
بسم الله الرحمن الرحیم

دانشگاه قم
دانشکده فنی و مهندسی
پایان‌نامه دوره کارشناسی ارشد مهندسی فناوری اطلاعات

عنوان:

استفاده از داده‌کاوی برای کشف الگوهای پنهان در داده‌های سرطان سینه (بیمارستان شهید رجایی آران و بیدگل)

استاد راهنما:
دکتر بهروز مینایی بیدگلی

استاد مشاور:
دکتر فریبا مینایی بیدگلی

نگارنده:
فاطمه السادات طباطبائی عینکی

تابستان/ ۱۳۹۰

«صورت جلسه دفاع از پایان نامه کارشناسی ارشد»

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر(عجل الله تعالی فرجه الشریف)

جلسه دفاعیه پایان نامه کارشناسی ارشد آقای/ خانم: رشته:

تحت عنوان: .....

با حضور هیأت داوران در محل دانشگاه قم در تاریخ: / / تشکیل گردید.

در این جلسه، پایان نامه با نمره (به عدد ..... ، بـه حـروف .....) و

با درجه: عالی ○ بسیار خوب ○ خوب ○ قابل قبول ○ مورد دفاع قرار گرفت.

نام و نام خانوادگی	سمت	مرتبه علمی	امضاء
	استاد راهنما		
	استاد مشاور		
	استاد ناظر		
	استاد ناظر		
	نماینده کمیته تحصیلات تکمیلی		

معاون آموزشی و پژوهشی دانشکده
نام و امضاء

مدیر امور آموزش و تحصیلات تکمیلی
نام و امضاء

تقدیم به:

ساحت مقدس امام زمان روحی فداه

و همچنین

تقدیم به روح پدر بزرگوارم
که الطافش همیشه شامل حالم بوده و هست،

مادر مهربانم
که همواره بر مشکلاتم صبوری کرد،

و همسر فداکارم
که با گذشت و ایثارش مرا یاری نمود

تشکر و قدردانی:

خداوند منان را شاکرم که این عبد ناچیز و حقیرش را توفیق تحصیل علم داد، و تا این مراحل و مدارج علمی رساند، و نیز از اساتید محترم دکتر بهروز مینایی استاد راهنما و دکتر فریبا مینایی استاد مشاور که بنده را در خلق این اثر کمک و راهنمایی کردند، و دیگر اساتید دوران کارشناسی ارشد که با مشاوره‌های خویش بنده را مساعدت نمودند، کمال تشکر و سپاس را دارم و برای این عزیزان توفیق روزافزون و عمر با برکت از درگاه احدیت مسئلت دارم.

چکیده:

سرطان، یکی از دلایل اصلی مرگ در سراسر جهان است. در این بیماری، سلول‌های بدن به صورت غیر قابل کنترل رشد می‌کنند. گروهی از سلول‌هایی که به سرعت تکثیر شده‌اند، ممکن است یک توده، جرم یا بافت اضافی ایجاد کنند. این توده‌ها، تومور نامیده می‌شوند. توده-ها می‌توانند سرطانی یا غیرسرطانی باشند. سرطان سینه، دومین عامل مرگ زنان است. این بیماری در میان مردان و زنان دیده شده، گرچه موارد ابتلا در میان مردان نادر است. در ایران نیز سرطان سینه، شایع‌ترین سرطان در میان زنان است. البته میزان ابتلا به این بیماری در ایران، بسیار کمتر از آمریکا و سایر کشورهای اروپایی است. همچنین سن ابتلا به این بیماری در ایران، نسبت به کشورهای توسعه‌یافته، یک دهه پائین‌تر است. فاکتور اصلی در کاهش مرگ و میر ناشی از این بیماری، تشخیص سریع و صحیح است. در این تحقیق، یک سیستم رده-بندی ارائه شده که می‌تواند در تشخیص تومور به پزشکان کمک کند. به علت چند رده‌ای و نامتوازن بودن داده‌های مورد استفاده در این تحقیق، رهیافت جدیدی برای یادگیری در داده-های چند رده‌ای و نامتوازن ارائه شده است. حساسیت سیستم رده‌بندی ارائه شده، ۸۶.۵۳ درصد و دقت آن، ۸۹.۲۵ درصد می‌باشد. همچنین در این تحقیق با استفاده از الگوریتم Apriori، به شناسایی عوامل موثر در بروز توده‌های خوش‌خیم و بدخیم پرداخته شده و عوامل افزایش سن، منارک زودرس، یائسگی دیررس، داشتن سن بالا هنگام اولین بارداری، سابقه‌ی طولانی مدت مصرف قرص‌های ضدبارداری، چاقی، نداشتن سابقه‌ی بارداری، زایمان و شیردهی و تعداد دفعات شیردهی بالا، به عنوان عوامل موثر در بروز توده، شناخته شده‌اند. اگر چه بیشتر این عوامل قبلا به عنوان فاکتورهای خطرزا در بروز سرطان سینه معرفی شده بودند، فاکتور تعداد دفعات شیردهی بالا، در این مطالعه به عنوان یک فاکتور خطرزا تعیین گردیده که با یافته‌های محققان قبلی در تعارض می‌باشد. این مسئله می‌تواند به دلیل تعداد بسیار اندک نمونه‌های مبتلا به تومور بدخیم باشد.

کلمات کلیدی: داده‌کاوی، رده‌بندی، قوانین انجمنی، تشخیص سرطان سینه، داده‌های نامتوازن و چند رده‌ای

فهرست مطالب

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست تصاویر

عنوان **صفحه**

فهرست نمودارها

عنوان **صفحه**

فهرست نشانه‌ها و اختصارها (Abreviations)

PSO	Particle Swarm Optimization
DPSO	Discrete Particle Swarm Optimization
SVM	Support Vector Machine
LS-SVM	Least Square Support Vector Machine
k-NN	k Nearest Neighbor
C&RT	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detector
QUEST	Quick, Unbiased, Efficient Statistical Tree
G-mean	Geometric mean
SMOTE	Synthetic Minority Over-Sampling
RUS	Random Under-Sampling
ROS	Random Over-Sampling
OSS	One-Sided Selection
EE	EasyEnsemble
MB	ModifiedBagging
MEE	Multi-class EasyEnsemble
NN	Neural Network
KDD	Knowledge discovery in databases
WHO	World Health Organization
SEER	Surveillance, Epidemiology and End Results
NCI	National Cancer Institute
BMI	Body Mass Index
FNA	Fine Needle Aspiration
MHT	Menopausal Hormone Therapy
CRM	Customer Relationship Management

فصل ۱: مقدمه

۱-۱. موضوع تحقیق

سرطان[1] نوعی بیماری است که در آن سلول‌های بدن، به صورت غیر قابل کنترل رشد می‌کنند، تغییر می‌کنند و تکثیر می‌شوند. سرطان، معمولا با بخشی از بدن که درگیر این بیماری است، نامگذاری می‌شود. در نتیجه، سرطان سینه[2] به معنای رشد غیرعادی سلول‌ها در بافت سینه است. گروهی از سلول‌هایی که به سرعت تکثیر شده‌اند، ممکن است یک توده[3]، جرم[4] یا بافت اضافی[5] ایجاد کنند. این توده‌ها، تومور[6] نامیده می‌شوند. توده‌ها می‌توانند سرطانی (بدخیم[7]) یا غیرسرطانی (خوش‌خیم[8]) باشند. تومورهای سرطانی، سرایت می‌کنند و بافت‌های سالم بدن را نابود می‌کنند. در واقع سرطان سینه به معنای تومورهای بدخیمی است که از سلول‌های سینه گسترش یافته‌اند. این سرطان، در میان مردان و زنان دیده شده است، گرچه موارد ابتلا در میان مردان نادر است. دلیل واقعی بروز سرطان سینه، مشخص نیست اما برخی فاکتورها، ریسک ابتلا به سرطان سینه را افزایش می‌دهند[9].

سرطان سینه، دلیل اصلی مرگ زنان ۴۰ تا ۵۵ ساله و دومین عامل مرگ زنان (پس از سرطان ریه[10]) است[11]. طبق آمار سازمان بهداشت جهانی[1]، سالانه بیش از ۱.۲ میلیون زن در

---

[1] Cancer
[2] Breast Cancer
[3] Lump
[4] Mass
[5] Extra tissue
[6] Tumor
[7] Malignant
[8] Benign
[9] Cancer Council Victoria. **Breast cancer: for people with cancer, their families and friends**. (Melbourne: Cancer Council Victoria, 2009), Page 1.
American Cancer Society. **Cancer Facts & Figures 2009**. (Atlanta: American Cancer Society, 2009), Page 1.
American Cancer Society. **Breast Cancer Facts & Figures 2009-2010**. (Atlanta: American Cancer Society, 2010), Page 1.
[10] Lung Cancer
[11] http://www.imaginis.com/breasthealth/breast_cancer.asp

سراسر دنیا، مبتلا به سرطان سینه، تشخیص داده می‌شوند[۲]. در ایران نیز سرطان سینه، شایع-ترین سرطان در میان زنان است[۳]. خوشبختانه در سال‌های اخیر، نرخ مرگ و میر ناشی از سرطان سینه، به دلیل تاکید بیشتر روی تکنیک‌های موثرتر تشخیص و معالجه، کاهش یافته است. فاکتور اصلی در این روند، تشخیص سریع و صحیح است[۴].

تکنیک‌های داده‌کاوی[۵] علاوه بر این که می‌توانند در تشخیص سریع این بیماری موثر باشند، می‌توانند از طریق شناسایی عوامل موثر در بروز توده‌های خوش‌خیم و بدخیم، به پیش-گیری از این بیماری کمک کنند و به این ترتیب باعث ارتقای سلامت جامعه شده و از تحمیل هزینه‌های سنگین تشخیص و درمان بیماری، بر بیماران جلوگیری کنند.

۱-۲. اهمیت و ضرورت تحقیق

پس از ۲۱ قرن، هنوز سرطان دلیل اصلی مرگ در جهان است و سرطان سینه به دومین عامل اصلی مرگ سرطانی در میان زنان تبدیل شده است[۶]. گرچه، گسترش تکنولوژی‌های پزشکی در دهه‌ی گذشته، موجب کاهش میزان مرگ و میر ناشی از این بیماری شده است. به دلیل تشخیص زود هنگام و درمان بهینه، میزان بهبودی افزایش یافته است. حدود ۹۷ درصد زنان می‌توانند به مدت ۵ سال یا بیشتر بهبود یابند[۷]. بعلاوه میزان بهبودی ۵ ساله برای بیماران با هر دو نوع تومور به میزان ۱۰ تا ۱۵ درصد افزایش یافته است و این به دلیل تشخیص زودهنگام بیماری است. بنابراین روش موثر برای کاهش مرگ در اثر سرطان سینه، تشخیص زودهنگام این بیماری است[۸]. تشخیص زودهنگام نیز به یک روش تشخیص صحیح و قابل

---

[1] World Health Organization (WHO)
[2] Mehmet Fatih Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". **Expert Systems with Applications**, 36, 2(2009):3240–3247.
[3] Shadi Kolahdoozan et al. "Five Common Cancers in Iran". **Archives of Iranian Medicine**, 13, 2(2010):143 – 146.
[4] Mehmet Fatih Akay, 2009.
[5] Data Mining
[6] Wei-Chang Yeh, Wei-Wen Chang & Yuk Ying Chung. "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method". **Expert Systems with Applications**, 36, 4(2009):8204–8211.
[7] Dursun Delen, Glenn Walker & Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods". **Artificial Intelligence in Medicine**, 34, 2(2005):113-127.
[8] Wei-Chang Yeh, Wei-Wen Chang & Yuk Ying Chung, 2009.

اعتماد نیاز دارد که به پزشکان اجازه می‌دهد تومورهای خوش‌خیم را از نوع بدخیم آن شناسایی کنند. بنابراین یافتن یک متد تشخیص صحیح و موثر و همچنین تشخیص عوامل خطر در بروز این بیماری، بسیار با اهمیت است.

مشکلی که در تشخیص این بیماری وجود دارد این است که در ابتدا ممکن است توده-های سرطانی، بسیار کوچک بوده و قابل لمس نباشند. بدیهی است تا زمانی‌که یک توده‌ی کوچک، بزرگ شده و قابل لمس شود، بیماری نیز پیشرفت کرده و این تأخیر در تشخیص بیماری، روند درمان را با مشکل مواجه خواهد کرد تا جایی که حتی ممکن است به مرگ بیمار منجر شود. بنابراین نیاز است تا از روش‌های پیشرفته‌تری برای تشخیص زودهنگام سرطان سینه استفاده شود. روش‌های رایج برای تشخیص زودهنگام سرطان سینه عبارتند از: ماموگرافی[۱]، سونوگرافی[۲] و نمونه‌برداری. این روش‌ها علاوه بر هزینه‌ی بالایی که بر بیمار تحمیل می‌کنند، ممکن است مضراتی را برای بیمار در پی داشته باشند. مخصوصا در روش ماموگرافی که از اشعه‌ی ایکس برای تولید تصاویر استفاده می‌شود. به طوری‌که این روش، معمولا برای افراد زیر ۴۰ سال که احتمال ابتلا به سرطان سینه در آنان کمتر است، توصیه نمی‌شود[۳].

ماموگرافی، یکی از رایج‌ترین روش‌های تشخیص سرطان سینه است. این روش می‌تواند ناهنجاری‌ها را با صحت[۴] ۸۵ تا ۹۰ درصد شناسایی کند[۵]. اما مشکلی که در این روش وجود دارد، این است که نتیجه‌ای که متخصصان مختلف از این تصاویر استنباط می‌کنند، ممکن است متفاوت باشد و این موضوع چالش دیگری را در امر تشخیص سرطان سینه ایجاد می‌کند[۶]. در چنین مواردی، از روش آسپیراسیون سوزنی ظریف[۷] که یکی از روش‌های نمونه‌برداری است،

---

[1] Mamography
[2] Ultrasound
[3] محمدرضا صفایی کشتگر و راب استین. **اطلاعات عمومی برای پیش‌گیری و نقاشی‌های بالینی برای بیماران مبتلا به سرطان پستان**. مریم برزگر و عطیه اکبری، تهران: پیام سفید، چاپ چهارم، (۱۳۸۸)، صفحه ۱۰
[4] Accuracy
[5] Shieu-Ming Chou et al. "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines". **Expert Systems with Applications**, 27, 1(2004):133–142.
[6] Joann G. Elmore et al. "Variability in radiologists interpretation of mammograms". **New England Journal of Medicine**, 331, 22(1994):1493-1499.
[7] Fine Needle Aspiration (FNA)