مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره اول. ص ۲۹-۱۹، ۱۳۹۹

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

ام‌البنین عباسی[۱]، محمدرضا رمضان پور[۲]*، ریحانه خورسند[۳]

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

مقدمه: سرطان ریه منبع اصلی مرگ‌ومیر برای مردان و زنان در سراسر جهان می‌باشد. بیماری ریه توسعه و رشد غیرقابل‌کنترل سلول‌ها در یک یا هر دو ریه می‌باشد. تشخیص زودرس سرطان آسان نیست؛ اما اگر سریع تشخیص داده شود، قابل‌درمان است. هدف از این مطالعه، ساخت مدل بهینه پیش‌بینی کننده بقاء بیماران مبتلابه 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 مبتلابه سرطان ریه استفاده شد. در این مطالعه، از پایگاه داده معتبر برنامه نظارت، اپیدمی‌شناسی و نتایج نهایی SEER دانشگاه لوییزول آمریکا استفاده شد. برای ارزیابی روش پیشنهادی از معیارهای دقت، صحت، خطا و جذر خطای میانگین مربعات استفاده شد.

نتایج: نتایج نهایی به دست‌آمده در این مطالعه نشان‌دهنده برتری روش بهینه‌سازی ANFIS با الگوریتم PSO نسبت به سایر روش‌ها، در راستای پیش‌بینی بقاء بیماران مبتلابه سرطان ریه با متوسط صحت برابر ۹۹/۸۰٪ برای بقاء یک‌ساله، ۹۹/۷۴٪ برای بقاء دوساله و ۹۹٪/۶۶ برای بقا 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 به‌کارگیری ایـن مـدل درزمینه پیش‌بینی بقا پیشنهاد می‌شود.

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

•ارجاع: عباسی ام‌البنین، 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
•شماره تماس: ۰۳۱۵۲۴۹۲۰۵۸
• **Email:** mr.ramezanpoor@gmail.com

مقدمه

سرطان ریه نوعی بیماری است که مشخصه آن رشد کنترل‌نشده 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 به‌منظور 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 به‌درستی کلاسه‌بندی نگردند [۳].

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

Karhan و Tunç، از مدل K- نزدیک‌ترین همسایه (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 موردبررسی 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 درصورتی‌که مقدار k کم باشد نمونه‌ها به‌صورت 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 تجزیه‌وتحلیل با استفاده از الگوریتم‌های طبقه‌بندی و بعدازآن همان روش در ۹۶ نمونه و ۷۱۳۰ ویژگی برای پیش‌بینی سرطان ریه صورت گرفته است. مجموعه داده‌های مورداستفاده 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 میشیگان 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 مورداستفاده درخت تصمیم‌گیری و تکنیک شبکه عصبی است. نتایج به‌دست‌آمده 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

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 (Surveillance SEER (Epidemiology and End Results برای طبقه‌بندی بیماران مبتلابه سرطان ریه، از جمله رگرسیون خطی، درختان تصمیم‌گیری، ماشین‌های تقویت گرادیان (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 RMSE ۱۵/۳۲ 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 مبتلابه سرطان ریه پیشنهاد دادند که از هشت الگوریتم NN, Mark of Bayse, TAN Bayse, Quest, C&R , C5, Chaid NN Dyna Mic , Ouick استفاده کردند. داده‌های استفاده‌شده در این پژوهش از دیتاست SEER جمع‌آوری شد. درروش 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٪ از آسیایی‌ها و ۶۷٪ از ساکنان جزایر هاوایی/اقیانوس آرام را دربردارد. بر مبنای اطلاعاتی که در اختیار محققان قرارگرفته، رکوردهای دادهSEER 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 روی‌داده‌ها انجام شد. داده‌های خام به‌صورت مقادیر عددی پشت سرهم ذخیره‌شده‌اند که بر اساس موقعیت‌شان جداسازی شده و عملیات پیش‌پردازش داده‌ها شامل پاک‌سازی داده(حذف داده‌های پرت، نامرتبط 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 محاسبه‌شده 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

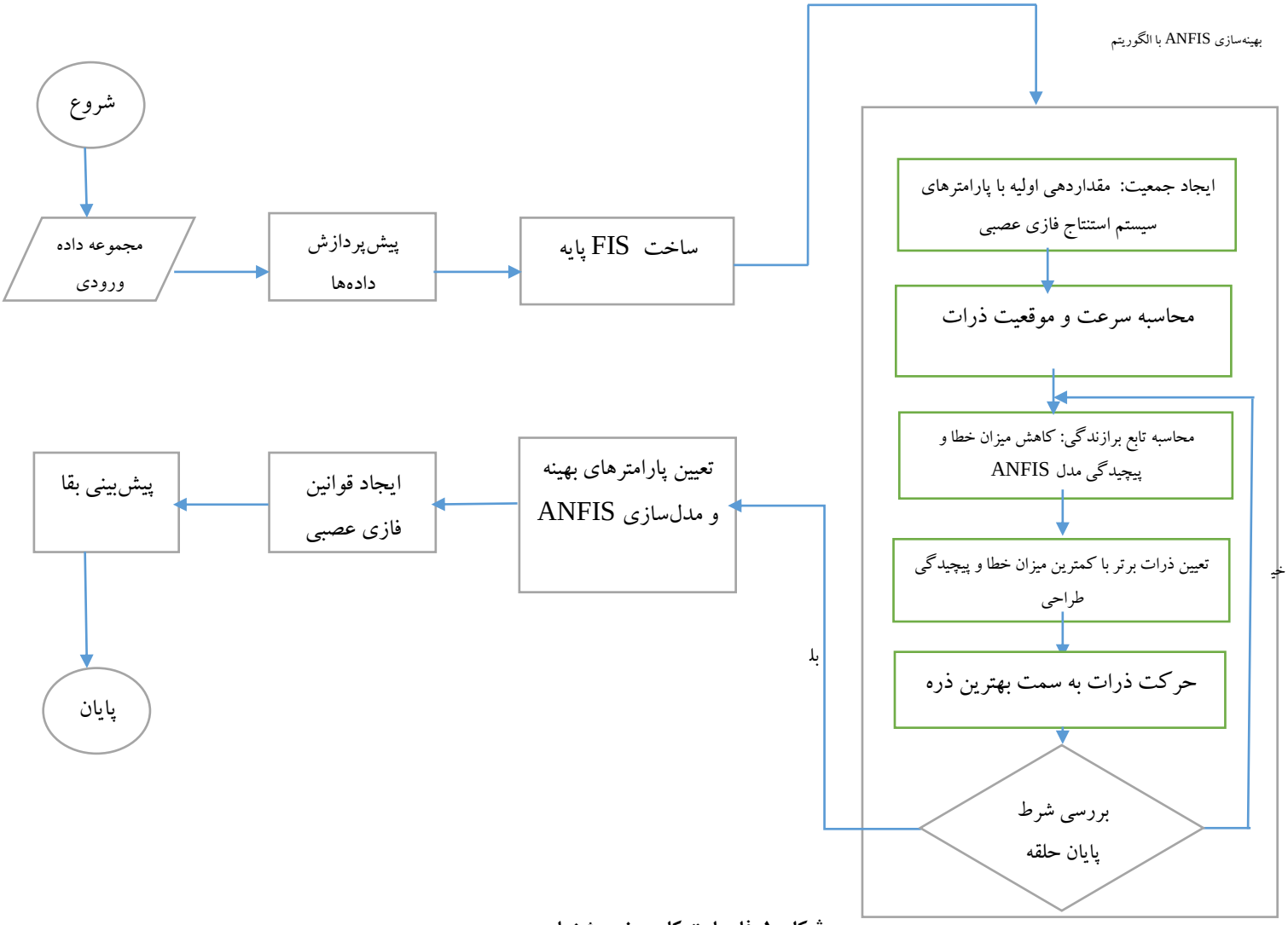
$$Z = \frac{x - min(x)}{max(x) - min(x)}$$

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 با الگوریتم (Particle Swarm Optimization) PSO

سیستم استنتاج فازی تطبیقی به‌عنوان یک سیستم خود آموزنده با تنظیم‌پذیری پارامترهای سیستم فازی سریع‌تر آموزش‌دیده 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

رادارند، تحت عنوان ANFIS معرفی می‌شود [۱۴].



شکل ۱: فلوچارت کلی روند پیشنهادی

ANFIS به‌طور عادی با مقادیر پیش‌فرض قابل طراحی می‌باشد؛ اما مشکلات زیادی ازجمله پیچیدگی زمانی در هنگام اجرا دارد و پیچیدگی شبکه عصبی در حالت عادی بر اساس تعداد لایه‌ها، تعداد نورون‌های موجود در هر لایه به‌صورت 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 به‌صورت بهینه از آن استفاده کرد که به‌صورت خودکار بهترین مقادیر را برای تعیین توابع عضویت، پارامترهای توابع عضویت، تعداد لایه‌ها محاسبه کند، در این پژوهش به‌منظور بهینه‌سازی ANFIS از رویکرد PSO با استفاده از تابع هدف مناسب به‌منظور کاهش خطا 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 خوشه‌بندی c میانگین FCM(Fuzzy Clustring -Mean)، این کار صورت خواهد گرفت که این FIS اولیه بهینه نیست و جهت بهینه شدن آن از الگوریتم بهینه‌ساز PSO به‌صورت زیر استفاده خواهد شد [۱۷].

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 به‌صورت آرایه‌هایی هستند که شامل توابع عضویت و تعداد لایه‌های شبکه عصبی می‌باشند که به‌صورت تصادفی مقداردهی خواهند شد و بر اساس مرحله اول که فازی‌سازی اولیه است مقداردهی اولیه به پارامترهای ANFIS انجام می‌شود.

ب) محاسبه سرعت و موقعیت ذرات: با توجه به ذرات ایجادشده، برای هر ذره i مقدار برازندگی x_i ارزیابی می‌شود و درصورتی‌که مقدار برازندگی بهتری حاصل شود بهترین موقعیت ذره، به‌هنگام می‌شود. بهترین موقعیت جدید کل گروه پیدا می‌شود. اگر برازندگی بهترین موقعیت جدید بهتر از گروه قبل است، بهترین موقعیت کل جمعیت، به‌هنگام می‌شود.

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

رابطه ۲

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

مقادیر α, β ثابت و برابر ۱۰۰ در نظر گرفته شد. $Accuracy$ مقایسه بین مقادیر پیش‌بینی‌شده 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 خواهند شد. ممکن است در این تغییر جهت به سمت ذره برتر 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.

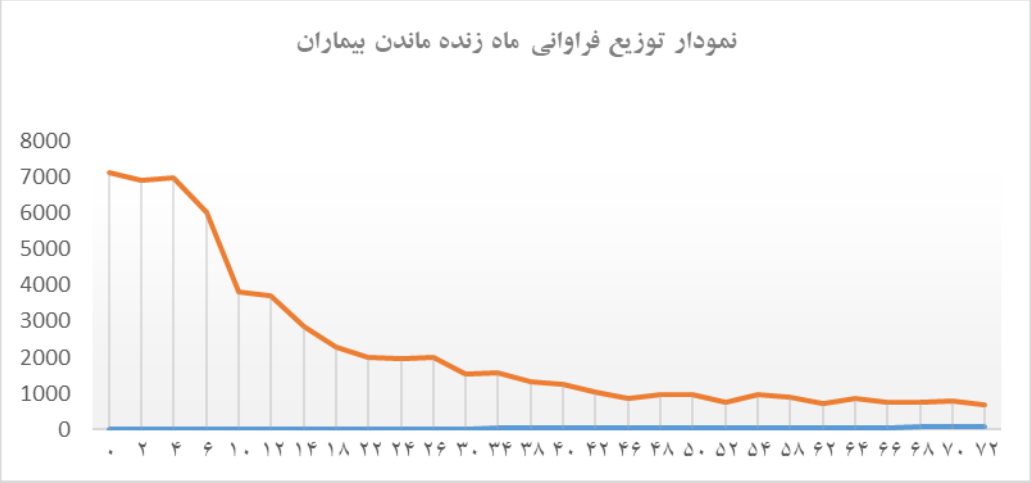
بر اساس پارامترهای بهینه تعیین‌شده (توابع عضویت، تعداد پارامترها و درجه توابع عضویت و ...) ANFIS بهینه مدل‌سازی شده بر روی‌داده‌های آموزشی اجرا شد. داده‌های موجود در مجموعه داده این پژوهش به‌صورت تصادفی ۷۰٪ داده‌های آموزش و ۱۰٪ اعتبار سنجی و ۲۰٪ داده‌های تست استفاده شد. اعتبارسنجی از روش K-لایه (K-Fold) استفاده شد که ابزاری است که نحوه آموزش داده‌ها را مدیریت می‌کند. داده‌های اعتبارسنجی به K زیرمجموعه افراز می‌شوند و از این زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1 برای آموزش به کار می‌روند. این روال K بار تکرار شده و همه داده‌ها یک‌بار برای آموزش و یک‌بار برای اعتبارسنجی استفاده می‌شوند. مقدار K برابر ۱۰ در نظر گرفته شد [۱۹]. داده‌های آموزشی به‌عنوان ورودی به ANFIS بهینه‌شده وارد شده و در نهایت داده‌های تست به‌منظور بررسی دقت مدل، استفاده شدند. بر اساس آن معیارهای ارزیابی با استفاده از مقادیر پیش‌بینی‌شده 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 ۲۰۱۳-۲۰۰۸ جمع‌آوری گردید که شامل ۱۲۴۶۳ رکورد بود. بازه زنده ماندن بیماران بین ۷۲-۰ ماه بود که توزیع زنده ماندن در شکل ۲ نشان داده شد.

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 خوش‌خیم 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

نام الگوریتم	معیارهای ارزیابی		
	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 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 زنده مانده‌اند یا فوت‌شده‌اند. در کلاس دوساله بیمارانی هستند که یک سال تا دو سال زنده مانده‌اند و یا زیر دو سال فوت‌شده‌اند. در کلاس پنج‌ساله بیمارانی هستند که یک سال تا پنج سال زنده مانده‌اند یا زیر پنج سال فوت‌شده‌اند (جدول ۳).

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. اگر ماه بقاء کوچک‌تر از ۱۲ ماه بود بیمار زیر یک سال فوت کرده است.
کلاس دوساله	اگر ماه بقاء بزرگ‌تر یا مساوی ۲۴ بود بیمار دو سال زنده است. اگر ماه بقاء کوچک‌تر از ۲۴ ماه بود بیمار زیر دو سال فوت کرده است.
کلاس پنج‌ساله	اگر ماه بقاء بزرگ‌تر یا مساوی ۶۰ بود بیمار پنج سال زنده است. 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 خطا ارزیابی شد همان‌طور که در جدول ۴ مشاهده شد روش پیشنهادی نسبت به سایر روش‌های مورد مقایسه دارای دقت و صحت بیشتری بود.

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

فیلد کلاس	معیارهای ارزیابی	الگوریتم‌های مورد مقایسه					
		J48	Logistic	Random Forest	NavieBayes	ANFIS	ANFIS+PSO
بقاء یک‌ساله	دقت	۹۲/۳۲	۹۵/ ۱	۹۵/۳۵	۸۵,۶۱	۹۳/۴۵	۹۹/۸۰
	بازخوانی	۹۳/۰۹	۹۷/۴۶	۹۳/۲۷	۸۷/۷	۹۲/۶	۹۵/۵۲
	صحت	۸۷/۲۶	۷۸/۹۴	۸۵/۷۵	۸۴/۴۴	۹۲/۳۴	۹۹/۶۷
	معیار F-	۹۰/۴۵	۹۲/۵۶	۹۱/۷۸	۹۳/۷۶	۹۴/۱	۹۷/۵۵
بقاء دوساله	دقت	۹۱/۶۵	۹۱/۴۸	۹۱/۷۷	۸۶/۶۳	۹۳,۹۷	۹۹/۷۴
	بازخوانی	۸۷/۴۱	۹۲/۳۲	۹۳/۰۲	۹۰/۸۹	۹۳/۱۲	۹۷/۱
	صحت	۷۳/۹۴	۷۸/۵۴	۸۱/۶۹	۸۶/۱۱	۹۲/۵۶	۹۸/۹۷
	معیار F-	۸۷	۸۹	۸۹/۵	۸۷	۹۴/۴۶	۹۸/۰۲
بقاء پنج‌ساله	دقت	۷۶	۸۲/۴۲	۸۶/۳۸	۹۰/۳۱	۹۱/۸۹	۹۹/۶۶
	بازخوانی	۸۲	۷۸/۱۹	۸۸۶/۲۵	۷۸,۸	۹۰/۳۲	۹۶/۰۳
	صحت	۶۴	۸۳/۰۲	۸۶/۵۱	۸۷/۸	۹۰/۴۵	۹۹,۱۱
	معیار F-	۸۲	۷۴	۸۷	۸۶/۷	۹۰/۱۹	۹۷,۵۴

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، نرمال‌سازی داده و انتخاب ویژگی، ANFIS با الگوریتم PSO بهینه‌سازی شد و ۱۹ویژگی مؤثرتر انتخاب گردید و سپس با استفاده از طبقه‌بندی به دو کلاس زنده و فوت‌شده طبقه‌بندی شدند. به‌طورکلی در بهترین حالت، دقت پاسخگویی ANFIS با الگوریتم PSO برای مجموعه داده SEER در فیلد کلاس یک‌ساله برابر ۹۹/۸۰٪، فیلد کلاس دوساله برابر ۹۹/۷۴٪ و فیلد کلاس پنج‌ساله برابر۹۹/۶۶٪ 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.

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

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

در این مطالعه، به بهینه‌سازی ANFIS با الگوریتم PSO پرداخته شد و کارایی آن روی مجموعه داده استاندارد SEER مربوط به بیماران مبتلابه سرطان ریه مورد ارزیابی قرار گرفت. نتایج به‌دست‌آمده نشان داد این تکنیک روشی مناسب برای پیش‌بینی بقاست؛ 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 وارده 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 باعث جلوگیری از بیماری و افزایش زمان بقاء بیماران و ارزیابی روش‌های درمانی گردید. ازاین‌رو 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 به‌دقت بالای روش پیشنهادی و نیاز در حوزه پزشکی، توسعه نرم‌افزاری نتایج تحقیق حاضر به‌صورت 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، نویسندگان 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.

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-hajiabadi 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

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 7(1): 19-29

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

Abbasi Ommolbanin[1], Ramezanpour Mohammadreza[2]*, Khorsand Reihaneh[3]

• **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