز عمل جراحی لاپاراسکوپی می‌باشد. از این طریق بافت اندومتریوز به‌طور مستقیم قابل رویت است. این عمل نیاز به بی‌هوشی کامل دارد و پس از آن بیمار می‌بایست دوره نقاهت دردناکی را سپری کند. ولی انجام عمل لاپاراسکوپی برای تشخیص و درمان اندومتریوز از نظر مالی

برای بیمار مقرون به صرفه نیست. گفتنی است برای کنترل و درمان این بیماری از داروهایی مانند پروژستین و قرص جلوگیری از بارداری استفاده می‌شود (۷). این داروها ساخت استروژن را از تخمدان‌ها کنترل کرده و باعث کاهش سایز بافت اندومتریوز می‌شوند. نکته‌ی قابل ملاحظه در این بیماری این است که معمولاً کنترل و درمان این بیماری آسان است. فقط مشکل تشخیص به موقع این بیماری است و معمولاً زمانی تشخیص داده می‌شود که اثرات نامطلوب خود را چون نابری در شخص ایجاد کرده باشد.

با توجه به اینکه توارث ژنتیکی به‌عنوان یکی از عوامل مهم خطر ساز بیماری اندومتریوز می‌باشد لزوم تشخیص درست و زودهنگام این بیماری با توجه به عوارضی که در پی دارد دو چندان می‌شود. از این رو مطالعه‌ی حاضر جهت تعیین اثر برخی از موتاسیون‌های ژنی در بیماری اندومتریوز با روش‌های داده‌کاوی انجام گردید، لازم به ذکر است این مدل

(۱۴/۹ درصد) می‌باشد (۱۸). لیکن در این تحقیق (ما) با نگاهی به کسب و کار (در اینجا بیماری اندومتریوز) برای اولین بار با آرایه‌ی یک مدل ترکیبی در تشخیص بیماری اندومتریوز پرداخته شده است. مدل پیشنهادی دارای اعتبار است. به گونه‌ای که با اجرای هر بخش از مدل به نتایج جدیدی دست خواهیم یافت. این نتایج هم می‌توانند به خودی خود مورد استفاده در کسب و کار قرار گیرند و هم این که نتایج آن می‌تواند در بخش‌های بعدی مدل مورد استفاده قرار گیرد. در حقیقت این بررسی بر روی این ژن‌ها با این روش آنالیز برای اولین بار در ایران انجام شده است.

نتیجه‌گیری

کشف بیماری‌ها توسط کامپیوتر از جمله مباحثی است که امروزه مورد توجه پژوهشگران قرار گرفته است. داده کاوی یکی از بهترین روش‌ها در پیش‌بینی بیماری اندومتریوز می‌باشد. در این پژوهش اثر ۹ ژن دخیل در بیماری اندومتریوز بر روی افراد مبتلا به اندومتریوز و افراد سالم بررسی گردید و روش‌های Naïve bayes, MLP, J48, K-Star بر روی ژن افراد پیاده‌سازی شد. در مقایسه‌ی روش‌های فوق مدل K-Star بدون اعمال هیچگونه فیلتر با نتیجه‌ای برابر ۹۹/۶۱ بالاترین میزان دقت را در تشخیص زود هنگام بیماری اندومتریوز دارا می‌باشد. امید است که در بلند مدت سیستم‌های کاملاً خودکار تشخیص اندومتریوز ایجاد شوند که بتوانند داده‌ها را تجزیه و تحلیل کرده و در تشخیص بیماری در مرحله‌ی اولیه به پزشکان و متخصصان زنان و زایمان و مراکز ناباروری کمک کنند.

تقدیر و تشکر

این پژوهش در قالب پایان نامه‌ی دانشجویی در گروه نرم‌افزار کامپیوتر دانشگاه آزاد اسلامی واحد زنجان انجام شده است. بدین وسیله برخود لازم می‌دانیم از زحمات این عزیزان

برگرفته از نتایج تحقیقات قبلی است. برخی از تحقیقات تنها به استفاده صرف از یک تکنیک بسنده کرده‌اند و یا به بررسی آلل‌های دو ژن TNF α و VEGF پرداخته‌اند. به‌عنوان نمونه پژوهشی با هدف بررسی این که آیا این دو پلی مورفیسم، TNF- α 1031T/C و VEGFp450G/C با حساسیت به اندومتریوز در جمعیت ایرانی همراه است انجام شد که در مجموع، ۱۳۵ زن با تشخیص اندومتریوز و ۱۷۳ زن با هیچ شواهدی از بیماری مورد مطالعه قرار گرفتند (۱۵).

در مطالعه‌ای مشابه اثر ۱۹ ژن که قادر به طبقه‌بندی سرطان اندومتر در دو زیر گروه بافت شناسی اندومتریوید و سروز بودند، شناسایی شد. مدل سازی، تجزیه و تحلیل داده‌ها و داده‌کاوی با استفاده از آرایه BRB ابزار R-program انجام شد (۹). بیماری مورد مطالعه سرطان اندومتر بود که ژن‌های مرتبط تجزیه و تحلیل گردیدند.

در مطالعه‌ای دیگر ژنوتیپ و فرکانس آلل VEGF+405 بین زنان مبتلا به اندومتریوز کنترل و با استفاده از روش PCR-RFLP مقایسه شدند (۱۶). در پژوهشی دیگر برای پرداختن به نقش تعاملات سرطان و استروما، ژن بیان را در هر دو سرطان و استروما با استفاده از نمونه‌های تطبیق سرطان اندومتر آنالیز نموده و ارتباط بین الگوهای بیان ژن و پیش آگهی در سرطان اندومتر را تجزیه و تحلیل کرده‌اند و نتیجه‌ی بررسی این است که پروفایل‌های ژن بیان سرطان و استروما می‌تواند بازگشت سرطان اندومتر و استروما پیش‌بینی کند که ژن بیان در درجه سرطان بستگی ندارد (۱۷). همچنین یکسری از مطالعات از طریق جمع‌آوری پرششنامه به بررسی کیفیت زندگی و یا عوامل برگشت بیماری و عود مجدد آن پرداخته شده است به‌عنوان نمونه در مقاله‌ای به تجزیه و تحلیل بیماران مبتلا به کیست لگن مکرر (اندومتریوز در لگن) پرداخته شده است و گزارش شده است که رفتار القای اتانول برای بیش از ۱۰ دقیقه به ویژه برای یک مورد با یک کیست اندومتر در نظر گرفته شده بسیار موثر از نظر عود

سینا که نهایت همکاری را داشته‌اند تقدیر و تشکر نماییم.

و پژوهشگاه فناوری‌های نوین علوم زیستی جهاد دانشگاه ابن

References

- 1- Williams HE, Barsky S, Storino W. Umbilical endometrioma (silent type). *Arch Dermatol*. 1976; 112: 1435-6.
- 2- Bektas. H. Abdominal wall endometrioma; a 10-year experience and brief review of the literature. *J Surg Res*. 2010; 164: 77-81.
- 3- Sajjadpour M, Tebyan health sector. Available from: <http://www.tebyan.net/newindex>.
- 4- Batt RE. A history of endometriosis. Springer, 2011: 13-38.
- 5- Revised American society for reproductive medicine classification of edometriosis. *Fertil Steril*. 1997; 67: 817.
- 6- Hadfield R. Delay in the diagnosis of endometriosis: a survey of women from the USA and the UK. *Hum Reprod*. 1996; 11: 878-80.
- 7- AAoF E. Diagnosis and treatment of endometriosis. 1999.
- 8- Brosens I. Early stage management of ovarian endometrioma to prevent infertility. *Facts Views Vis Obgyn*. 2013; 5: 309-14.
- 9- Rahmioglu N. Genetic variants underlying risk of endometriosis: insights from meta-analysis of eight genome-wide association and replication datasets. *Hum Reprod Update*. 2014. 20: 702-16.
- 10- Tan PN, Steinbach M, Kumer V. Introduction to data mining (eBook - EnG) 2005.
- 11- Mary K, Obenshain M. Application of data mining techniques to healthcare data. *Infect Control Hosp Epistamol*. 2004: 690.
- 12- Stefansson H, Steinhorsdottir V, Jonsson H, et al. Genetic factors contribute to the risk of developing endometriosis. *Hum Reprod*. 2002; 17: 555-9.
- 13- Zondervan KT, Kennedy SH. The genetic basis of endometriosis. *Curr Opin Obstet Gynecol*. 2001; 13: 309-14.
- 14- Kennedy S. Genetics of endometriosis: a review of the positional cloning approaches. *Semin Reprod Med*. 2003. 21: 111-8.
- 15- Saliminejad K. Association study of the TNF-a 1031T/C and VEGFp450G/C polymorphisms with susceptibility to endometriosis. *Gynecal Endocrinal*. 2013; 29: 974-7.
- 16- Memariani T, Association of vascular endothelial growth factor (VEGF) +405 G>C polymorphism with endometriosis in Iranian population. *J Clin Oncol*. 2008; 26: 4672-8.
- 17- Iguchi Y. Simultaneous analysis of the gene expression profiles of cancer and stromal cells in endometrial cancer. *Genes Chromosomes & Cancer*. 2014; 53: 725-37.
- 18- Wang YF. Mining medical data: a case study of endometriosis. *J Med Syst*. 2013; 37: 9899.

The Effect of some Genetic Mutations on Endometriosis Using Data Mining Techniques

Bahrami L¹, Sadeghi Bigham B², Kamali K³

¹Dept. of Computer, Islamic Azad University, Zanjan Branch, Zanjan, Iran

²Dept. of Computer, Faculty of Computer, University of Basic Sciences, Zanjan, Iran

³Dept. of Epidemiology, Metabolic Disease Research Center, Zanjan University of Medical Sciences, Zanjan, Iran

Corresponding Author: Sadeghi Bigham B, Dept. of Computer, Faculty of Computer, University of Basic Sciences, Zanjan, Iran

E-mail: b_sadeghi_b@iasbs.ac.ir

Received: 26 Jul 2015 **Accepted:** 21 Jan 2016

Background and Objective: Endometriosis is a prevalent disease in women which may lead to infertility or low fertility. Grasping the genetic grounds for the disease may contribute to its treatment because it is presumed that genetic factors predispose to endometriosis risk factors.

Materials and Methods: 9 genes involved in endometriosis in patients suffering from endometriosis and also in healthy individuals (total 260 samples) were examined. The data were obtained from Ibn Sina (Jahad Daneshgahi) Research Center of new Technologies in Biological Sciences Institute. The study incorporated standard process Crisp for data mining. Weka data mining and software modeling were implemented with the aid of four algorithms.

Results: Comparison of four algorithms implied prominent accuracy of K-Star model. Meanwhile, filtering, while reducing the percentage of models, presented a positive impact on the MLP model. The lowest percentage pertained to Naïve Bayes.

Conclusion: K-Star model without any filtering proved to have the highest accuracy in the early diagnosis of endometriosis.

Keywords: Data Mining, Endometriosis, Classification, Weka, MLP, K-star, Naïve bayes, J48