

پیش‌بینی بقاء بیماران مبتلا به سرطان ریه با استفاده از سیستم استنتاج عصبی - فازی تطبیقی بهبود یافته

ام‌البین عباسی^۱، محمدرضا رمضان پور^{۲*}، ریحانه خورسند^۳

• پذیرش مقاله: ۱۳۹۸/۱۲/۴

• دریافت مقاله: ۱۳۹۸/۸/۵

مقدمه: سرطان ریه منبع اصلی مرگ و میر برای مردان و زنان در سراسر جهان می‌باشد. بیماری ریه توسعه و رشد غیرقابل کنترل سلول‌ها در یک یا هر دو ریه می‌باشد. تشخیص زودرس سرطان آسان نیست؛ اما اگر سریع تشخیص داده شود، قابل درمان است. هدف از این مطالعه، ساخت مدل بهینه پیش‌بینی کننده بقاء بیماران مبتلا به سرطان ریه بر اساس ویژگی‌های بیماران با رویکرد داده‌کاوی می‌باشد.

روش: در این مطالعه توصیفی- کاربردی، از الگوریتم سیستم استنتاج عصبی فازی تطبیقی ANFIS و الگوریتم بهینه‌سازی ازدحام ذرات PSO برای پیش‌بینی بقاء بیماران مبتلا به سرطان ریه استفاده شد. در این مطالعه، از پایگاه داده معتبر برنامه نظارت، اپیدمی‌شناسی و نتایج نهایی SEER دانشگاه لویزول آمریکا استفاده شد. برای ارزیابی روش پیشنهادی از معیارهای دقت، صحت، خطا و جذر خطای میانگین مربعات استفاده شد.

نتایج: نتایج نهایی به دست آمده در این مطالعه نشان‌دهنده برتری روش بهینه‌سازی ANFIS با الگوریتم PSO نسبت به سایر روش‌ها، در راستای پیش‌بینی بقاء بیماران مبتلا به سرطان ریه با متوسط صحت برابر ۹۹/۸۰٪ برای بقاء یک‌ساله، ۹۹/۷۴٪ برای بقاء دو ساله و ۹۹/۶۶٪ برای بقاء پنج‌ساله بر روی مجموعه داده SEER بود.

نتیجه‌گیری: استفاده از مدل بهینه‌سازی شده ANFIS با الگوریتم PSO در پیش‌بینی بقاء بیماران مبتلا به سرطان ریه بسیار قدرتمند است. مدل پیشنهادی نسبت به سایر مدل‌های مورد مقایسه دارای بیشترین صحت، دقت و کمترین میزان خطا بوده است؛ بنابراین به کارگیری این مدل در زمینه پیش‌بینی بقاء پیشنهاد می‌شود.

کلید واژه‌ها: داده‌کاوی، پیش‌بینی بقاء، سرطان ریه، نرخ بقاء

• ارجاع: عباسی ام‌البین، رمضان پور محمدرضا، خورسند ریحانه. پیش‌بینی بقاء بیماران سرطان ریه با استفاده از سیستم استنتاج عصبی - فازی تطبیقی بهبود یافته. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۱): ۱۹-۲۹.

۱. دانشجوی کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد مبارکه، دانشگاه آزاد اسلامی، اصفهان، ایران

۲. استادیار، گروه مهندسی کامپیوتر، واحد مبارکه، دانشگاه آزاد اسلامی، اصفهان، ایران

۳. استادیار، گروه مهندسی کامپیوتر، واحد دولت آباد، دانشگاه آزاد اسلامی، اصفهان، ایران

* نویسنده مسئول: محمدرضا رمضانپور

آدرس: اصفهان، مبارکه، بلوار معلم، میدان فردوسی، بلوار شهید نصوحی، دانشگاه آزاد اسلامی واحد مبارکه

• Email: mr.ramezanpoor@gmail.com

• شماره تماس: ۰۳۱۵۲۴۹۲۰۵۸

مقدمه

سرطان ریه نوعی بیماری است که مشخصه آن رشد کنترل نشده سلول در بافت‌های ریه است. اگر این بیماری درمان نشود، رشد سلولی می‌تواند در فرایند متاستاز به بیرون از ریه گسترش پیدا کند و به بافت‌های اطراف یا سایر اعضای بدن برسد. اکثر سرطان‌هایی که از ریه شروع می‌شوند، به نام سرطان‌های ابتدایی ریه، کارسینوماهایی هستند که از بافت پوششی نشأت می‌گیرند. انواع اصلی سرطان ریه، سرطان‌های ریه سلول کوچک و سلول غیرکوچک هستند. شایع‌ترین علائم شامل سرفه و خلط خونی، کاهش وزن و تنگی نفس می‌باشند [۱،۲].

داده‌کاوی به منظور یافتن الگو از درون پایگاه داده‌های بزرگ داده و نیز پیش‌بینی سرطان‌ها با استفاده از الگوریتم‌های پیش‌بینی کننده به کار می‌رود. هدف از داده‌کاوی پیش‌بینی کننده در حوزه پزشکی و بالینی رسیدن به مدلی است که با استفاده از اطلاعات بیمار پیامدها را پیش‌بینی نموده و بر اساس آن تصمیم گرفت. در این زمینه می‌توان به مطالعه Massion و همکاران اشاره کرد، که جهت پیش‌بینی بقای سرطان ریه از ماشین بردار پشتیبان استفاده نمودند. از آنجایی که ماشین بردار پشتیبان یک روش خطی در جداسازی داده‌ها می‌باشد، بیشتر تمرکز خود را بر روی کاهش ریسک و افزایش فاصله میان داده‌های کلاس‌های باینری قرار می‌دهد؛ لذا یکی از مشکلات این کلاسه‌بند، عدم کارایی در حجم پایین دیتاست ورودی و همچنین میزان حساسیت بالا در انتخاب ابر صفحه‌های جداکننده با بیشترین میزان حاشیه می‌باشد که باعث می‌شود نقاط مرزی به درستی کلاسه‌بندی نگردند [۳].

Win و همکاران، با استفاده از استخراج ویژگی ژن‌های بیمار ریوی به صورت جداگانه و بیان میزان ناهنجاری به صورت یک تابع لگاریتمی با استفاده از کلاسه‌بند باینری درخت تصمیم‌گیری عملیات پیش‌بینی انجام دادند [۴].

Tunç و Karhan، از مدل K- نزدیک‌ترین همسایه (K-Nearest Neighbor) در پیش‌بینی بقای سرطان ریه استفاده کردند. اساس کار این روش بر پایه میزان فاصله میان نمونه‌های مختلف است که با یک عدد فرد همانند k تعداد همسایگان را مورد بررسی قرار می‌دهد. اگر تعداد همسایگان یک نمونه مثبت باشد، آن نمونه نیز مثبت در نظر گرفته می‌شود و بالعکس. یکی از مشکلات این روش، نبود فاز آموزشی داده‌ها می‌باشد که باعث می‌شود در صورتی که مقدار k کم باشد نمونه‌ها به صورت تصادفی و در غیر این صورت اگر

عدد k بالا باشد بر اساس همسایگان یک نمونه مقداردهی شوند که این کار بدون هیچ مرز به خصوصی میان داده‌های مختلف انجام می‌پذیرد [۵].

Yu و همکاران در پژوهشی با عنوان پیش‌بینی پروفیل‌های مولکولی سرطان ریه‌های غیرسلولی با استفاده از شبکه‌های عصبی مصنوعی ارائه دادند. در این پژوهش یک شبکه عصبی کانولوشن تشخیصی را برای یادگیری ویژگی‌های عمیق برای پیش‌بینی وضعیت رشد اپیدمی با رشد سلول‌های سرطانی مرتبط استفاده شده است. در این رویکرد از ۵۹۵ بیمار برای آموزش و ۸۹ بیمار جهت اعتبار سنجی و ارزیابی استفاده شده است [۶].

Murty و همکاران در پژوهشی با عنوان پیش‌بینی بیماری سرطان ریه ارائه دادند. در این پژوهش، برای پیش‌بینی سرطان ریه با استفاده از الگوریتم‌های طبقه‌بندی مانند بیزین، شبکه عصبی شعاعی پایه یا (RBF (Radial Basis Function)، انجام شده است. در ابتدا ۳۲ نمونه سرطانی و غیر سرطانی داده‌ها با ۵۷ ویژگی جمع‌آوری شد، قبل از پردازش و تجزیه و تحلیل با استفاده از الگوریتم‌های طبقه‌بندی و بعد از آن همان روش در ۹۶ نمونه و ۷۱۳۰ ویژگی برای پیش‌بینی سرطان ریه صورت گرفته است. مجموعه داده‌های مورد استفاده در این مطالعه از مجموعه داده‌های دانشگاه کالیفرنیا در ارواین UCI (University of California at Irvine) برای یادگیری ماشین‌های سرطان ریه و بیماران مبتلابه سرطان ریوی می‌شیرگان صورت گرفته است [۷].

Wang و همکاران، به پیش‌بینی نرخ بقاء سرطان ریه با داده‌های ۲۰۰ بیمار مبتلابه سرطان ریه پرداختند. در این پژوهش طبقه‌بندهای مورد استفاده درخت تصمیم‌گیری و تکنیک شبکه عصبی است. نتایج به دست آمده با استفاده از طبقه بند شبکه عصبی دارای ۹۲٪ و با درخت تصمیم برابر با ۸۹٪ است [۸].

میرعباسی و برزگری‌نژاد در پژوهشی در زمینه پیش‌بینی طول عمر بیماران مبتلابه سرطان ریه بعد از عمل جراحی با استفاده از تکنیک‌های داده‌کاوی، انجام دادند. هدف از انجام این مطالعه ارائه نتایج عمل جراحی بر میزان بقاء بیماران مبتلابه تومورهای ریوی است که تحت عمل جراحی قرار گرفته بودند. در این پژوهش پیشگویی جهت پیش‌بینی بقاء بیماران مبتلابه سرطان ریه بعد از عمل به کمک الگوریتم‌های شبکه عصبی، درخت تصمیم و K- نزدیک‌ترین همسایه صورت گرفت. نتایج ارزیابی مدل‌های پیشنهادی نشان داد، الگوریتم

از آسیایی‌ها و ۶۷٪ از ساکنان جزایر هاوایی/اقیانوس آرام را دربردارد. بر مبنای اطلاعاتی که در اختیار محققان قرار گرفته، رکوردهای داده SEER از ساختاری مشتمل بر فیلدهای داده حاوی ۱۴۹ ویژگی و بیش از ۵۰۰۰۰۰ رکورد است [۱۲].

شکل ۱ فلوچارت روش پیشنهادی را نشان داد. در ابتدا مرحله پیش‌پردازش و انتخاب ویژگی بر روی داده‌ها انجام شد. داده‌های خام به صورت مقادیر عددی پشت سرهم ذخیره شده‌اند که بر اساس موقعیت‌شان جداسازی شده و عملیات پیش‌پردازش داده‌ها شامل پاک‌سازی داده (حذف داده‌های پرت، نامرتب و گم‌شده)، کاهش داده و تبدیل داده بر روی آن‌ها انجام خواهد گرفت.

جهت پیش‌پردازش داده‌ها از نرم افزار Knime که یکی از نرم افزارهای پر قدرت در حوزه داده کاوی است، استفاده گردید [۱۳]. در ادامه جهت انتخاب ویژگی‌های مؤثرتر در افزایش کارایی مدل با استفاده از عملیات ریاضی، واریانس ویژگی‌ها محاسبه شده و ویژگی‌هایی که دارای واریانس کمتری هستند و باعث ایجاد اختلال و انحراف مدل می‌شوند توسط فیلتر واریانس حذف می‌شوند و در نهایت داده‌ها نرمال‌سازی شدند. نرمال‌سازی داده‌ها عبارت است از روشی که داده‌هایی که در یک دامنه نیستند را در دامنه مشابه قرار می‌دهد. که این عملیات با استفاده از رابطه (۱) انجام می‌شوند [۱۳]:

رابطه ۱

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

که در رابطه فوق مقداری است که باید نرمال شود و کمترین مقدار در آن مجموعه و بیشترین مقدار را بر می‌گرداند.

بهینه‌سازی (Adaptive Neuro-Fuzzy Inference System) ANFIS (Particle Swarm Optimization) PSO

سیستم استنتاج فازی تطبیقی به‌عنوان یک سیستم خود آموزنده با تنظیم‌پذیری پارامترهای سیستم فازی سریع‌تر آموزش دیده و دقت بالایی دارد. از طرفی، شبکه‌های عصبی مصنوعی به دلیل قابلیت آموزش‌پذیری با استفاده از الگوهای مختلف آموزشی می‌تواند ارتباط مناسبی بین متغیرهای ورودی و خروجی ایجاد نماید؛ لذا استفاده ترکیب از سیستم استنتاج فازی و شبکه عصبی مصنوعی به‌عنوان ابزاری قدرتمند که قابلیت پیش‌بینی نتایج با استفاده از داده‌های عددی موجود

درخت تصمیم با دقت ۸۲/۳۵٪ کارایی بهتری نسبت به دو الگوریتم دیگر داشت [۹].

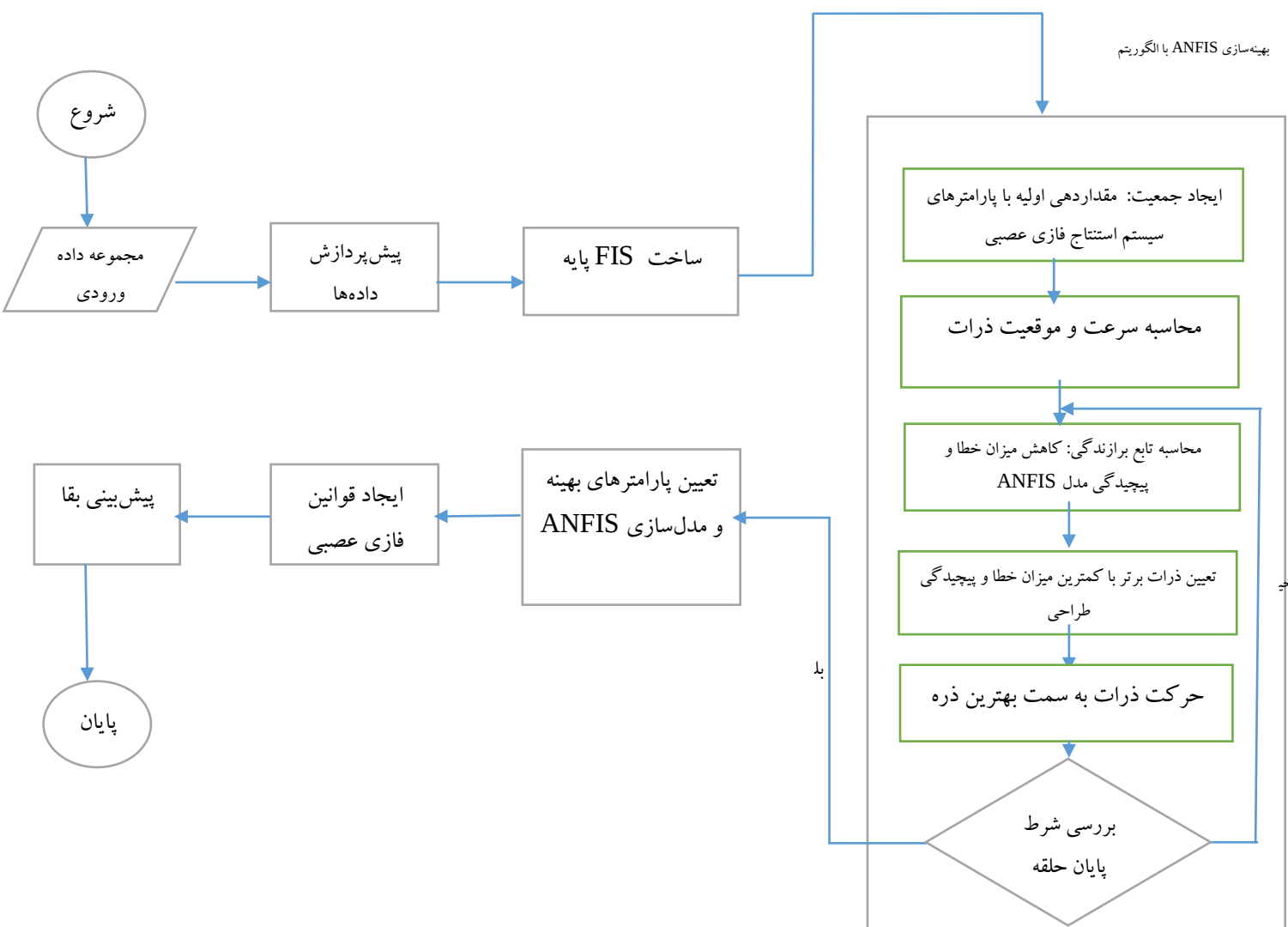
Lynch و همکاران در مطالعه‌ای از تکنیک‌های یادگیری تحت نظارت به پایگاه داده (Epidemiology and End Results Surveillance) SEER برای طبقه‌بندی بیماران مبتلایه سرطان ریه، از جمله رگرسیون خطی، درختان تصمیم‌گیری، ماشین‌های تقویت گرادین (Gradient Boosting Machine) GBM، ماشین بردار پشتیبان (Support Vector Machine) SVM و روش ترکیبی چندگانه استفاده کردند. جهت بهبود پیش‌بینی بقاء ویژگی‌های استفاده شده در این روش‌ها شامل درجه تومور، اندازه تومور، جنسیت، سن، مرحله و تعداد اولیه تومور در نظر گرفته می‌شود. بهترین تکنیک انجام آن روش ترکیبی چندگانه با خطای مربع میانگین ریشه (Root Mean Square Error) RMSE، ۱۵/۰۵ و GBM با مقدار ۱۵/۳۲ RMSE بوده و SVM مقدار RMSE برابر با ۱۵/۸۲ داشته است. هدف نهایی این پژوهش، ارزیابی روش‌های درمانی در راستای افزایش بقاء بیماران و تصمیم‌گیری‌های لازم جهت مراقبت بیمار است [۱۰].

بنای و ساجدی یک مدل کارآمد برای پیش‌بینی بقای افراد مبتلایه سرطان ریه پیشنهاد دادند که از هشت الگوریتم NN, Mark of Bayse, TAN Bayse, Quest, C&R, Quick SEER, Chaid NN Dyna Mic, C5 استفاده کردند. داده‌های استفاده‌شده در این پژوهش از دیتاست SEER جمع‌آوری شد. در روش پیشنهادی داده‌ها زنده ماندن بیماران را بعد از ۶ماه- ۹ ماه، یک سال، دو سال، پنج سال به وسیله ۶۴ ویژگی بررسی کردند. که در این بین، الگوریتم C5 با دقت ۹۷/۹۳ درصد برای کلاس شش ماه، با دقت ۹۶/۹۹۴ درصد برای کلاس ۹ ماه و با دقت ۹۶/۹۱ درصد برای کلاس یک سال و با دقت ۹۶/۳۲ درصد برای کلاس دو سال و دقت ۹۳/۱۲ برای کلاس پنج سال جزء برترین الگوریتم بود [۱۱].

روش

در این مطالعه توصیفی-کاربردی از مجموعه داده SEER استفاده شد. در حال حاضر، SEER داده‌های ابتلاء و بقاء از سرطان را از بین بایگانی‌های مبتنی بر جمعیت جمع‌آوری می‌کند که تقریباً ۳۰٪ از جمعیت آمریکا را شامل می‌شود. حوزه پوشش SEER، ۲۶٪ از آفریقایی-آمریکایی‌ها، ۳۸٪ از اسپانیایی‌ها، ۴۴٪ از آمریکایی-هندی‌ها و بومیان آلاسکا، ۵۰٪

را دارند، تحت عنوان ANFIS معرفی می‌شود [۱۴].



شکل ۱: فلوچارت کلی روند پیشنهادی

پارامترهای توابع عضویت، تعداد لایه‌ها محاسبه کند، در این پژوهش به منظور بهینه‌سازی ANFIS از رویکرد PSO با استفاده از تابع هدف مناسب به منظور کاهش خطا و کاهش پیچیدگی استفاده شد [۱۶].

در ابتدا به منظور طراحی سیستم استنتاج فازی FIS (Fuzzy Inference Systems) اولیه نیاز به ورود داده‌ها به منطق فازی است که با استفاده از روش خوشه‌بندی C میانگین (Fuzzy Clustering - Mean)، این کار صورت خواهد گرفت که این FIS اولیه بهینه نیست و جهت بهینه شدن آن از الگوریتم بهینه‌ساز PSO به صورت زیر استفاده خواهد شد [۱۷].

ANFIS به طور عادی با مقادیر پیش فرض قابل طراحی می‌باشد؛ اما مشکلات زیادی از جمله پیچیدگی زمانی در هنگام اجرا دارد و پیچیدگی شبکه عصبی در حالت عادی بر اساس تعداد لایه‌ها، تعداد نورون‌های موجود در هر لایه به صورت نمایی افزایش می‌یابد. از طرفی در شبکه عصبی فازی تعیین میزان پارامترهای ثابت موجود برای تابع عضویت، تعداد لایه‌ها و نحوه ایجاد شبکه مشکل و دارای پیچیدگی می‌باشد؛ ولی به دلیل قدرت بسیار بالای این الگوریتم که به آن اشاره شد نمی‌توان این الگوریتم را برای مدل‌سازی در داده‌کاوی حذف کرد [۱۵]؛ بنابراین به منظور بهبود آن می‌توان این الگوریتم را باز طراحی کرد و به صورت بهینه از آن استفاده کرد که به صورت خودکار بهترین مقادیر را برای تعیین توابع عضویت،

به نتایج بهتری برسند. در پایان شرط پایان حلقه عدد تکرار یا خطای کمتر از ۰/۰۱٪ بررسی و اگر خاتمه یافت، پارامترهای بهینه تعیین شد.

بر اساس پارامترهای بهینه تعیین شده (توابع عضویت، تعداد پارامترها و درجه توابع عضویت و ...) ANFIS بهینه مدل سازی شده بر روی داده های آموزشی اجرا شد. داده های موجود در مجموعه داده این پژوهش به صورت تصادفی ۷۰٪ داده های آموزش و ۱۰٪ اعتبار سنجی و ۲۰٪ داده های تست استفاده شد. اعتبارسنجی از روش K-Fold (K-لايه) استفاده شد که ابزاری است که نحوه آموزش داده ها را مدیریت می کند. داده های اعتبارسنجی به K زیرمجموعه افراز می شوند و از این زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1 برای آموزش به کار می روند. این روال K بار تکرار شده و همه داده ها یکبار برای آموزش و یکبار برای اعتبارسنجی استفاده می شوند. مقدار K برابر ۱۰ در نظر گرفته شد [۱۹]. داده های آموزشی به عنوان ورودی به ANFIS بهینه شده وارد شده و در نهایت داده های تست به منظور بررسی دقت مدل، استفاده شدند. بر اساس آن معیارهای ارزیابی با استفاده از مقادیر پیش بینی شده و مقادیر واقعی محاسبه شد که نتایج خروجی آن در جداول ۳ و ۴ نشان داده شد. با توجه به این که داده ها از پایگاه داده استاندارد SEER گرفته شد تمامی مجوزهای لازم جهت انجام این مطالعه گرفته شد و ملاحظات اخلاقی در کلیه مراحل اجرایی رعایت گردید.

نتایج

در این مطالعه مجموعه ای از ویژگی ها کلیدی شامل ۱۸ متغیر پیوسته و گسسته که تأثیر بسزایی در پیش بینی بقاء سرطان ریه دارند، انتخاب شده اند (جدول ۱). با استفاده از روش پیشنهادی ارتباط بین ۱۸ ویژگی ورودی و تعداد ماه زنده ماندن ارزیابی شد. داده ها از پایگاه داده SEER بین سال های ۲۰۰۸-۲۰۱۳ جمع آوری گردید که شامل ۱۲۴۶۳ رکورد بود. بازه زنده ماندن بیماران بین ۰-۷۲ ماه بود که توزیع زنده ماندن در شکل ۲ نشان داده شد.

(الف) ایجاد جمعیت اولیه: در این مرحله جمعیت اولیه ایجاد خواهد شد و اندازه هر ذره بر اساس تعداد توابع عضویت فازی و بر اساس تعداد ویژگی ها تعیین می شود. این ذرات به صورت آرایه هایی هستند که شامل توابع عضویت و تعداد لایه های شبکه عصبی می باشند که به صورت تصادفی مقداردهی خواهند شد و بر اساس مرحله اول که فازی سازی اولیه است مقداردهی اولیه به پارامترهای ANFIS انجام می شود.

(ب) محاسبه سرعت و موقعیت ذرات: با توجه به ذرات ایجاد شده، برای هر ذره i مقدار برازندگی fit_i ارزیابی می شود و در صورتی که مقدار برازندگی بهتری حاصل شود بهترین موقعیت ذره، به هنگام می شود. بهترین موقعیت جدید کل گروه پیدا می شود. اگر برازندگی بهترین موقعیت جدید بهتر از گروه قبل است، بهترین موقعیت کل جمعیت، به هنگام می شود.

(ج) محاسبه تابع برازندگی: بر اساس تابع هدف تعیین شده که دقت بالای مدل و کاهش پیچیدگی محاسباتی ANFIS و میزان خطای مدل طراحی شده است ذرات برتر تعیین خواهند شد و بر اساس آن ها جمعیت جدید بر اساس تعداد تکرار ایجاد خواهند شد. رابطه (۲) نشان دهنده تابع برازندگی است [۱۸].

رابطه ۲

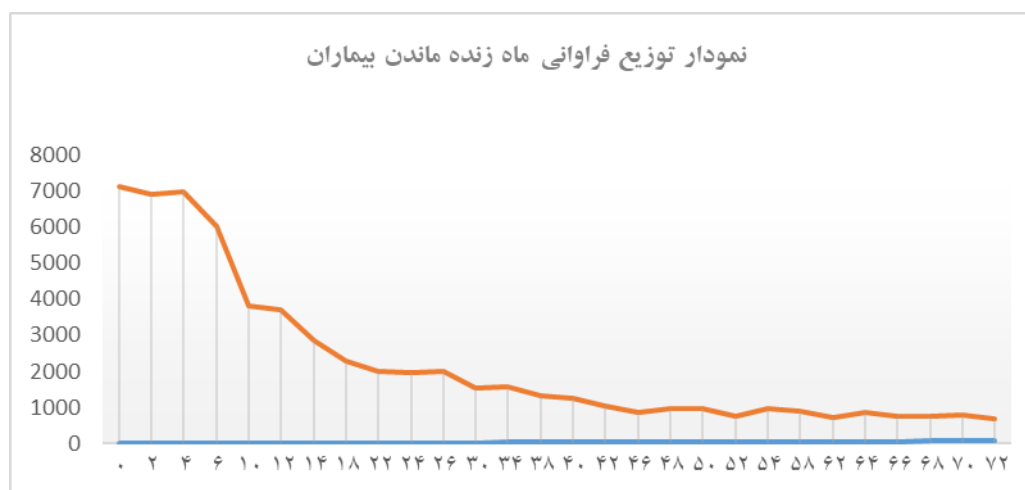
$$Fitness = \alpha(Accuracy) + \beta(ABS((n - s)/n))$$

مقادیر α ، β ثابت و برابر ۱۰۰ در نظر گرفته شد. $Accuracy$ مقایسه بین مقادیر پیش بینی شده توسط سیستم و مقادیر واقعی است. n کل ویژگی و s ویژگی مؤثرتر در پیش بینی است. جمله اول ضربی از دقت و جمله دوم ضربی از کاهش ویژگی می باشد.

(د) حرکت ذرات به سمت بهترین ذره: بر اساس منطق الگوریتم ازدحام ذرات، ذره برتر از جمعیت تعیین خواهد شد و بر اساس آن ذرات دیگر با یک زاویه خاص به ذره برتر نزدیک خواهند شد. ممکن است در این تغییر جهت به سمت ذره برتر

جدول ۱: ویژگی‌های مؤثر استفاده شده در مجموعه داده

ردیف	نام ویژگی	توصیفی برای ویژگی	نوع
۱	سن		عددی
۲	درجه		عددی
۳	توالی پرتودرمانی با جراحی		عددی
۴	دلیل عدم انجام جراحی		عددی
۵	تعداد اولیه		عددی
۶	T	اندازه تومور	عددی
۷	N	درگیری غدد لنفاوی	عددی
۸	M	انتشار تومور به اندام‌های دیگر	عددی
۹	شماره توالی	ترتیب وقوع سرطان ریه با توجه به سایر سرطان‌ها برای این بیمار	عددی
۱۰	مرحله	مرحله تومور - بر اساس T، N و M	عددی
۱۱	پرتودرمانی	آیا بیمار پرتودرمانی داشته یا خیر	عددی
۱۲	سایت اصلی	محل تومور در ریه‌ها	عددی
۱۳	غدد لنفاوی CS	تعداد غدد لنفاوی درگیر	عددی
۱۴	تاریخ تشخیص	زمان تشخیص بیماری	عددی
۱۵	تاریخ تحت نظر	زمانی که بعد از تشخیص بیمار تحت درمان قرار گرفت.	عددی
۱۶	دلیل مرگ غیر از سرطان ریه		عددی
۱۷	دلیل مرگ سرطان ریه		عددی
۱۸	درصد بدخیم یا خوش خیم بودن سرطان		عددی
۱۹	زمان بقا	تعداد ماه‌هایی که بیمار از زمان تشخیص زنده است.	عددی



شکل ۲: توزیع زنده ماندن در بازه ۰-۷۲ ماه

۲- پارامترهای مورد ارزیابی

بعد از اعمال روش پیشنهادی به ویژگی‌های مؤثر، کارایی روش پیشنهادی در جدول ۲ نشان داد. ستون RMSE در جدول ۲ نشان‌دهنده جذر میانگین مربعات خطا بین مقدار پیش‌بینی‌شده

و مقدار واقعی است. ستون انحراف معیار نشان‌دهنده انحراف معیار بین ماه زنده ماندن پیش‌بینی‌شده و واقعی می‌باشد. ستون میانگین همچنین میانگین ماه زنده ماندن پیش‌بینی‌شده توسط روش پیشنهادی را نشان داد.

جدول ۲: نتایج خروجی ANFIS و روش پیشنهادی

معیارهای ارزیابی			
نام الگوریتم	جذر میانگین مربعات خطا	انحراف معیار	میانگین بقا
ANFIS	۵/۰۲۵۳	۶/۴۲۱۱	۲۰/۱۷
ANFIS+PSO	۴/۴۸۲۵	۴/۴۸۵۶	۲۲/۲۳

با توجه به عدم توزیع یکسان سال زنده ماندن بیماران، جهت ارزیابی دقیق تر روش پیشنهادی، بیماران بر اساس میزان سال زنده ماندن به کلاس های یک ساله، دو ساله و پنج ساله طبقه بندی شدند. در کلاس یک ساله بیماران هستند یک سال زنده مانده اند یا فوت شده اند. در کلاس دو ساله بیماران هستند که یک سال تا دو سال زنده مانده اند و یا زیر دو سال فوت شده اند. در کلاس پنج ساله بیماران هستند که یک سال تا پنج سال زنده مانده اند یا زیر پنج سال فوت شده اند (جدول ۳).

جدول ۳: تقسیم کلاس داده ها (بقا)

ویژگی بقا	تقسیم مقادیر داده
کلاس یک ساله	اگر ماه بقا بزرگتر یا مساوی ۱۲ بود بیمار یک سال زنده است. اگر ماه بقا کوچکتر از ۱۲ ماه بود بیمار زیر یک سال فوت کرده است.
کلاس دو ساله	اگر ماه بقا بزرگتر یا مساوی ۲۴ بود بیمار دو سال زنده است. اگر ماه بقا کوچکتر از ۲۴ ماه بود بیمار زیر دو سال فوت کرده است.
کلاس پنج ساله	اگر ماه بقا بزرگتر یا مساوی ۶۰ بود بیمار پنج سال زنده است. اگر ماه بقا کوچکتر از ۶۰ ماه هست بیمار زیر پنج سال فوت کرده است.

نتایج روش پیشنهادی با چند الگوریتم دیگر در حوزه داده کاوی با معیارهای دقت، بازخوانی، صحت، پارامتر F و خطا ارزیابی شد همان طور که در جدول ۴ مشاهده شد روش پیشنهادی نسبت به سایر روش های مورد مقایسه دارای دقت و صحت بیشتری بود.

جدول ۴: مقایسه نتایج در چند الگوریتم داده کاوی با الگوریتم پیشنهادی

فیلد کلاس	معیارهای ارزیابی	الگوریتم های مورد مقایسه				
		ANFIS+PSO	ANFIS	NavieBayes	Random Forest	Logistic J48
بقا یک ساله	دقت	۹۹/۸۰	۹۳/۴۵	۸۵/۶۱	۹۵/۳۵	۹۵/۱
	بازخوانی	۹۵/۵۲	۹۲/۶	۸۷/۷	۹۳/۲۷	۹۷/۴۶
	صحت	۹۹/۶۷	۹۲/۳۴	۸۴/۴۴	۸۵/۷۵	۷۸/۹۴
	معیار F-	۹۷/۵۵	۹۴/۱	۹۳/۷۶	۹۱/۷۸	۹۲/۵۶
بقا دو ساله	دقت	۹۹/۷۴	۹۳/۹۷	۸۶/۶۳	۹۱/۷۷	۹۱/۴۸
	بازخوانی	۹۷/۱	۹۳/۱۲	۹۰/۸۹	۹۳/۰۲	۹۲/۳۲
	صحت	۹۸/۹۷	۹۲/۵۶	۸۶/۱۱	۸۱/۶۹	۷۸/۵۴
	معیار F-	۹۸/۰۲	۹۴/۴۶	۸۷	۸۹/۵	۸۹
بقا پنج ساله	دقت	۹۹/۶۶	۹۱/۸۹	۹۰/۳۱	۸۶/۳۸	۸۲/۴۲
	بازخوانی	۹۶/۰۳	۹۰/۳۲	۷۸/۸	۸۸۶/۲۵	۷۸/۱۹
	صحت	۹۹/۱۱	۹۰/۴۵	۸۷/۸	۸۶/۵۱	۸۳/۰۲
	معیار F-	۹۷/۵۴	۹۰/۱۹	۸۶/۷	۸۷	۷۴

بحث و نتیجه گیری

نتایج پژوهش حاضر به اثربخشی بهینه سازی ANFIS با استفاده از الگوریتم PSO در پیش بینی بقاء بیماران سرطان ریه اشاره دارد. در روش پیشنهادی پس از ورود داده های خام و انجام عملیات پیش پردازش، نرمال سازی داده و انتخاب ویژگی، ANFIS با الگوریتم PSO بهینه سازی شد و ۱۹ ویژگی مؤثرتر انتخاب گردید و سپس با استفاده از طبقه بندی به دو کلاس زنده و فوت شده طبقه بندی شدند. به طور کلی در بهترین حالت، دقت پاسخگویی ANFIS با الگوریتم PSO برای

مجموعه داده SEER در فیلد کلاس یک ساله برابر ۹۹/۸۰٪، فیلد کلاس دوساله برابر ۹۹/۷۴٪ و فیلد کلاس پنج ساله برابر ۹۹/۶۶٪ است. همچنین برای ارزیابی بیشتر، روش پیشنهادی با سایر روش های مشابه که از یک پایگاه داده یکسان استفاده کرده اند مقایسه گردیده است که در جدول ۵ نشان داده شد. چنانچه در این جدول مشاهده شد روش پیشنهادی نسبت به دو روش مورد مقایسه دارای دقت بالاتری بود که دلیل آن استفاده از ترکیب دو روش ANFIS و PSO به صورت هم زمان می توان اشاره کرد.

جدول ۶: مقایسه دقت روش پیشنهادی با سایر روش های مشابه

روش پیشنهادی	Lynch و همکاران [۱۰]	بنای و سجادی [۱۱]	کلاس طبقه بندی داده
۹۹/۸۰	۹۲/۴۷	۹۶/۹۱	یک سال
۹۹/۷۴	۹۱/۹۷	۹۶/۳۲	دو سال
۹۹/۶۶	۹۰/۸۴	۹۳/۱۲	پنج سال

باعث جلوگیری از بیماری و افزایش زمان بقاء بیماران و ارزیابی روش های درمانی گردید. از این رو پیشنهاد می گردد این موضوع در مطالعات آینده مدنظر قرار گیرد. همچنین پیشنهاد می گردد از ویژگی های بیشتری با سایر تکنیک های داده کاوی برای طراحی مدل های پیش بینی استفاده گردد و کارایی آن ها بررسی گردد. با توجه به دقت بالای روش پیشنهادی و نیاز در حوزه پزشکی، توسعه نرم افزاری نتایج تحقیق حاضر به صورت برنامه های کاربردی توصیه می گردد.

تشکر و قدردانی

بدین وسیله از تمامی اساتیدی که در انجام مطالعه حاضر همکاری نمودند، تشکر و قدردانی به عمل می آید.

تعارض منافع

در انجام مطالعه حاضر، نویسندگان هیچ گونه تعارض منافی نداشته اند. این مقاله حاصل پایان نامه کارشناسی ارشد با شماره ۱۹۰۴۱۰۰۹۹۶۲۰۰۴ است که با حمایت دانشگاه آزاد اسلامی واحد مبارکه انجام شده است.

از محدودیت های موجود در مطالعه حاضر عدم دسترسی به داده های واقعی بیماران در ایران می باشد. از طرفی عدم ثبت اطلاعات به صورت دیجیتال، جمع آوری آن ها را با مشکل مواجه کرد که در صورت رفع این محدودیت ها، به طور قطع می توان نتایج بهتری برای بیماران ایرانی به دست آورد.

در این مطالعه، به بهینه سازی ANFIS با الگوریتم PSO پرداخته شد و کارایی آن روی مجموعه داده استاندارد SEER مربوط به بیماران مبتلا به سرطان ریه مورد ارزیابی قرار گرفت. نتایج به دست آمده نشان داد این تکنیک روشی مناسب برای پیش بینی بقا است؛ بنابراین می توان گفت با توجه به حساسیت علم پزشکی در رابطه با حفظ جان انسان و کمبود نیروی انسانی متخصص در نظام سلامت و در راستای مدیریت بهینه منابع، روش پیشنهادی در بهبود ارائه خدمات کمک شایانی می کند و می توان با تشخیص زودتر بیماری هزینه های وارده بر بیمار، بیمارستان و صنعت بیمه را کاهش داد و کمک کرد تا بیمار عمر بیشتر با زندگی مطلوب تری را سپری کند.

با استفاده از ترکیب الگوریتم های فرا ابتکاری دیگر در کنار الگوریتم پیشنهادی می توان اقدام به شناسایی ریسک فاکتورهای تأثیرگذار بر بقای بیماران شد و با تحلیل این عوامل

References

1. Levy A, Hendriks LE, Berghmans T, Faivre-Finn C, GiajLevra M, GiajLevra N, et al. EORTC Lung Cancer Group Survey on the Definition of NSCLC Synchronous Oligometastatic Disease. *Eur J Cancer* 2019;122:109-14. doi: 10.1016/j.ejca.2019.09.012.
2. Fathi-hajabadi F, Farzi S. Prediction of lung cancer in Kermanshah province based on artificial neural network, 1st National Conference of New Outlook on Electrical and Computer Engineering; 2017 Mar 19-20; Kermanshah: Islamic Azad University Kermanshah Branch; 2017. [In Persian]
3. Aliferis CF, Tsamardinos I, Massion PP, Statnikov AR, Fananapazir N, Hardin DP. Machine Learning Models for Classification of Lung Cancer and Selection of Genomic Markers Using Array Gene Expression Data. *American Association for Artificial Intelligence*; 2003.
4. Win SL, Htike ZZ, Yusof F, Noorbatcha IA. Gene expression mining for predicting survivability of patients in early stages of lung cancer. *International Journal on Bioinformatics & Biosciences* 2014;4(2): 1-9. doi:10.5121/ijbb.2014.4201
5. Karhan Z, Tunç T. Lung Cancer Detection and Classification with Classification Algorithms. *IOSR Journal of Computer Engineering (IOSR-JCE)* 2016;18(6):71-7.
6. Yu D, Zhou M, Yang F, Dong D, Gevaert O, Liu Z, et al. Convolutional neural networks for predicting molecular profiles of non-small cell lung cancer. 14th International Symposium on Biomedical Imaging; 2017 Apr 18-21; (Melbourne, VIC, Australia: IEEE; 2017. p. 69-572. doi: 10.1109/ISBI.2017.7950585
7. Murty NR, Babu MP. A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis. *International Journal of Computational Intelligence Research* 2017;13(5):1041-8.
8. Wang Z, Feng F, Zhou X, Duan L, Wang J, Wu Y, et al. Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. *Oncotarget*. 2017;8(55):94793-804. doi: 10.18632/oncotarget.21935.
9. Mirabbasi SE, Barzegarinejad A. Predicting Survival Cancer Patients After Surgery Using Data Mining Techniques. *National Conference on Knowledge and Technology of Engineering Sciences of Iran*; 2017 Mar 10; Tehran: Farzanegan Institute of Higher Education; 2017. [In Persian]
10. Lynch CM, Abdollahi B, Fuqua JD, Alexandra R, Bartholomai JA, Balgmann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform* 2018; 108: 1-8. doi: 10.1016/j.ijmedinf.2017.09.013
11. Banay M, Sajedi H. Predicting Survival of Patients with Lung Cancer Using Classification Algorithms in Data Mining. 2nd National Conference on Computer Science and Information Technology, Najafabad: Islamic Azad University Najafabad Branch; 2015. [In Persian]
12. SEER Research Data Record Description Cases Diagnosed in 1973-2014 [cited 2017 May 20]. Available from: <https://seer.cancer.gov/data-software/documentation/seerstat/nov2016/TextData.FileDescription.pdf>
13. Dormishi A, Ataei M, Khaloo-Kakaie R, Mikaeil R, Shaffiee-Haghshenas S. Performance evaluation of gang saw using hybrid ANFIS-DE and hybrid ANFIS-PSO algorithms. *Journal of Mining and Environment* 2019; 10(2):543-57. doi: 10.22044/JME.2018.6750.1496
14. Esfe MH. Thermal Conductivity Modeling of Aqueous CuO Nanofluids by Adaptive Neuro-Fuzzy Inference System (ANFIS) Using Experimental Data. *Periodica Polytechnica Chemical Engineering* 2018; 62(2): 202-8. doi.org/10.3311/PPch.9670
15. Gholami A, Bonakdari H, Ebtehaj I, Gharabaghi B, Khodashenas SR, Talesh SH, et al. A methodological approach of predicting threshold channel bank profile by multi-objective evolutionary optimization of ANFIS. *Engineering Geology*, 2018: 298-309. doi.org/10.1016/j.enggeo.2018.03.030
16. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. *Transl Lung Cancer Res* 2018;7(3):304-12. doi:10.21037/tlcr.2018.05.15
17. Karaboga D, Kaya E. Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey. *Artificial Intelligence Review* 2018;52(4):1-31. doi: 10.1007/s10462-017-9610-2
18. Jiang L, Huang W, Liu J, Harris K, Yarmus L, Shao W, Chen H, et al. Endosonography With Lymph Node Sampling for Restaging the Mediastinum in Lung Cancer: A Systematic Review and Pooled Data Analysis. *J Thorac Cardiovasc Surg*. 2020;159(3):1099-108.e5. doi: 10.1016/j.jtcvs.2019.07.095.
19. Ahmadi M, Ramezani M, Khorsand R. Diagnosis of Liver Disorders Using a Combination of Adaptive NeuronFuzzy Inference System and Particle Swarm Optimization Algorithm. *Health Inf Manage* 2019; 16(3): 115-21. [In Persian] doi: 10.22122/him.v16i3.3886

Predicting Survival of Patients with Lung Cancer Using Improved Adaptive Neuro-Fuzzy Inference System

Abbasi Ommolbanin¹, Ramezanpour Mohammadreza^{2*}, Khorsand Reihaneh³

• Received: 27 Oct, 2019

• Accepted: 23 Feb, 2020

Introduction: Lung cancer is the main cause of mortality in both genders worldwide. This disease is caused by the uncontrollable growth and development of cells in both or one of the lungs. Although the early diagnosis of this cancer is not an easy task, the earlier it is diagnosed, the higher will be the chance of treating. The objective of this study was to develop an optimized prediction model of the survival of patients with lung cancer based on patients' characteristics through data mining approach.

Method: In this applied-descriptive study, the Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm and the Particle Swarm Optimization (PSO) algorithm were applied to predict the survival rate of patients with lung cancer. The Surveillance, Epidemiology and End-Results (SEER) database of Louisville University, USA was also utilized. The evaluation of this proposed model was conducted based on certain criteria including accuracy, precision, error and root-mean-square error.

Results: The obtained finding indicate the outperformance of ANFIS through PSO algorithm vs. its counterparts in this context with a 99.80 accuracy for one-year survival, 99.74% for two-years and 99.66% for five-years on SEER dataset.

Conclusion: Applying ANFIS through PSO in predicting the survival of patients with lung cancer is a strong measure. Compared with other models, this newly proposed model was of the highest accuracy and precision and of the lowest error rate. Therefore, it is suggested to apply this model for predicting survival of patient.

Keywords: Data Mining, Survival Prediction, Lung Cancer, Survival Rate

• **Citation:** Abbasi O, Ramezanpour MR, Khorsand R. Predicting Survival of Patients with Lung Cancer Using Improved Adaptive Neuro-Fuzzy Inference System. Journal of Health and Biomedical Informatics 2020; 7(1): 19-29. [In Persian]

1. M.Sc. Student in Computer Engineering, Department of Computer Engineering, Mobarakeh Branch, Islamic Azad University, Isfahan, Iran

2. Assistant Professor, Department of Computer Engineering, Mobarakeh Branch, Islamic Azad University, Isfahan, Iran

3. Assistant Professor, Department of Computer Engineering, Dolatabad Branch, Islamic Azad University, Isfahan, Iran

***Corresponding Author:** Mohammadreza Ramezanpour

Address: Mobarakeh Branch, Islamic Azad University, Shahid Nasouhi Blvd, Ferdowsi Square, Moallem Blvd, Mobarakeh, Isfahan

• **Tel:** 031-52492058

• **Email:** mr.ramezanpour@gmail.com