مجله دانشگاه علوم پزشکی شهرکرد/ دوره ۱۵، شماره ۵/ آذر و دی ۱۳۹۲/ ۵۶-۴۷

مقاله پژوهشی

مدل پیش‌بینی بیماری عروق کرونر با استفاده از شبکه‌های عصبی و گزینش متغیر مبتنی بر درخت رگرسیون و طبقه بندی

عیسی محمودی[۱]، دکتر رضا عسکری مقدم[۲]*، دکتر محمد هادی معظم[۳]، دکتر سعید صادقیان[۴]

[۱] دانشجوی کارشناسی ارشد نرم افزار، دانشگاه پیام نور مرکز تهران، تهران، ایران؛ [۲] گروه مکاترونیک، دانشگاه تهران، تهران، ایران.؛ [۳] گروه برق، دانشگاه پیام نور مرکز تهران، تهران، ایران؛ [۴] گروه قلب و عروق، دانشگاه علوم پزشکی تهران، تهران، ایران.

تاریخ دریافت: ۹۱/۱۰/۵ اصلاح نهایی: ۹۲/۲/۲۸ تاریخ پذیرش: ۹۲/۳/۲۶

چکیده:

زمینه و هدف: **با توجه به آنکه خطرات اجرای روش‌های تشخیص تهاجمی در بیماری عروق کرونر از جمله آنژیوگرافی قابل ملاحظه می‌باشد و از طرفی تجارب موفقیت آمیزی در مورد روش‌های داده کاوی در پزشکی حاصل شده است؛ لذا این مطالعه با هدف تولید مدلی مبتنی بر تکنیک داده کاوی شبکه‌های عصبی که قابلیت پیش‌بینی بیماری عروق کرونر را داشته باشد انجام شده است.**

روش بررسی: **در این مطالعه توصیفی-تحلیلی، مجموعه داده‌ای شامل ۹ ریسک فاکتور از اطلاعات ۱۳۲۲۸ نفر که در مرکز قلب تهران آنژیوگرافی شده بودند (۴۰۵۹ نفر فاقد بیماری عروق کرونر و ۹۱۶۹ نفر مبتلا به این بیماری) مورد استفاده قرار گرفت. تولید مدل پیش‌بینی بیماری عروق کرونر بر اساس شبکه عصبی پرسپترون چند لایه و روش گزینش متغیر، مبتنی بر درخت رگرسیون و طبقه بندی می‌باشد که هر دو با استفاده از نرم افزار Statistica انجام شده است. برای مقایسه و انتخاب بهترین مدل از آنالیز منحنی راک استفاده گردید.**

یافته‌ها: **پس از هفت مرتبه مدل سازی و مقایسه مدل‌های تولید شده، مدل نهایی تشکیل شده از کل ریسک فاکتورهای موجود با سطح زیر منحنی راک ۰/۷۵۴، دقت ۷۴/۱۹٪، حساسیت ۹۲/۴۱٪ و ویژگی ۳۳/۲۵٪ بدست آمد. در نتیجه انجام گزینش متغیر نیز مدلی متشکل از ٤ ریسک فاکتور با سطح زیر منحنی راک ۰/۷۳۷، دقت ۷٤/۱۹٪، حساسیت ۹۳/۳٤٪ و ویژگی ۳۱/۱۷٪ تولید شد.**

نتیجه‌گیری: **در این مطالعه مدل بدست آمده مبتنی بر شبکه‌های عصبی، علاوه بر توانایی بالا در تشخیص افراد بیمار، تعداد قابل قبولی از افرادی که فاقد بیماری عروق کرونر بودند را نیز شناسایی کرد. همچنین، بکارگیری تکنیک‌های گزینش متغیر در این مطالعه نیز نتایج خوبی در زمینه کاهش پیچیدگی مدل به همراه داشت و منجر به تولید مدلی متشکل از تنها چهار ریسک فاکتور سن، جنس، دیابت و فشارخون بالا گردید.**

واژه‌های کلیدی: بیماری عروق کرونر، مدل سازی بیماری، شبکه‌های عصبی، گزینش متغیر.

مقدمه:

بیماری‌های قلبی عروقی امروزه مهمترین عامل مرگ و میر انسان‌ها می‌باشند (۱). یکی از شایع‌ترین بیماری‌های قلبی-عروقی، بیماری عروق کرونر است که طبق گزارش سازمان بهداشت جهانی در سال ۲۰۰۸ با ۷/۲۵ میلیون مرگ در جهان، در صدر ده عامل اول مرگ و میر جهان قرار گرفته است (۲). در ایران نیز بر اساس اطلاعات حاصل از یک مطالعه اپیدمیولوژیک که به بررسی علت مرگ و میر در سال ۱۳۸۸ پرداخته است از بین ۳۲۱۵۷۰ فوتی، تعداد ۸۲۳۰۷ مورد ناشی از بیماری عروق کرونر بوده است که با ۲۵/۶ درصد اولین عامل منجر به مرگ محسوب می‌شود (۳).

به طور کلی می‌توان گفت بیماری عروق کرونر نتیجه همگرایی تعدادی از ریسک فاکتورهای مسبب این بیماری می‌باشد (۴). بررسی و مطالعه منابع مختلف نشان می‌دهد که ریسک فاکتورهای موثر بیماری عروق کرونر شامل: مصرف سیگار، فشار خون بالا، اختلالات

*نویسنده مسئول: دانشگاه تهران، دانشکده علوم و فنون نوین، گروه مکاترونیک، تلفن: ۰۹۱۲۳۰۶۰٤۱۷، E-mail:r.askari@ut.ac.ir

چربی خون (کلسترول تام بالا، تری گلیسیرید بالا، LDL بالا و HDL پایین)، دیابت، عدم فعالیت فیزیکی، چاقی، چاقی شکمی، رژیم غذایی ناسالم، سن، جنسیت، سابقه خانوادگی، مصرف الکل، عوامل روانی، یائسگی، بالا بودن گلوکز ناشتا، فیبرینوژن، لیپو پروتئین a، پروتئین مرحله حاد (CRP) و هموسیستئین می‌باشند (۱۰-۴).

بهترین روش ارزیابی بیماری عروق کرونر آنژیوگرافی می‌باشد و در واقع این روش به عنوان استاندارد طلایی برای تشخیص این بیماری به حساب می‌آید (۱۰). با این وجود آنژیوگرافی یک روش گران و تهاجمی بوده و همراه با ریسک‌هایی چون مرگ، سکته قلبی و سکته مغزی می‌باشد (۶)، لذا برای شناسایی و ارزیابی میزان وسعت این بیماری قبل از آنژیوگرافی عروق کرونر آزمایشات غیر تهاجمی انجام می‌گیرد (۱۰) که به علت حساس و ویژه نبودن این روش‌ها ممکن است در نتایج مثبت یا منفی کاذب داشته باشیم که برای بیمار خطر آفرین می‌باشد، از این رو وجود سیستم‌های پشتیبان تصمیم گیری در کنار روش‌های قبل از آنژیوگرافی برای کم کردن نتایج کاذب لازم به نظر می‌رسد (۱۱).

سیستم‌های پشتیبان تصمیم‌گیری که در حل مسائل و تصمیم گیری‌های پیچیده به کمک انسان‌ها آمده‌اند (۱۲)، اخیراً برای تشخیص بیماری‌ها مورد توجه تعداد زیادی از محققین قرار گرفته‌اند. این سیستم‌ها با استفاده از تکنیک‌های داده کاوی می توانند به کشف الگوها در داده‌های پزشکی پرداخته و فرآیند تصمیم‌گیری را بهبود بخشند و در نتیجه، هزینه‌ها را تحت تأثیر خود قرار داده (۱۳) و کیفیت مراقبت بهداشتی را افزایش دهند (۱۴). تکنیک‌های مختلفی برای داده کاوی وجود دارد که از متداول‌ترین آن‌ها می‌توان درخت تصمیم‌گیری، دسته بندی‌کننده بیزین (Bayesian)، شبکه‌های عصبی، ماشین بردار پشتیبان، قوانین انجمنی، دسته‌بندی مبتنی بر قانون، نزدیک‌ترین همسایه k ام، مجموعه‌های خشن، الگوریتم‌های خوشه‌بندی و الگو ریتم‌های ژنتیک را نام برد (۱۵). در این میان شبکه‌های عصبی مصنوعی که الهام گرفته از شبکه عصبی بیولوژیکی می‌باشند- یک مدل ریاضی از سیستم تشخیصی انسان به حساب می‌آیند و به طور وسیعی در زمینه‌های مختلف به ویژه پزشکی مورد استفاده قرار گرفته اند (۱۶). شبکه‌های عصبی با توجه به کاربردهای متنوع، در انواع مختلفی وجود دارند که یکی از پرکاربردترین آن ها در حل مسائل دسته بندی، شبکه عصبی پرسپترون چند لایه (Multi Layer Perceptron= MLP) می‌باشد (۱۷).

تحقیقات متعددی برای پیش‌بینی بیماری عروق کرونر با استفاده از تکنیک‌های داده کاوی و مجموعه داده‌های مختلف انجام شده است؛ به طور مثال در تحقیقی از سه تکنیک رگرسیون لجستیک، درخت تصمیم گیری و طبقه بندی و شبکه‌های عصبی به همراه داده‌های ۸ ریسک فاکتور مربوط به ۱۲۴۵ نفر استفاده شد و در نهایت مدل شبکه عصبی پرسپترون چند لایه با دقت ۷۸/۷ درصد بهترین مدل معرفی گردید (۱۸). در مطالعات دیگر با استفاده از شبکه عصبی، بر روی مجموعه داده‌هایی با ریسک فاکتورها و اندازه‌های متفاوت به مدل سازی بیماری عروق کرونر پرداخته شد و مدل‌هایی با دقت ۸۹ درصد و ۷۲ درصد بدست آمد (۱۱). در برخی تحقیقات نیز از مجموعه داده های بیماران عروق کرونر موجود در مخزن یادگیری ماشین دانشگاه ایروین کالیفرنیا (۱۹) استفاده شده است که بر اساس تکنیک‌های داده کاوی مختلف، نتایج متفاوتی بدست آمده است (۲۲-۲۰).

با توجه به خطرات اجرای روش‌های تشخیص تهاجمی از جمله آنژیوگرافی و تجارب موفقیت آمیزی که در مورد روش‌های داده کاوی حاصل شده است؛ لذا در این مطالعه با استفاده از تکنیک شبکه‌های عصبی به ارائه یک مدل تشخیصی که قابلیت پیش‌بینی بیماری عروق کرونر را داشته باشد پرداخته شده است.

روش بررسی:

در این مطالعه توصیفی-تحلیلی برای دستیابی به مدل پیش‌بینی بیماری عروق کرونر از بانک اطلاعاتی

آنژیوگرافی مرکز قلب تهران شامل ۱۳۲۲۸ رکورد اطلاعاتی استفاده شده است. این بانک اطلاعاتی شامل ریسک فاکتورهای: سن، جنس، چاقی، چاقی شکمی، سابقه خانوادگی، مصرف سیگار، چربی خون بالا، دیابت و فشارخون بالا می‌باشد. (جدول شماره ۱).

برای جلوگیری از پدیده انطباق بیش از حد (Overfitting) و ارزیابی قابلیت تعمیم مدل، قبل از شروع فرآیند مدل سازی، مجموعه داده‌های موجود به سه زیر مجموعه آموزش (۶۰٪)، آزمایش (۲۰٪) و اعتبارسنجی (۲۰٪) تقسیم شدند (۲۳).

جدول شماره ۱: آماره‌های توصیفی مجموعه داده بدست آمده از بانک اطلاعات آنژیوگرافی مرکز قلب تهران

متغیر وابسته	دامنه	تعریف عملیاتی			
بیماری عروق	۰ , ۱	فاقد بیماری (۰)، مبتلا (۱)			
متغیرهای مستقل	دامنه	تعریف عملیاتی	مبتلا به بیماری عروق کرونر (٪۶۹)	فاقد بیماری عروق کرونر (٪۳۱)	*Pvalue
سن	۱۰۰-۱۸	سال	**۶۱/۳۲±۱۰/۵۲	**۵۶/۲۸±۱۱/۲۶	<۰/۰۰۱
جنس	۰ , ۱	مرد (۰)، زن (۱)	مرد (٪۶۹)	مرد (٪۴۵)	<۰/۰۰۱
سابقه خانوادگی	۰ , ۱	ندارد (۰)، دارد (۱)	ندارد (٪۸۲)	ندارد (٪۸۵)	<۰/۰۰۲
مصرف سیگار	۰ , ۱ , ۲	ندارد (۰)، دارد (۱)، ترک	ندارد (٪۶۰)، دارد (٪۲۴)	ندارد (٪۷۶)، دارد (٪۱۴)	<۰/۰۰۱
چربی خون بالا	۰ , ۱	ندارد (۰)، دارد (۱)	ندارد (٪۳۳)	ندارد (٪۴۱)	<۰/۰۰۱
فشار خون بالا	۰ , ۱	ندارد (۰)، دارد (۱)	ندارد (٪۴۱)	ندارد (٪۴۹)	<۰/۰۰۱
دیابت قندی	۰ , ۱	ندارد (۰)، دارد (۱)	ندارد (٪۶۴)	ندارد (٪۷۹)	<۰/۰۰۱
چاقی شکمی	۰ , ۱	ندارد (۰)، دارد (۱)	ندارد (٪۴۵)	ندارد (٪۳۱)	<۰/۰۰۱
چاقی (BMI)	۰	کمتر از ۱۸/۵ (۰)	(٪۱)	(٪۱)	<۰/۰۰۱
	۱	بین ۱۸/۵ – ۲۴/۹ (۱)	(٪۲۵)	(٪۲۱)	
	۲	بین ۲۵ – ۲۹/۹ (۲)	(٪۴۶)	(٪۴۰)	
	۳	بزرگ تر از ۳۰ (۳)	(٪۲۸)	(٪۳۸)	

*نشان دهنده اختلاف معنی دار بین افراد مبتلا به بیماری عروق کرونر و فاقد بیماری؛ **داده ها به صورت "میانگین±انحراف معیار" می باشند.

در این پژوهش برای مدل سازی از یک شبکه عصبی MLP سه لایه استفاده شد. همچنین الگوریتم آموزشی این شبکه، الگوریتم (Broyden-Fletcher-Goldfarb-Shanno= BFGS) بود که در واقع توسعه یافته الگوریتم شبه نیوتن می‌باشد. این الگوریتم نسبت به الگوریتم‌های شیب توام و گرادیان مزدوج، سریع‌تر بوده و در تعداد گردش کمتری همگرا می‌شود؛ از این رو یکی از مناسب‌ترین الگوریتم‌های آموزشی می‌باشد (۲۴). از آنجا که رابطه‌ای برای برآورد پارامترهایی چون تعداد نرون‌های لایه پنهان، تابع فعالیت لایه‌ها و تابع خطا، در یک مدل شبکه عصبی موجود نبوده و با تکرار آزمایش مقادیر مناسب آن ها یافت می‌شود؛ لذا مطابق سود و کد موجود در تصویر شماره ۱ در هر مرتبه مدل سازی ۱۰۰ شبکه عصبی انجام گرفت.

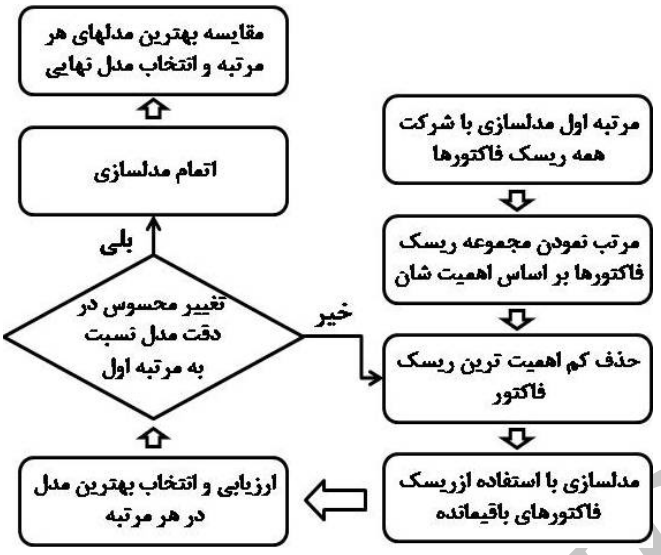
```
Create model ()
{
    for (int i=0; i<100; i++)
    {
        //select activation functions for hidden layer randomly
        AFHL = Random (Identity, Logistic, Tanh, Exponential)
        //select activation functions for output layer randomly
        AFOL = Random (Identity, Logistic, Tanh, Exponential, Softmax)
        //select error functions randomly
        EF = Random (Sum of squares, Cross entropy);
        //number of neuron in hidden layer
        NN = Random (Between 3 and 13)
        //create neural network
        Network [i] = CreateNeuralNetwork (AFHL , AFOL, EF, NN)
    }
}
```

تصویر شماره ۱: سود و کد تولید شبکه عصبی

برای تولید مدل مورد نظر جهت پیش بینی بیماری عروق کرونر اولین مرتبه مدل سازی شرکت همه ریسک فاکتورهای موجود بود (تصویر شماره ۲). در ادامه برای حذف متغیرهای اضافی و نامربوط و تولید مدل‌های ساده تر از گزینش متغیر استفاده شد (۲۵،۲۶). بدین ترتیب که ابتدا با استفاده از تکنیک درخت رگرسیون و طبقه‌بندی (Classification and Regression Tree= CART) ترتیب اهمیت ریسک فاکتورها مشخص گردید؛ سپس بر اساس روش حذف رو به عقب مرحله به مرحله (Stepwise Backward Elimination) در هر مرتبه مدل سازی، کم اهمیت‌ترین ریسک فاکتور حذف و مدل سازی با ریسک فاکتورهای باقیمانده انجام پذیرفت. این کار تا زمانی که در دقت مدل‌های به دست آمده در هر مرتبه نسبت به مرتبه اول تغییر محسوسی مشاهده نگردید، ادامه یافت. این روند هفت مرتبه تکرار شد تا اینکه در مرتبه هفتم مدل سازی دقت به دست آمده نسبت به مرتبه اول دارای تغییرات محسوس بود. بدین معنی که ریسک فاکتورهای باقیمانده برای مدل سازی کافی نبود؛ لذا فرآیند مدل سازی متوقف گردید و پس از اتمام مدل سازی، بهترین مدل‌های به دست آمده در هر مرتبه مقایسه و مدل نهایی انتخاب شد.

در این مطالعه از سطح زیر منحنی مشخصه عملکرد (ROC) برای مقایسه کارآیی مدل‌ها استفاده گردید. این روش در سال های اخیر به طور گسترده برای ارزیابی الگوریتم‌های یادگیری ماشین مورد استفاده قرار گرفته است (۲۷) و در زمینه پزشکی نیز به عنوان یک روش موثر برای ارزیابی کارآیی تست‌های تشخیصی در برابر استانداردهای طلایی، به کار گرفته می‌شود (۲۸).

برای مشخص کردن اینکه بعد از حذف یک ریسک فاکتور و مدل سازی با ریسک فاکتورهای باقیمانده، تغییرات به دست آمده در سطح زیر منحنی راک محسوس می‌باشد یا خیر، از روش دیلانگ (Delong) استفاده شده است که معنی‌دار بودن (یا نبودن) تفاوت میان دو سطح زیر منحنی راک را نشان می‌دهد (۲۹). برای اجرای روش دیلانگ از بسته pROC موجود در نرم افزار R نسخه ۲،۳،۱۵ استفاده شده است.



تصویر شماره ۲: روند ایجاد مدل پیش‌بینی بیماری عروق کرونر مبتنی بر شبکه‌های عصبی و گزینش متغیر

یافته ها:

در اولین مرتبه مدل سازی که با شرکت کلیه ریسک فاکتورهای موجود انجام شد، ۱۰۰ شبکه عصبی مطابق با سود و کد تولید شد (تصویر شماره ۱) که پس از ارزیابی آن ها بهترین شبکه با سطح زیر منحنی ۰/۷۵۴ انتخاب گردید. در ادامه با استفاده از تکنیک درخت رگرسیون و طبقه بندی اهمیت ریسک فاکتورها به ترتیب: سن (۱۰۰٪)، دیابت (۸۶٪)، فشارخون (۵۲٪)، جنس (۴۹٪)، چربی خون بالا (۳۷٪)، مصرف سیگار (۳۶٪)، چاقی (۱۷٪)، چاقی شکمی (۱۷٪) و سابقه خانوادگی (۱۳٪) بدست آمد.

در مرتبه دوم مدل سازی، با توجه به ترتیب اهمیت ریسک فاکتورها، کم اهمیت‌ترین ریسک فاکتور که سابقه خانوادگی بود، حذف و مدل سازی با ریسک فاکتورهای باقیمانده تکرار گردید که پس از ارزیابی ۱۰۰ شبکه عصبی تولید شده، بهترین آن ها با سطح زیر منحنی راک ۰/۷۵۲ انتخاب شد. مقایسه سطح زیر منحنی راک بهترین مدل به دست آمده در مرتبه دوم نسبت به مرتبه اول با استفاده از روش دیلانگ نشان داد که تفاوت معنی‌دار بین دو مدل وجود ندارد (P=۰/۷۸۹)؛ لذا مدل سازی برای سومین مرتبه و بدون ریسک فاکتور چاقی شکمی ادامه یافت. پس از ارزیابی ۱۰۰ شبکه عصبی تولید شده در این مرتبه، بهترین آن ها با سطح زیر منحنی راک ۰/۷۵۱ انتخاب شد. مقایسه سطح زیر منحنی راک بهترین مدل‌های بدست آمده در مرتبه‌های سوم و اول با استفاده از روش دیلانگ نشان داد که تفاوت معنی‌دار بین دو مدل وجود ندارد (P=۰/۴۷۰). به همین ترتیب روند حذف ریسک فاکتورهای کم اهمیت و مدل سازی با ریسک فاکتورهای باقیمانده ادامه یافت. نتایج حاصل از روش دیلانگ برای مقایسه بهترین مدل‌های بدست آمده در مرتبه‌های چهارم و اول (P=۰/۴۰۲)، مرتبه‌های پنجم و اول (P=۰/۳۹۸)، مرتبه‌های ششم و اول (P=۰/۱۳۳) نشان از عدم وجود تفاوت معنی‌دار بین این مدل‌ها نسبت به مدل مرتبه اول داشت، ولی نتیجه مقایسه بهترین مدل‌های بدست آمده در مرتبه‌های هفتم و اول (P<۰/۰۰۱) نشان داد که تفاوت معنی‌دار بین این مدل‌ها وجود دارد و این بدین معنا بود که تعداد ریسک فاکتورهای باقیمانده برای مدل سازی کافی نمی‌باشد؛ لذا فرآیند مدل سازی متوقف و مشخصات مدل‌های به دست آمده در هفت مرتبه مدل سازی ترسیم شد (جدول شماره ۲).

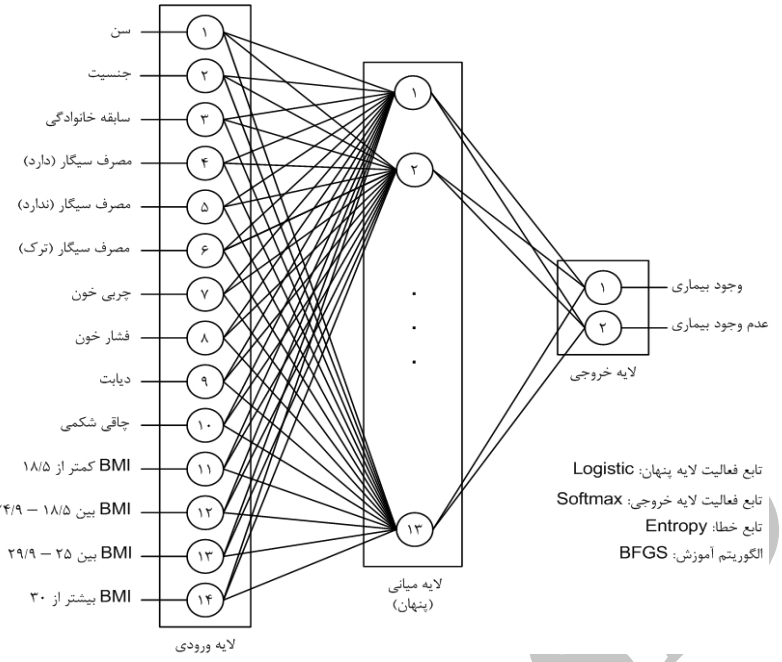
جدول شماره ۲: نتایج هفت مرتبه مدل سازی برای پیش‌بینی بیماری عروق کرونر مبتنی بر شبکه‌های عصبی

شماره مدل	ریسک فاکتورها	مجموعه داده	دقت	سطح زیر راک	ویژگی	حساسیت
۱	سن، جنس، چاقی، چاقی شکمی، سابقه خانوادگی، مصرف سیگار، چربی‌خون بالا، دیابت، فشارخون بالا	آموزش	٪۷۵/۶۰	۰/۷۴۲	٪۳۶/۷۷	٪۹۲/۵۱
		ارزیابی	٪۷۴/۱۹	۰/۷۵۴	٪۳۳/۲۵	٪۹۲/۴۱
۲	سن، جنس، چاقی، چاقی شکمی، مصرف سیگار، چربی خون بالا، دیابت، فشارخون بالا	آموزش	٪۷۵/۶۶	۰/۷۴۱	٪۳۶/۳۹	٪۹۲/۷۵
		ارزیابی	٪۷۳/۷۷	۰/۷۵۲	٪۳۲/۳۹	٪۹۲/۱۹
۳	سن، جنس، چاقی، مصرف سیگار، چربی خون بالا، دیابت، فشارخون بالا	آموزش	٪۷۵/۷۱	۰/۷۴۰	٪۳۶/۹۳	٪۹۲/۵۸
		ارزیابی	٪۷۳/۲۴	۰/۷۵۱	٪۳۲/۶۴	٪۹۲/۴۱
۴	سن، جنس، مصرف سیگار، چربی خون بالا، دیابت، فشارخون بالا	آموزش	٪۷۵/۷۹	۰/۷۳۶	٪۳۶/۷۳	٪۹۲/۸۰
		ارزیابی	٪۷۴/۰۴	۰/۷۴۹	٪۳۲/۱۵	٪۹۲/۶۸
۵	سن، جنس، چربی خون بالا، دیابت، فشارخون بالا	آموزش	٪۷۵/۴۵	۰/۷۳۰	٪۳۶/۱۴	٪۹۲/۵۷
		ارزیابی	٪۷۴/۱۵	۰/۷۴۰	٪۳۲/۲۷	٪۹۲/۷۹
۶	سن، جنس، دیابت، فشارخون بالا	آموزش	٪۷۵/۲۱	۰/۷۲۷	٪۳۴/۱۱	٪۹۳/۱۱
		ارزیابی	٪۷۴/۱۹	۰/۷۳۷	٪۳۱/۱۷	٪۹۳/۳۴
۷	سن، دیابت و فشارخون بالا	آموزش	٪۷۰/۶۹	۰/۶۵۴	٪۱۱/۹۲	٪۹۶/۲۷
		ارزیابی	٪۶۹/۵۰	۰/۶۶۵	٪۱۲/۰۲	٪۹۵/۸۵

همانگونه که در جدول شماره ۲ مشخص شده است، مدل شماره ۱ از نظر سطح زیر منحنی راک و دقت، بهترین مدل می‌باشد؛ لذا به عنوان مدل منتخب انتخاب می‌گردد. با این وجود مدل شماره ۶ که از نظر دقت و سطح زیر منحنی راک تفاوت معنی داری با مدل منتخب ندارد، تنها از چهار ریسک فاکتور تشکیل شده است و شبکه عصبی حاصل از آن ساده‌تر می‌باشد.

مجموعه داده مورد استفاده در تحقیق حاضر شامل متغیرهایی از نوع دسته‌ای (Categorical) بود. برای مثال متغیر چاقی دارای چهار گروه می‌باشد که با اعداد ۰ تا ۳ نشان داده شده‌اند. همانگونه که مشخص است این متغیرها از جنس عدد نیستند و نمایش عددی آن ها صحیح نمی‌باشد؛ لذا به طور معمول بهترین روش برای استفاده از آن ها در مدل سازی، تبدیل هر متغیر به چند متغیر ساختگی (Dummy) می‌باشد که از نوع دودویی بوده و فقط مقادیر ۰ و ۱ می‌گیرد (۳۰،۳۱). پس از تبدیل این نوع متغیرها به نوع متغیرهای ساختگی تعداد ورودی‌های شبکه به ۱۴ افزایش پیدا کرد.

شبکه عصبی تولید شده برای مدل منتخب، یک شبکه پرسپترون دو لایه بدست آمد که در واحد ورودی دارای ۱۴ نرون بود. این شبکه شامل یک لایه پنهان با تعداد ۱۳ نرون بوده و در واحد خروجی نیز دارای دو نرون است که یکی بیانگر وجود بیماری و دیگری عدم وجود بیماری عروق کرونر می‌باشد (تصویر شماره ۳). از دیگر ویژگی‌های این شبکه پیاده سازی راحت‌تر سخت افزاری و تهیه یک نمونه قابل حمل می‌باشد.



تصویر شماره ۳ : شبکه عصبی پیش‌بینی کننده بیماری عروق کرونر

بحث:

در برخی از تحقیقاتی که از مجموعه داده موجود در مخزن یادگیری ماشین دانشگاه ایروین کالیفرنیا استفاده شده است (۱۹)، مدل‌هایی با دقت بالا به دست آمده که علت آن استفاده از نتایج معاینات فیزیکی، الکتروکاردیوگرافی، تصویر برداری و استرس تست علاوه بر ریسک فاکتورهای بیماری عروق کرونر می‌باشد. این مسأله نشان می‌دهد استفاده از نتایج آزمایشات تشخیصی قبل از آنژیوگرافی در دستیابی به مدل با دقت بالا بسیار موثر می‌باشد، البته باید در نظر داشت که این آزمایشات علاوه بر ریسک‌های احتمالی برای بیمار، مستلزم صرف زمان و هزینه می باشند. در تحقیق حاضر و تحقیقات دیگری که مدل‌های به دست آمده تنها مبتنی بر ریسک فاکتورهای بیماری عروق کرونر می باشند، به علت تفاوت در ریسک فاکتورهای مورد استفاده، مدل‌هایی با دقت متفاوت گزارش شده است (۳۲،۱۸،۱۱). محدودیت در استفاده از ریسک فاکتورهای کافی در برخی از این تحقیقات (۳۲،۱۸) باعث کم شدن دقت مدل گردیده است که این مسأله در تحقیق حاضر نیز مشهود است. مسأله دیگری که حائز اهمیت می‌باشد این است که در تحقیق حاضر و دیگر تحقیقات مورد بررسی، میزان ویژگی مدل‌ها کمتر از حساسیت می‌باشد و این بدین معنی است که مدل در تشخیص افراد بیمار نسبت به تشخیص افراد سالم تواناتر است.

استفاده از گزینش متغیر در تحقیق حاضر نتایج جالب توجهی به همراه داشت، زیرا منجر به تولید مدلی متشکل از تنها چهار ریسک فاکتور سن، جنس، دیابت و فشار خون شد که تفاوت معنی داری با مدل نهایی ارائه شده نداشت؛ لازم به ذکر است که در هیچ یک از تحقیقات بررسی شده از گزینش متغیر استفاده نشده بود تا با مطالعه حاضر مورد مقایسه قرار گیرد.

در برخی از تحقیقات بررسی شده، از آنالیز منحنی‌های راک برای ارزیابی مدل بدست آمده استفاده شده است و در برخی دیگر ملاک ارزیابی، دقت مدل در مجموعه داده آزمایش بوده است. با توجه به اینکه دقت مدل به تنهایی معیار مناسبی برای ارزیابی مدل نمی باشد، در تحقیق حاضر از آنالیز منحنی‌های

راک که یکی از بهترین معیارهای ارزیابی مدل به حساب می‌آید، استفاده شد.

نتیجه گیری:

در این مطالعه برای پیش‌بینی بیماری عروق کرونر از تکنیک قدرتمند شبکه‌های عصبی استفاده گردید که مدل نهایی به دست آمده دارای دقت ۷۴/۱۹ درصد، ویژگی ۳۳/۲۵ درصد و حساسیت ۹۲/۴۱ درصد بود و توانست با درصد بالایی بیماران را تشخیص داده و تعداد ۲۷۱ نفر از ۸۱۵ فردی که فاقد بیماری عروق کرونر بودند را مشخص نماید. آنچه در این تحقیق و سایر تحقیقات مشابه مشهود است، این است که میزان ویژگی مدل‌ها کمتر از حساسیت می‌باشد و این بدین معنی است که مدل‌ها در تشخیص افراد بیمار نسبت به افراد سالم تواناتر هستند. در برخی تحقیقات استفـاده از نتایج معاینات فیزیکی، الکتروکاردیوگرافی، تصویربرداری و استرس تست به همراه ریسک فاکتورهای بیماری عروق کرونر منجر به تولید مدل‌هایی با دقت بالا شده است. این مسأله نشان می دهد استفاده از نتایج آزمایشات تشخیصی قبل از آنژیوگرافی در دستیابی به مدل با دقت بالا بسیار موثر می‌باشد، البته باید در نظر داشت که این آزمایشات علاوه بر ریسک‌های احتمالی برای بیمار، مستلزم صرف زمان و هزینه می باشند. در این تحقیق از تکنیک درخت رگرسیون و طبقه بندی و روش حذف رو به عقب مرحله به مرحله برای گزینش متغیر استفاده شد که نتایج جالب توجهی به همراه داشت و منجر به تولید مدلی با استفاده از چهار ریسک فاکتور سن، جنس، دیابت، فشار خون بالا و با دقت ۷۴/۱۹ درصد گردید که تفاوت معنی داری با مدل نهایی نداشت.

تشکر و قدردانی:

بدینوسیله از پرسنل بخش تحقیقات مرکز قلب تهران که داده‌های این مطالعه را در اختیار پژوهشگران قرار دادند، سپاسگزاری می‌نماییم.

منابع:

1. Mendis S, Puska P, Norrving B. Global atlas on cardiovascular disease prevention and control. Geneva: World Health Organization; 2011.

2. Fact sheet No. 310 :The top ten causes of death. Geneva: World Health Organization; 2011.

3. Amani F, Kazemnejad A, Habibi R, Hajizadeh E. Pattern of mortality trend in Iran during 1970-2009. J Gorgan Univ Med Sci. 2011; 12(4): 85-90.

4. Bonow RO, Mann DL, Zipes DP, Libby P. Braunwald's Heart Disease: a textbook of cardiovascular medicine. 1st ed. Philadelphia: Saunders; 2012.

5. Cecil RL, Lee Goldman M, Schafer AI. Goldman's cecil medicine. Philadelphia: Elsevier/Saunders; 2012.

6. Fauci A, Braunwald E, Kasper D, Hauser S, Longo D, Jameson J, et al. Harrison's principles of internal medicine. 17th ed. NewYork: McGraw-Hill Companies; 2008.

7. Erqou S, Kaptoge S, Perry PL, Di Angelantonio E, Thompson A, White IR, et al. Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA. 2009; 302(4): 412-23.

8. Emerging Risk Factors C, Kaptoge S, Di Angelantonio E, Lowe G, Pepys MB, Thompson SG, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet. 2010 Jan; 375(9709): 132-40.

9. Humphrey LL, Fu R, Rogers K, Freeman M, Helfand M. Homocysteine level and coronary heart disease incidence: a systematic review and meta-analysis. Mayo Clinic Proceedings Mayo Clinic. 2008 Nov; 83(11): 1203-12.

10. Crawford M. Current diagnosis & treatment in cardiology. 3rd ed. NewYork: McGraw-Hill Medical; 2009.

11. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med. 2000 Mar; 18(3): 187-203.

12. Shim JP, Warkentin M, Courtney JF, Power DJ, Sharda R, Carlsson C. Past, present, and future of decision support technology. Dec Supp Sys. 2002; 33(2): 111-26.

13. Koh HC, Tan G. Data mining applications in healthcare. J Healthcare Info Manag. 2005 Spring; 19(2): 64-72.

14. Chae YM, Kim HS, Tark KC, Park HJ, Ho SH. Analysis of healthcare quality indicator using data mining and decision support system. Exp Sys Applic. 2003; 24(2): 167-72.

15. Gorunescu F. Data mining: concepts, models and techniques. Verlag Berlin Heidelberg: Springer; 2011.

16. Fausett LV. Fundamentals of neural networks: architectures, algorithms, and applications. New Jersey: Prentice-Hall; 1993.

17. Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 2008.

18. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Exp Sys Applic. 2008; 34(1): 366-74.

19. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE. Neural network predictions of significant coronary artery stenosis in men. Artif Intell Med. 2005 Jun; 34(2): 151-61.

20. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl. 2009; 36(4): 7675-80.

21. Khatibi V, Montazer GA. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Syst Appl. 2010; 37(12): 8536-42.

22. Muthukaruppan S, Er MJ. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Syst Appl. 2012;39(14):11657-65.

23. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer; 2009.

24. Bishop CM, Hinton G. Neural networks for pattern recognition. Oxford: Clarendon Press; 1995.

25. Han J, Kamber M, Pei J. Data mining, concepts and techniques. 2nd ed. Morgan Kaufman; 2006.

26. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Minnesota: Pearson Addison Wesley; 2006.

27. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006; 27(8): 861-74.

28. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr. 2011 Apr; 48(4): 277-87.

29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep; 44(3): 837-45.

30. Shmueli G, Patel NR and Bruce PC. Data mining for business intelligence: concepts, techniques, and applications in microsoft office excel with xlminer: Wiley; 2011.

31. Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications: Elsevier Science; 2009.

32. Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV, Naka KK, et al. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE transactions on information technology in biomedicine. IEEE Trans Inf Technol Biomed. 2008 Jul; 12(4): 447-58.

Journal of Shahrekord University of Medical Sciences
(J Shahrekord Univ Med Sci) 2013 Dec, Jan; 15(5): 47-56.

Original article

Prediction model for coronary artery disease using neural networks and feature selection based on classification and regression tree

Mahmoudi I (MA Student)[1], Askari moghadam R (PhD)[2]*, Moazzam MH (PhD)[3], Sadeghian S (MD)[4]
[1]Student of software engineering, Tehran Payame Noor University, Tehran, I.R. Iran; [2]Mechatronic Dept., Tehran University, Tehran, I.R. Iran; [3]Electronic Dept., Tehran Payame Noor University, Tehran, I.R. Iran; [4]Cardiology Dept, Tehran University of Medical Sciences, Tehran, I.R. Iran.
Received: 25/Dec/2012 Revised: 18/May/2013 Accepted: 16/June/2013

Background and aims: Risk of implementing invasive diagnostic procedures for coronary artery disease (CAD) such as angiography is considerable. On the other hand, Successful experience has been achieved in medical data mining approaches. Therefore this study has been done to produce a model based on data mining techniques of neural networks that can predict coronary artery disease.
Methods: In this descriptive- analytical study, the data set includes nine risk factors of 13228 participants who were undergone angiography at Tehran Heart Center. (4059 participants were not suffering from CAD but 9169 were suffering from CAD). Producing model for predicting coronary artery disease was done based on multilayer perceptron neural networks and variable selection based on classification and regression tree (CART) using of Statistica software. For comparison and selection of best model, the ROC curve analysis was used.
Results: After seven-time modeling and comparing the generated models, the final model consists of all existing risk factors obtained with the area under ROC curve of 0.754, accuracy of 74.19%, sensitivity of 92.41% and specificity of 33.25% .Also, variable selection results in producing a model consists of four risk factors with area under ROC curve of 0.737, accuracy of 74.19%, sensitivity of 93.34% and specificity of 31.17% was produced.
Conclusion: The obtained model is produced based on neural networks. The model is able to identify both high risk patients and acceptable number of healthy subjects. Also, utilizing the feature selection in this study ends up in production of a model which consists of only four risk factors as: age, sex, diabetes and high blood pressure.

Keywords: **Coronary artery disease, Feature selection, Neural networks, Modeling.**

Cite this article as: Mahmoudi I, Askari moghadam R, Moazzam MH, Sadeghian S. Prediction model for coronary artery disease using neural networks and feature selection based on classification and regression tree. J Shahrekord Univ Med Sci. 2013 Dec, Jan; 15(5): 47-56.

***Corresponding author:**
Mechatronics Group, Faculty of New Sciences and Technologies, University of Tehran, Tehran, I.R. Iran. Tel: 00989123060417, E-mail:r.askari@ut.ac.ir