

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

ام البنین عباسی<sup>۱</sup>، محمدرضا رمضان پور<sup>۲\*</sup>، ریحانه خورسند<sup>۳</sup>

• دریافت مقاله: ۱۳۹۸/۸/۵ • پذیرش مقاله: ۱۳۹۸/۱۲/۴

**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 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، ایدمی‌شناسی و نتایج نهایی SEER دانشگاه لوییزول آمریکا استفاده شد. برای ارزیابی روش پیشنهادی از معیارهای دقت، صحت، خطأ 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، ۷۴/۹۹٪ برای بقاء دو‌ساله و ۶۶/۹۹٪ 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 خطأ بوده است؛ بنابراین به کارگیری این مدل درزمینه پیش‌بینی بقاء پیشنهاد می‌شود.

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

**ارجاع:** عباسی ام البنین، رمضان پور محمدرضا، خورسند ریحانه. پیش‌بینی بقاء بیماران سرطان ریه با استفاده از سیستم استنتاج عصبی- فازی تطبیقی بهبودیافته. 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 سرطان‌های ریه سلول کوچک و سلول غیرکوچک هستند. شایع‌ترین عالائم 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 پشتیبانی 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 دیتابست ورودی و همچنین میزان حساسیت بالا در انتخاب ابر صفحه‌های جداکننده با بیشترین میزان حاشیه می‌باشد که باعث می‌شود نقاط مرزی به درستی کلاسه‌بندی نگردد [۳].

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

K-Nearest Neighbor (KNN) 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 موردنرسی 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

Wang و همکاران، به پیش‌بینی نرخ بقاء سرطان ریه با داده‌های ۲۰۰ بیمار مبتلا به سرطان ریه پرداختند. در این پژوهش طبقه‌بندی‌های مورداستفاده درخت تصمیم‌گیری و تکنیک شبکه عصبی است. نتایج به دست آمده با استفاده از طبقه‌بند شبکه عصبی دارای ۹۶٪ و با درخت تصمیم برابر با ۸۹٪ است [۵].

میرعباسی و بروزگرانی نژاد در پژوهشی در زمینه پیش‌بینی طول عمر بیماران مبتلا به 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 مبتلا به تومورهای ریوی است که تحت عمل جراحی قرار گرفته بودند. در این پژوهش پیشگویی جهت پیش‌بینی بقاء بیماران مبتلا به سرطان ریه بعد از عمل به کمک الگوریتم‌های شبکه عصبی، درخت تصمیم و K-نژدیک‌ترین 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 ۵۰۰۰۰ 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 پرقدرت 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.

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

سیستم استنتاج فازی تطبیقی به عنوان 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 به عنوان 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].

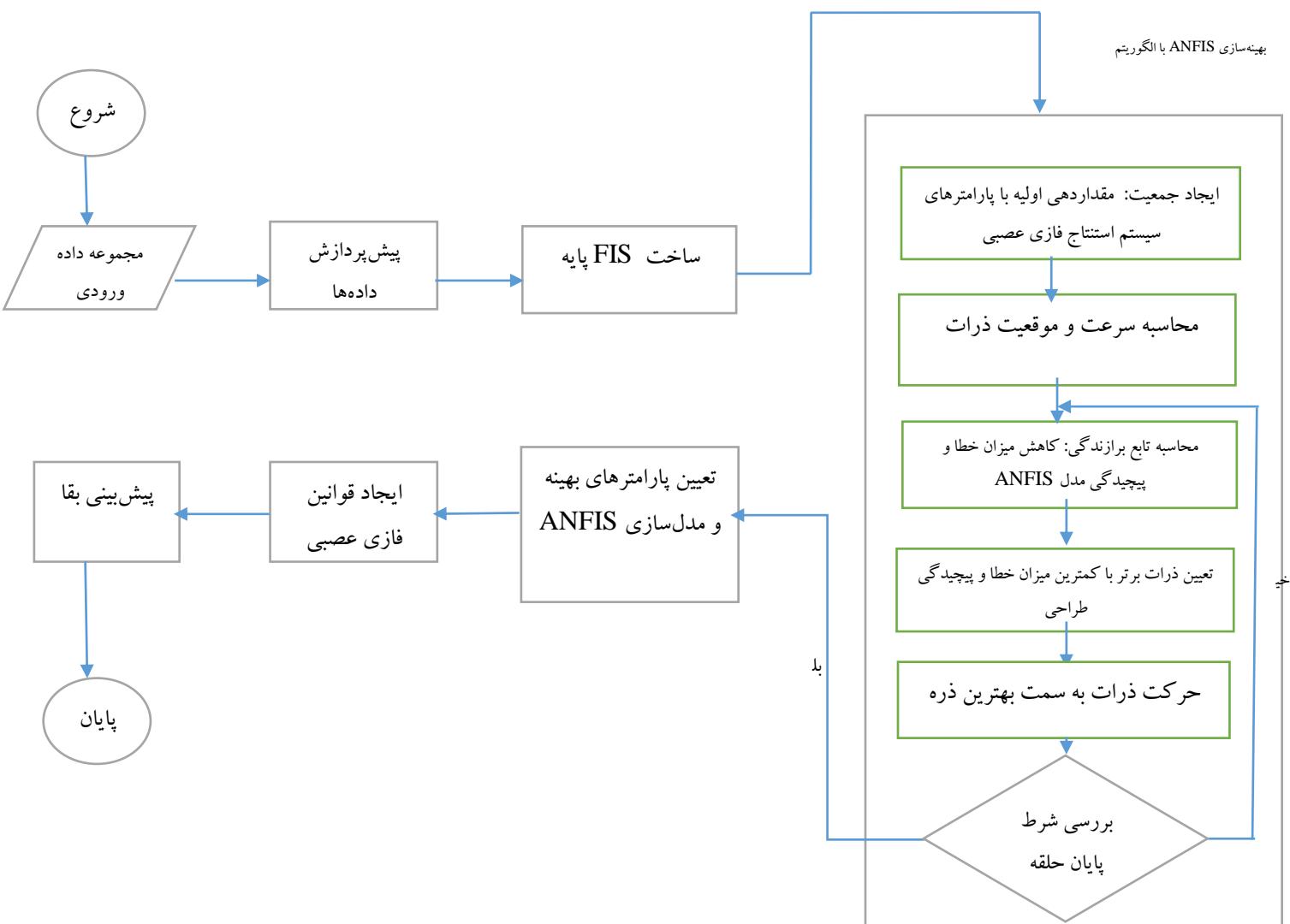
Lynch و همکاران در مطالعه‌ای از تکنیک‌های یادگیری Surveillance تحت نظر از پایگاه داده (SEER Epidemiology and End Results طبقه‌بندی بیماران مبتلا به سرطان ریه، از جمله رگرسیون خطی، درختان تصمیم‌گیری، ماشین‌های تقویت گرادیان GBM (Gradient Boosting Machine)، ماشین بردار پشتیبان SVM (Support Vector Machine) 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 Root Mean Square با خطای مربع میانگین ریشه (RMSE) با مقدار ۱۵/۳۲ RMSE (Error ۱۵/۰۵)، GBM با مقدار ۱۵/۰۵ 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 بقاء افراد مبتلا به سرطان ریه پیشنهاد دادند که از هشت الگوریتم NN, Mark of Bayse, TAN Bayse, Quest, C&R, C5, Chaid NN Dyna Mic, Ouick SEER داده‌های استفاده شده در این پژوهش از دیتا است جمع‌آوری شد. در روش پیشنهادی داده‌ها زنده ماندن بیماران را بعد از ۶۴-۹ ماه، یک سال، دو سال، پنج سال به وسیله ویژگی بررسی گردند. که در این بین، الگوریتم C5 با دقت ۹۶/۹۹۴ درصد برای کلاس شش ماه، با دقت ۹۶/۹۱ درصد برای کلاس ۹ ماه 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، 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-هنگی‌ها و بومیان آلاسکا، ۵۰٪

رادارند، تحت عنوان ANFIS معرفی می‌شود [۱۴].



شکل ۱: فلوچارت کلی روند پیشنهادی

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

در ابتدا بهمنظور طراحی سیستم استنتاج فازی FIS (Fuzzy Inference Systems) (اولیه نیاز به ورود داده‌ها به منطق فازی است که با استفاده از روش خوشبندی C میانگین FCM(Fuzzy Clustering -Mean) 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 پیش‌فرض 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 بهمنظور 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  $0.01\%$  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 داده‌های آموزشی اجرا شد. داده‌های موجود در مجموعه داده این پژوهش به صورت تصادفی  $70\%$  داده‌های آموزش و  $10\%$  اعتبار سنجی و  $20\%$  داده‌های تست استفاده شد. اعتبار سنجی از روش K-Fold استفاده شد که ابزاری است که نحوه آموزش داده‌ها را مدیریت می‌کند. داده‌های اعتبار سنجی به K زیر مجموعه افزار می‌شوند و از این زیر مجموعه، هر بار یکی برای اعتبار سنجی و K-1 برای آموزش به کار می‌روند. این روال K بار تکرار شده و همه داده‌ها یکبار برای آموزش و یکبار برای اعتبار سنجی استفاده می‌شوند. مقدار K برابر  $10$  در نظر گرفته شد [۱۹]. داده‌های آموزشی به عنوان ورودی به ANFIS بهینه شده وارد شده و در نهایت داده‌های تست به منظور بررسی دقیق استفاده از مقادیر شدن. بر اساس آن معیارهای ارزیابی با استفاده از مقادیر پیش‌بینی شده و مقادیر واقعی محاسبه شد که نتایج خروجی آن در جداول ۳ و ۴ نشان داده شد. با توجه به این که داده‌ها 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 مجموعه‌ای از ویژگی‌ها کلیدی شامل ۱۸ متغیر پیوسته و گسسته که تأثیر بسزایی در پیش‌بینی بقاء سرطان ریه دارند، انتخاب شده‌اند (جدول ۱). با استفاده از روش پیشنهادی ارتباط بین ۱۸ ویژگی ورودی و تعداد ماه زنده ماندن ارزیابی شد. داده‌ها از پایگاه داده SEER بین سال‌های ۲۰۰۸-۲۰۱۳ جمع‌آوری 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 ویژگی‌ها تعیین می‌شود. این ذرات به صورت آرایه‌هایی هستند که شامل توابع عضویت و تعداد لایه‌های شبکه عصبی می‌باشند که به صورت تصادفی مقداردهی خواهد شد و بر اساس مرحله اول که فازی سازی اولیه است مقداردهی اولیه به پارامترهای ANFIS انجام می‌شود.

**(ب) محاسبه سرعت و موقعیت ذرات:** با توجه به ذرات ایجاد شده، برای هر ذره  $i$  مقدار برازنده  $\alpha_i$  ارزیابی می‌شود و در صورتی که مقدار برازنده  $i$  بهتری حاصل شود بهترین موقعیت ذره، به نگام می‌شود. بهترین موقعیت جدید کل گروه پیدا می‌شود. اگر برازنده  $i$  بهترین موقعیت جدید بهتر از گروه قبل است، بهترین موقعیت کل جمعیت، به نگام می‌شود.

**(ج) محاسبه تابع برازنده:** 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 خواهد شد و بر اساس آن‌ها جمعیت جدید بر اساس تعداد تکرار ایجاد خواهد شد. رابطه (۲) نشان‌دهنده تابع برازنده است [۱۸].

رابطه ۲

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

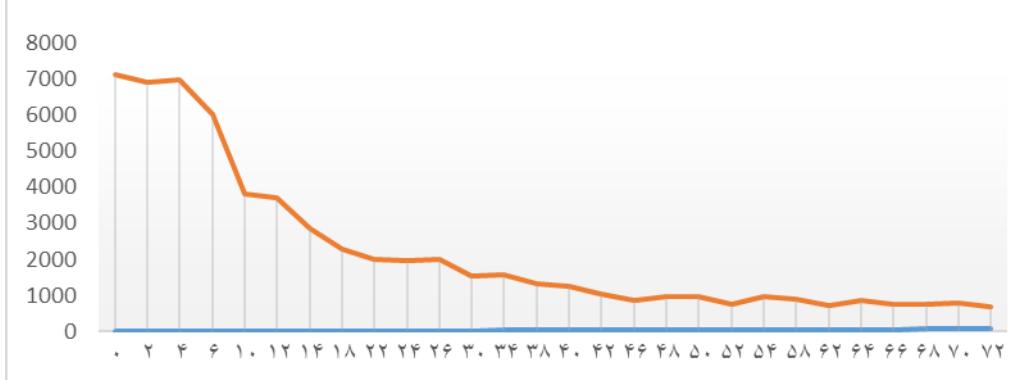
مقادیر  $\alpha$  و  $\beta$  ثابت و برابر  $100$  در نظر گرفته شد. Accuracy مقایسه بین مقادیر پیش‌بینی شده توسط سیستم و مقادیر واقعی است.  $n$  کل ویژگی و  $s$  ویژگی مؤثرتر در پیش‌بینی است. جمله اول ضریبی از دقت و جمله دوم ضریبی از کاهش ویژگی می‌باشد.

**(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 خواهد شد. ممکن است در این تغییر جهت به سمت ذره برتر

جدول ۱: ویژگی‌های مؤثر استفاده شده در مجموعه داده

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

نمودار توزیع فراوانی ماه زنده ماندن بیماران



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

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

## 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 خطأ بین مقدار پیش‌بینی شده

جدول ۲: نتایج خروجی ANFIS و روش پیشنهادی

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

زنده مانده‌اند یا فوت شده‌اند. در کلاس دو ساله بیمارانی هستند که یک سال تا دو سال زنده مانده‌اند و یا زیر دو سال فوت شده‌اند. در کلاس پنج ساله بیمارانی هستند که یک سال تا پنج سال زنده مانده‌اند یا زیر پنج سال فوت شده‌اند (جدول ۳).

با توجه به عدم توزیع یکسان سال زنده ماندن بیماران، جهت ارزیابی دقیق‌تر روش پیشنهادی، بیماران بر اساس میزان سال زنده ماندن به کلاس‌های یک‌ساله، دو‌ساله و پنج‌ساله طبقه‌بندی شدند. در کلاس یک‌ساله بیمارانی هستند یک سال

جدول ۳: تقسیم کلاس داده‌ها(بقا)

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

نتایج روش پیشنهادی با چند الگوریتم دیگر در حوزه داده‌کاوی با معیارهای دقت، بازخوانی، صحت، پارامتر F و خطای ارزیابی بیشتری بود.

نتایج روش پیشنهادی با چند الگوریتم دیگر در حوزه داده‌کاوی با معیارهای دقت، بازخوانی، صحت، پارامتر F و خطای ارزیابی شد همان‌طور که در جدول ۴ مشاهده شد روش پیشنهادی

جدول ۴: مقایسه نتایج در چند الگوریتم داده‌کاوی با الگوریتم پیشنهادی

فیلد کلاس						معیارهای ارزیابی
الگوریتم‌های مورد مقایسه						
ANFIS+PSO	ANFIS	NavieBayes	Random Forest	Logistic	J48	
۹۹/۸۰	۹۳/۴۵	۸۵,۶۱	۹۵/۳۵	۹۵/۱	۹۲/۳۲	دقت
۹۵/۵۲	۹۲/۶	۸۷/۷	۹۳/۲۷	۹۷/۴۶	۹۳/۰۹	بقاء یک‌ساله
۹۹/۶۷	۹۲/۳۴	۸۴/۴۴	۸۵/۷۵	۷۸/۹۴	۸۷/۲۶	بازخوانی
۹۷/۵۵	۹۴/۱	۹۳/۷۶	۹۱/۷۸	۹۲/۵۶	۹۰/۴۵	صحت
۹۹/۷۴	۹۳/۹۷	۸۶/۶۳	۹۱/۷۷	۹۱/۴۸	۹۱/۶۵	F- معيار
۹۷/۱	۹۳/۱۲	۹۰/۸۹	۹۳/۰۲	۹۲/۳۲	۸۷/۴۱	دقت
۹۸/۹۷	۹۲/۵۶	۸۶/۱۱	۸۱/۶۹	۷۸/۵۴	۷۳/۹۴	بازخوانی دو ساله
۹۸/۰۲	۹۴/۴۶	۸۷	۸۹/۵	۸۹	۸۷	صحت
۹۹/۶۶	۹۱/۸۹	۹۰/۳۱	۸۶/۳۸	۸۲/۴۲	۷۶	F- معيار
۹۶/۰۳	۹۰/۳۲	۷۸,۸	۸۸/۲۵	۷۸/۱۹	۸۲	دقت
۹۹,۱۱	۹۰/۴۵	۸۷/۸	۸۶/۵۱	۸۳/۰۲	۶۴	بازخوانی پنج ساله
۹۷,۵۴	۹۰/۱۹	۸۶/۷	۸۷	۷۴	۸۲	صحت

مجموعه داده SEER در فیلد کلاس یک ساله برابر ۹۹/۸۰٪، فیلد کلاس دو ساله برابر ۹۹/۷۴٪ و فیلد کلاس پنج ساله برابر ۶۶/۹۹٪ است. همچنین برای ارزیابی بیشتر، روش پیشنهادی با سایر روش‌های مشابه که از یک پایگاه داده یکسان استفاده کرده‌اند 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 همزمان می‌توان اشاره کرد.

## بحث و نتیجه گیری

نتایج پژوهش حاضر به اثربخشی بهینه‌سازی ANFIS با استفاده از الگوریتم PSO در پیش‌بینی بقاء بیماران سرطان ریه اشاره دارد. درروش پیشنهادی پس از ورود داده‌های خام و انجام عملیات پیش‌پردازش، نرم‌افزاری داده و انتخاب ویژگی، ANFIS با الگوریتم PSO بهینه‌سازی شد و ۹۱٪ دقت پیشنهادی از طبقه‌بندی به دو کلاس زنده و فوت‌شده طبقه‌بندی شدند. به طور کلی در بهترین حالت، دقت پاسخگویی ANFIS با الگوریتم PSO برای

جدول ۶: مقایسه دقت روش پیشنهادی با سایر روش‌های مشابه

روش پیشنهادی	Lynch و همکاران [۱۰]	بنای و سجادی [۱۱]	کلاس طبقه‌بندی 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 روش‌های درمانی گردید. از این‌رو پیشنهاد می‌گردد این موضوع در مطالعات آینده مدنظر قرار گیرد. همچنین پیشنهاد می‌گردد از ویژگی‌های بیشتری با سایر تکنیک‌های داده‌کاوی برای طراحی مدل‌های پیش‌بینی استفاده گردد و کارایی آن‌ها بررسی گردد. با توجه به دقت بالای روش پیشنهادی و نیاز در حوزه پژوهشی، توسعه نرم‌افزاری نتایج تحقیق حاضر به صورت برنامه‌های کاربردی توصیه می‌گردد.

## تشکر و قدردانی

بدین‌وسیله از تمامی اساتیدی که در انجام مطالعه حاضر همکاری نمودند، تشکر و قدردانی به عمل می‌آید.

## تعارض منافع

در انجام مطالعه حاضر، نویسنده‌گان هیچ‌گونه تعارض منافعی نداشته‌اند. این مقاله حاصل پایان‌نامه کارشناسی ارشد با شماره ۱۹۰۴۱۰۰۹۹۶۲۰۰۴ است که با حمایت دانشگاه آزاد اسلامی واحد مبارکه انجام شده است.

از محدودیت‌های موجود در مطالعه حاضر عدم دسترسی به داده‌های واقعی بیماران در ایران می‌باشد. از طرفی عدم ثبت اطلاعات به صورت دیجیتال، جمع‌آوری آن‌ها را با مشکل مواجه کرد که در صورت رفع این محدودیت‌ها، به طور قطع می‌توان نتایج بهتری برای بیماران ایرانی به دست آورد.

در این مطالعه، به بهینه‌سازی ANFIS با الگوریتم SEER پرداخته شد و کارایی آن روی مجموعه داده استاندارد مربوط به بیماران مبتلا به سرطان ریه مورد ارزیابی قرار گرفت. نتایج به دست آمده نشان داد این تکنیک روشی مناسب برای پیش‌بینی بقا است؛ بنابراین می‌توان 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 می‌کند و می‌توان با تشخیص زودتر بیماری هزینه‌های واردہ بر بیمار، بیمارستان و صنعت بیمه را کاهش داد و کمک کرد تا بیمار عمر بیشتر بازندگی مطلوب‌تری را سپری کند.

با استفاده از ترکیب الگوریتم‌های فرا ابتکاری دیگر در کنار الگوریتم پیشنهادی می‌توان 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-hajibadi 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, Balgemann 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, Ramezanpour 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 Ommolbanin<sup>1</sup>, Ramezanpour Mohammadreza<sup>2\*</sup>, Khorsand Reihaneh<sup>3</sup>

• 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.ramezanpoor@gmail.com