ORIGINAL ARTICLE

Iranian Quarterly Journal of Breast Disease 2018; 11(3):55.

Breast Cancer Staging by Using TNM System and Ant Colony Algorithm

Naseri Norowzani S: Master of Computer Engineering, Department of Computer, Shiraz Branch, Islamic Azad University, Shiraz, Iran
Shayegan MA: Assistant Professor of Computer Engineering, Department of Computer, Shiraz Branch, Islamic Azad University, Shiraz, Iran

Corresponding Author: Mohammad Amin Shayegan, shayegan@iaushiraz.ac.ir

[DOR: 20.1001.1.17359406.1397.11.3.5.5]

[DOI: 10.30699/acadpub.ijbd..11.3.55]

Abstract

Introduction: Staging is one of the most important factors determining the survival of a patient suffering from breast cancer and plays a key role in choosing treatment modalities. The method of choice for cancer staging is surgery followed by histological evaluation. However, finding a predictive algorithm to replace surgery in the staging of breast cancer is both cost- and time-saving and helps physicians to provide appropriate therapeutic techniques. The present paper introduces a strong predictive model for breast cancer staging using data mining techniques.

Methods: We suggested a mechanized model based on the TNM staging system and the ant colony algorithm. This method would reduce the patient's mental stress and financial costs because it does not need a surgical operation. The SEER international dataset and a local data set of 1148 women with breast cancer were used to evaluate the system performance, and model accuracy and the area under the ROC curve were calculated for different classifications.

Results: Using the TNM system, the accuracy rates were 99.93% and 99.91% for the SEER international dataset and the local dataset, respectively. The accuracy rates were 99.43% and 98.95% for the SEER international dataset and local dataset, respectively, when the ant colony algorithm was applied. Our results indicated that in addition to commonly used features in the TNM system, other features such as vascular invasion, age, blood group, number of children, birthplace, histology, CS Extension, positive regional node, morphology, and Site-Specific Factors 2, 3, and 6 can be used as important factors for breast cancer staging.

Conclusion: Based on the results obtained, the two Logistic and Multi-Class Classifiers have the best accuracies for the SEER and local datasets in this study.

Keywords: Breast Cancer, Stage detection, Data mining, TNM system, Ant Colony Algorith

مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال یازدهم، شماره سوم، پاییز ۱۳۹۷؛ (۷۰-۵۶)

تاریخ ارسال: ۹۷/۰۶/۰۷ | تاریخ پذیرش: ۹۷/۰۹/۱۲

تعیین مرحله بالینی بیماران مبتلا به سرطان پستان با استفاده از سیستم TNM و الگوریتم کلونی مورچگان

سعیده ناصری نوروزانی: گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران

محمد امین شایگان *: گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران

چکیده

مقدمه: میزان پیشرفت سرطان پستان (مرحله، Staging)، یکی از مهم‌ترین عوامل تعیین کننده میزان بقای فرد بیمار و انتخاب روش‌های درمانی مناسب توسط پزشکان است. معمولا تعیین پیشرفت سرطان پستان، پس از عمل جراحی و از طریق ارزیابی بافت شناسی انجام می‌شود. از این رو یافتن الگوریتم مناسبی که بتواند میزان پیشرفت و همچنین مرحله (Staging) سرطان پستان را تعیین کند، به پزشکان در ارایه روش‌های درمانی مناسب کمک فراوانی خواهد کرد. لذا در این پژوهش تلاش شده است تا با استفاده از تکنیک‌های داده‌کاوی، یک مدل قوی پیش‌بینی مرحله سرطان پستان معرفی گردد.

روش بررسی: در این پژوهش، یک مدل مکانیزه با استفاده از سیستم TNM و همچنین استفاده از الگوریتم کلونی مورچگان، برای تشخیص مرحله سرطان پستان، پیشنهاد شده است. این روش‌ها به دلیل عدم نیاز به عمل جراحی، باعث کاهش زیاد هزینه‌ها و آسیب‌های روحی بیمار می‌شود. برای ارزیابی سیستم، از دیتاست بین‌المللی SEER و یک دیتاست محلی از اطلاعات ۱۱۴۸ بیمار زن مبتلا به سرطان پستان، استفاده شد و دو معیار «دقت» و «سطح زیر نمودار راک» برای طبقه‌بندهای مختلف محاسبه گردید.

یافته‌ها: با استفاده از سیستم TNM برای دیتاست SEER، دقت ۹۹/۹۳٪ و برای دیتاست محلی، دقت ۹۹/۹۱٪ و با استفاده از الگوریتم کلونی مورچگان، برای دیتاست SEER دقت ۹۹/۴۳٪ و برای دیتاست محلی، دقت ۹۸/۹۵٪ بدست آمد. همچنین مشخص گردید علاوه بر ویژگی‌های مورد استفاده مرسوم T ، N و M، ویژگی‌های دیگری همچون تهاجم عروقی، سن بیمار، گروه خونی، تعدا فرزندان، محل تولد، بافت شناسی سلولی، نوع بافت درگیر و Site-Specific Factor های شماره ۲ ، ۳ و ۶ نیز می‌توانند به عنوان عوامل مهم در تعیین مرحله بالینی بیماران مبتلا به سرطان پستان استفاده شوند.

نتیجه‌گیری: بر اساس نتایج حاصل، دو طبقه‌بند Logistic و Multi Class Classifier به ترتیب دارای بالاترین میزان دقت برای دیتاست‌های SEER و محلی این پژوهش هستند.

واژه‌های کلیدی: سرطان پستان، تشخیص مرحله، داده‌کاوی، سیستم TNM، الگوریتم کلونی مورچگان

* نشانی نویسنده مسئول:شیراز، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد شیراز، محمد امین شایگان.

نشانی الکترونیک: shayegan@iaushiraz.ac.ir

مقدمه

امروزه سرطان پستان، یکی از اصلی‌ترین دلایل مرگ و میر در بین زنان در سرتاسر جهان است (۱-۳). این سرطان یکی از شایع‌ترین سرطان‌ها در میان زنان بوده (۴-۷) که البته در مردان نیز رخ می‌دهد (۵). در بین انواع مختلف سرطان، ۲۷٪ آنها مربوط به سرطان پستان است و این بیماری روز به روز در حال افزایش است (۳). امروزه افراد زیادی در سراسر جهان از این بیماری رنج می‌برند به‌طوری‌که سرطان پستان، دومین علت مرگ و میر در بین زنان می‌باشد (۸، ۹). بر اساس اعلام سازمان بهداشت جهانی[1] (WHO)، سرطان پستان یکی از بیشترین سرطان‌های تشخیص داده شده بین زنان در کشورهای توسعه یافته و در حال توسعه است (۱۱،۱۰،۷). در دهه ۱۹۷۰ در ایالت متحده آمریکا، از هر ۱۳ زن یک نفر و در سال ۲۰۰۴ از هر هشت زن یک نفر مبتلا به سرطان پستان بوده است (۱۲). امروزه در کشورهای توسعه یافته نیز از هر هشت زن، یک زن در طول زندگی خود، سرطان پستان را تجربه می‌کند (۱۳). بر اساس آخرین آمار مرکز تحقیقات سرطان در ایران، سالیانه حدود ۸۵۰۰ مورد جدید سرطان پستان در کشور ثبت می‌شود و ۱۴۰۰ نفر به دلیل ابتلا به سرطان پستان فوت می‌کنند. همچنین تا سال ۱۳۹۴در حدود ۴۰,۰۰۰ نفر در ایران مبتلا به این بیماری بوده‌اند (۱۴).

بر اساس تحقیق سنترک و کارا (۱۱)، بالا رفتن انتظارات زندگی، شهرنشینی، هجوم و پذیرش زندگی غربی و... باعث وقوع و پیشرفت سرطان پستان شده است. بیشتر این سرطان‌ها در فاز آخر و در سنین ۱۵ الی ۴۹ سالگی تشخیص داده می‌شوند (۱۵)، درحالی‌که پیش‌بینی و تشخیص زودهنگام و قبل از پیشرفت بیماری، بسیار مهم بوده و باعث زنده ماندن و افزایش بقای شخص بیمار می‌گردد (۱۶). اگر این بیماری در مراحل[2] اولیه به درستی تشخیص داده شود، می‌توان تا حدود زیادی از پیشرفت آن جلوگیری کرد و افراد را از خطر مرگ نجات داد. آمارها نشان می‌دهند که میزان بقای بیماران مبتلا به سرطان پستان تا پنج سال پس از تشخیص به موقع، ۸۸٪ و تا ده سال پس از تشخیص به موقع ۸۰٪ است (۱۷). همچنین راست قلم و پورقاسم (۳) معتقدند که اگر توده سرطانی خیلی زود و در مراحل اولیه تشخیص داده شود، یعنی زمانی‌که اندازه آن کوچک و کمتر از mm ۱۰ است، شانس نجات و درمان شخص، ۸۵٪ خواهد بود. اما اگر دیر و در مراحل آخر تشخیص داده شود، شانس درمان و نجات شخص، فقط ۱۰٪ می‌باشد.

پیش‌بینی و تشخیص زود هنگام و همچنین تعیین دقیق مرحله سرطان پستان، از جمله اهداف مهم و بزرگ پیش روی محققان در این زمینه می‌باشد. لذا پژوهش‌های زیادی توسط محققین در زمینه تشخیص زودهنگام و بقای بیماران مبتلا به سرطان پستان انجام شده است (۲). لیکن پژوهش‌های بسیار اندکی در زمینه تشخیص مرحله سرطان پستان صورت گرفته است که با توجه به اهمیت این موضوع در اتخاذ نحوه درمان و نجات جان افراد، ضروری است که به آن نیز پرداخته شود.

اتحادیه سرطان آمریکا[3](AJCC) در سال ۱۹۵۹ سازمان‌دهی شد تا بتواند سیستمی جهت تشخیص مرحله کلینیکی سرطان را گسترش دهد. پس از گذشت شش سال، اعضای این کمیته چندین راهنما در حوزه سرطان‌های مختلف از جمله پستان، حنجره، گردن، حلق و دهانه رحم منتشر کردند (۱۸). سیستم استاندارد TNM یکی از مرسوم‌ترین سیستم‌های طبقه‌بندی برای تعیین میزان گستردگی و مرحله سرطان است (۱۹). این سیستم مبتنی بر سه نماد T (اندازه تومور سرطانی)، N (تعداد غدد لنفاوی درگیر سرطان) و M (متاستاز) است (۲۰). این سیستم برای اولین بار توسط پیر دنویکس در دهه ۱۹۵۰ معرفی شد و اولین نسخه آن توسط اتحادیه سرطان آمریکا در سال ۱۹۷۷ و دومین نسخه آن در سال ۱۹۸۳ منتشر شد که تولید کنندگان آن بیان کردند که دو فاکتور «درجه تمایز تومورهای اولیه» و «سن بیمار» نقش مهمی در مرحله‌بندی سرطان ایفا می‌کنند. نسخه‌های سوم تا هشتم این سیستم به ترتیب در سال‌های ۱۹۸۸، ۱۹۹۲ ، ۱۹۹۷، ۲۰۰۲ ، ۲۰۰۹ و ۲۰۱۶ منتشر شدند که هر نسخه شامل اصلاحاتی نسبت به نسخه قبلی خود بوده است (۲۱).

در سیستم TNM تومورهای اولیه بر اساس اندازه ، به پنج بخش T_0 ،T_1 ،T_2 ،T_3 و T_4 تقسیم می‌شوند. T_0 نشان‌دهنده این است که هیچ تومور سرطانی وجود ندارد، T_1 نشان‌دهنده تومور سرطانی کوچک‌تر از دو سانتی‌متر،

[1] World Health Organization (WHO)
[2] Stage
[3] American Joint Committee on Cancer (AJCC)

[DOR: 20.1001.1.17359406.1397.11.3.5.5]
DOI: 10.30699/acadpub.ijbd..11.3.55]

T_2 نشان‌دهنده تومور سرطانی بیشتر از دو سانتی‌متر و کمتر از پنج سانتی‌متر، T_3 نشان‌دهنده تومور سرطانی بزرگتر از پنج سانتی‌متر و کوچک‌تر از ۱۰ سانتی‌متر و T_4 نشان‌دهنده تومور سرطانی بزرگتر از ۱۰ سانتی‌متر می‌باشد، به‌طوری‌که پوست و قفسه سینه را هم درگیر کرده باشد (۲۲).

بر اساس تعداد غدد لنفاوی درگیر، تومورهای اولیه به چهار بخش N_0، N_1، N_2 و N_3 تقسیم می‌شوند که N_0 نشان‌دهنده این است که هیچ غده لنفاوی درگیر نشده است، N_1 نشان‌دهنده این است که یک تا سه عدد از غدد لنفاوی درگیر می‌باشند، N_2 نشان‌دهنده این است که چهار تا نه عدد از غدد لنفاوی درگیر می‌باشند و N_3 نشان‌دهنده این است که بیش از ۱۰ عدد از غدد لنفاوی درگیر هستند (۲۲-۲۴).

در سیستم TNM، متاستاز به دو بخش M_0 و M_1 تقسیم شده است که M_0 نشان‌دهنده این است که هیچکدام از ارگان‌ها و اعضای بدن، درگیر سرطان نیستند و M_1 نشان‌دهنده این است که علاوه بر پستان، ارگان‌ها و اعضای دیگری از بدن مانند استخوان، مغز، ریه و کبد نیز درگیر سرطان شده‌اند (۲۲، ۲۴).

با توجه به تقسیم‌بندی‌های اشاره شده، مراحل سرطان پستان از مرحله صفر[4] تا مرحله چهار[5] نام‌گذاری شده‌اند که برخی از این مراحل خود شامل زیر گروه‌هایی نیز هستند که با حروف A، B و C نشان داده می‌شوند. به عنوان قانون، پایین‌ترین شماره، نشان‌دهنده کمترین میزان پیشرفت سرطان و بالاترین شماره، نشان‌دهنده بیشترین میزان پیشرفت سرطان است. به عنوان مثال مرحله چهار به معنی سرطان بسیار پیشرفته می‌باشد. همچنین در زیر گروه‌ها، حرفی که زودتر آمده است نشان‌دهنده پایین‌ترین (کم خطرترین) مرحله می‌باشد.

علم امروزه، به خصوص در زمینه پزشکی، با حجم بسیار زیادی از داده‌ها روبرو بوده به‌طوری‌که کار کردن با جزئیات این داده‌ها و اتخاذ تصمیمات مرتبط با تشخیص و درمان، بسیار وقت‌گیر و هزینه‌بر می‌باشد (۲۵). با پیشرفت تکنولوژی‌های کلینیکی، اطلاعات متنوعی از سرطان پستان جمع‌آوری شده‌اند که پالایش و اداره کردن همه این اطلاعات، به یک چالش بزرگ در این زمینه تبدیل شده است. بنابراین نقش تکنیک‌های داده‌کاوی در تشخیص پزشکی بسیار حایز اهمیت است و این تکنیک به یک راه‌کار مفید برای پزشکان تبدیل شده تا بتوانند سرطان‌ها را در مراحل اولیه تشخیص داده و بهترین روش درمانی را برای بیمار اتخاذ کنند. یکی از مهم‌ترین مزایای داده‌کاوی، به ویژه در حجم وسیع داده‌ها، این است که استفاده از تکنیک‌های داده‌کاوی، نسبت به به‌کارگیری افراد خبره، بسیار کم هزینه‌تر است (۲۶).

تحقیقات نشان داده‌اند که روش‌های داده‌کاوی برای پیش‌بینی و تشخیص سرطان پستان، نسبت به روش‌های سنتی موفق‌تر بوده‌اند (۶ ، ۱۰)، به‌طوری‌که امروزه بخش سلامت بیشترین نیاز را به داده‌کاوی پیدا کرده و از روش‌های سنتی به سمت پزشکی مبتنی بر شواهد سوق پیدا کرده است (۲۶). لذا در این پژوهش برای ساخت مدل پیش‌بینی مرحله سرطان پستان، از تکنیک‌های داده‌کاوی استفاده گردید. هدف نهایی این پژوهش، پیشنهاد مدلی است که بتواند با انتخاب کمترین تعداد ویژگی‌ها، دقت تشخیص و تعیین مرحله سرطان پستان را بهبود دهد و نهایتا توسط کلینیک‌های پزشکی و درمانی، بیمارستان‌ها و پزشکان مورد استفاده قرار گیرد.

مواد و روش‌ها

متدولوژی این پژوهش از دو بخش عمده تشکیل شده است. هدف بخش اول، ایجاد مدل تشخیص مرحله سرطان پستان بر اساس دیتاست بین‌المللی SEER بوده و در بخش دوم، هدف بومی‌سازی و ایجاد مدل تشخیص مرحله سرطان پستان بر اساس یک دیتاست محلی بوده است. در هر دو بخش، عملیات پیش پردازش مناسب بر روی هر دو دیتاست صورت گرفت تا اطلاعاتی به‌دست آیند که قابل استفاده برای عملیات مراحل بعدی باشند. جهت پیاده‌سازی و تحلیل داده‌ها، از نرم‌افزارهای آماری رپیدماینر نسخه ۵-۳ و نرم‌افزار وکا نسخه ۳-۸-۱ استفاده گردید. از آنجا که انتخاب کلیدی‌ترین ویژگی‌ها، در داده‌های پزشکی، یکی از مهم‌ترین چالش‌ها در این زمینه است، لذا به منظور دستیابی به اهداف این مطالعه، دو سناریو در هر بخش در نظر گرفته شد. اولین سناریو، ایجاد مدل تشخیص و پیش‌بینی مرحله سرطان پستان بر اساس سیستم TNM و دومین سناریو، ایجاد مدل پیش‌بینی با استفاده از الگوریتم فراابتکاری کلونی مورچگان،

[4] Stage 0
[5] Stage IV

به‌طوری‌که بتوان ویژگی‌های موثر و قدرتمندی استخراج کرد، است. در نهایت، نتایج حاصل از هر دو حالت با هم مقایسه گردید.

داده‌های مورد استفاده: در این پژوهش دو دیتاست مورد استفاده قرار گرفته است. دیتاست اول، دیتاست بین‌المللی SEER بوده که داده‌های آن، طی سال‌های ۱۹۷۳ تا ۲۰۱۱ جمع‌آوری شده است. این دیتاست شامل ۵۱۱۸۷۸ نمونه و ۱۴۹ ویژگی است که از اطلاعات تمام بیماران سرطانی از جمله پستان، روده، مقعد و... تشکیل شده است که پس از انتخاب داده‌های سرطان پستان، حجم آن به ۱۶۹۷۲۱ نمونه و ۱۴۹ ویژگی و نهایتا پس از انجام یکسری عملیات پیش پردازش مناسب و حذف ویژگی‌های غیرضروری، حجم آن به ۱۴۰۴ نمونه و ۸۶ ویژگی تقلیل یافت.

دیتاست دوم، یک دیتاست محلی است که داده‌های آن، طی سال‌های ۲۰۰۰ الی ۲۰۰۵ از مراکز سرطان شیراز، که از جمله مراکز مهم در ارایه سرویس و خدمات برای سرطان در جنوب کشور می‌باشند، جمع‌آوری شده است. این دیتاست شامل ۱۱۴۸ نمونه و ۵۸ ویژگی است. از جمله این ویژگی‌ها می‌توان به اندازه تومور، تعداد غدد لنفاوی درگیر، سن بیمار، تاریخ تشخیص سرطان، درجه تمایز تومورهای سرطانی، محل تومور، سن اولین حاملگی شخص، تعداد فرزندان، مدت زمان شیردهی، مذهب، درآمد خانوادگی شخص، محل اقامت، تحصیلات، شغل، مصرف قرص ضدبارداری، سن یائسگی، مصرف سیگار، قد، وزن، تقسیم بندی بافتی سلول سرطانی، گروه خونی، مثبت یا منفی بودن گروه خونی، درگیری استخوان شخص به سرطان، درگیری کبد، ریه، مغز، تخمدان، پستان سمت چپ یا راست شخص به سرطان و... اشاره کرد. از آنجایی که هیچ نمونه تکراری در این دیتاست محلی وجود ندارد و تنها تعدادی از ویژگی‌ها (در بعضی از نمونه‌ها) دارای مقادیر مفقود هستند، بنابراین هیچ نمونه‌ای از این دیتاست حذف نگردید.

پیش پردازش داده‌ها: عموما داده‌های خام اولیه دیتاست‌ها، مستقیما قابل استفاده نیستند. لذا باید عملیات پیش پردازش مناسب مانند جداسازی داده‌های خام، حذف ویژگی‌های غیرضروری، فیلتر کردن نمونه‌ها و تبدیل نوع بر روی داده‌ها انجام شود. این عملیات باعث می‌شود که نتایج حاصل، از کیفیت و دقت بالاتری برخوردار باشند.

الف) انتخاب داده‌های سرطان پستان در دیتاست SEER: هدف این تحقیق، انجام عملیات بر روی داده‌های سرطان پستان بوده است. از آنجا که دیتاست SEER شامل اطلاعات مربوط به سایر سرطان‌ها نیز است، بنابراین باید فقط داده‌ها و نمونه‌هایی که وابسته به سرطان پستان هستند، استخراج شوند. از بین ویژگی‌های موجود در این دیتاست، ویژگیPrimary Site نشان‌دهنده نوع سرطان است که نمونه‌های مربوط به سرطان پستان در این فیلد، مقداری بین C500 تا C509 دارند. لذا در این مرحله، تعداد نمونه‌ها از ۵۱۱۸۷۸ به ۱۶۹۷۲۱ تقلیل یافت. سپس ۱۴۰۴ نمونه به صورت تصادفی انتخاب شد تا هر دو دیتاست بین‌المللی و محلی تقریبا از تعداد نمونه‌های یکسانی برخوردار باشند.

ب) حذف ویژگی‌های غیرضروری از دیتاست SEER: در این مرحله، ویژگی‌هایی غیرمفید از اطلاعات حذف شدند. برای نیل به این هدف، با تنظیم پارامتر Numerical Min Deviation با مقدار ۰/۹۹۸ در نرم‌افزار طراحی شده، تمام ویژگی‌هایی که مقدار واریانس آنها از این حد آستانه تجاوز می‌کند و همچنین فیلدهای تک مقداری، حذف خواهند شد. نتیجه این عملیات، کاهش تعداد ویژگی‌ها از ۱۴۹ به ۸۶ بوده است.

ج) تبدیل نوع داده‌ها در دیتاست SEER: شکل مناسب داده، به عنوان ورودی الگوریتم‌های داده‌کاوی، نقش مهمی در این فرآیند بازی می‌کند. بنابراین برای آماده سازی بعضی از ویژگی‌ها برای عملیات، نیاز به تغیر نوع آنها است. در دیتاست SEER، ویژگی‌های Derived AJCC T، Derived AJCC N و Derived AJCC M به‌ترتیب نشان‌دهنده سه ویژگی T، N و M می‌باشند. این سه ویژگی از نوع اسمی هستند، بنابراین برای اینکه قابل استفاده برای عملیات پردازش شوند به نوع عددی تبدیل شدند.

ایجاد متغیرها و مراحل سیستم TNM: برای پیاده‌سازی سیستم TNM و تعریف مراحل سرطان پستان، نیاز به ایجاد متغیرهای مربوطه در محیط شبیه‌سازی نرم‌افزار است. متغیرهای سیستم TNM شامل پنج متغیر T_0 الی T_4، چهار متغیر N_0 الی N_3 و

[DOR: 20.1001.1.17359406.1397.11.3.5.5]
DOI: 10.30699/acadpub.ijbd..11.3.55]

دو متغیر M_0 و M_1 و همچنین متغیرهای تعیین مراحل سرطان پستان، شامل هشت متغیر Stage 0 الی Stage IV هستند. تمام این متغیرها با استفاده از اطلاعات موجود در دیتاست بین‌المللی SEER و دیتاست محلی و با استفاده از نرم‌افزار رپیدماینر ساخته شدند.

آموزش و آزمایش سیستم TNM: برای ایجاد مدل پیش‌بینی و تشخیص مرحله بالینی سرطان پستان بر اساس سیستم استاندارد TNM، باید از ویژگی‌های T_0، T_1، T_2، T_3، T_4، N_0، N_1، N_2، N_3، M_0، M_1 به همراه فیلد کلاس (Staging) (که در مرحله قبل یعنی ایجاد متغیرها و مراحل سیستم TNM ایجاد شد) استفاده گردد.

ذکر یک نکته مهم بسیار ضروری است که رکوردهای هر دو دیتاست SEER و محلی به‌کار رفته در این پژوهش، فاقد مقدار برای آیتم Staging است (یعنی در نسخه‌های مختلف این دیتاست‌ها، مشخص نشده که یک بیمار در چه مرحله‌ای از سرطان پستان است و فیلد مربوطه، به صورت null است) لذا در ابتدا باید رکوردهای موجود در این دیتاست ، تعیین مرحله شوند (لیبل بگیرند، فرآیند Staging) تا بتوان مرحله سرطان پستان هر بیمار را مشخص کرد. با انجام این کار است که می‌توان در ادامه، از الگوریتم‌های مختلف داده‌کاوی و فراابتکاری (مثل کلونی مورچگان) استفاده کرد تا مرحله سرطان یک بیمار مشخص شود و نهایتا، صحت این نتایج را با مقایسه نتایج با مقادیر واقعی موجود در benchmark، مقایسه و قدرت و دقت روش پیشنهادی را محاسبه و معرفی کرد.

برای ایجاد مدل‌های مختلف دسته‌بندی در نرم‌افزار وکا، پس از انتخاب ویژگی‌های مورد نیاز، انواع مختلف طبقه‌بندها مانند درخت تصمیم، شبکه‌های عصبی و شبکه‌های بیزین و ... قابل مشاهده بوده که طبق نیاز می‌توان هر کدام از این طبقه‌بندها را انتخاب و بر روی داده‌ها اعمال کرد. برای بدست آوردن دقیق‌ترین تخمین، کمترین میزان خطا و تقسیم داده‌های موجود در دیتاست به دو دسته آموزشی و تست، تکنیک اعتبارسنجی متقاطع با ده تکرار[6] انتخاب گردید. پس از اجرا شدن مدل مورد نظر، معیارهای "میزان دقت"[7] و "سطح زیر نمودار راک"[8] طبقه‌بندها با هم مقایسه شد تا نهایتا بهترین مدل دسته‌بندی که بهترین کارایی را روی دیتاست پیشنهادی دارند، شناسایی و انتخاب گردد. این گام نیز در هر دو دیتاست SEER و محلی یکسان بوده است.

آموزش و آزمایش سیستم الگوریتم کلونی مورچگان: برای انتخاب ویژگی‌ها[9] (FS)، روش‌های مختلفی وجود دارد که از جمله آنها می‌توان به الگوریتم‌های فراابتکاری همچون الگوریتم کلونی مورچگان، که جهت بهینه‌یابی در حل مسایل مورد استفاده قرار می‌گیرد، اشاره کرد. این الگوریتم از رفتار اجتماعی مورچگان برای جستجوی کوتاه‌ترین مسیر برای یافتن غذا الهام گرفته است (۲۷). در این الگوریتم مسئله به صورت یک گراف نمایش داده شده و هر گره گراف، یک ویژگی را نمایش می‌دهد. گره‌ها توسط یال‌هایی بهم وصل شده‌اند که با استفاده از این یال‌ها می‌توان بر روی گره‌ها حرکت کرده و ویژگی‌های مناسبی را انتخاب کرد (۲۸). مورچه‌ها هنگام عبور از یک مسیر، از خود ردی از ماده شیمیایی فرومون بجا می‌گذارند که البته این ماده زود تبخیر می‌شود. اما در کوتاه مدت به عنوان رد مورچه بر سطح زمین باقی می‌ماند و مورچه‌های بعدی از مسیری عبور می‌کنند که میزان غلظت فرومون، بیشترین مقدار باشد (۲۹). جستجو برای زیر مجموعه‌ای بهینه از ویژگی‌ها، بر روی گراف ادامه دارد به‌گونه‌ای‌که از حداقل تعداد گره‌ها عبور شود. این عملیات با استفاده از یک معیار توقف مناسب (مانند معیار دقت)، پایان می‌پذیرد (۳۰).

در این بخش برای ایجاد مدل پیش‌بینی و تشخیص مرحله بالینی سرطان پستان، بجای استفاده از ویژگی‌های موجود در سیستم TNM، از ویژگی‌هایی استفاده شد که توسط الگوریتم کلونی مورچگان انتخاب شدند. در این گام نیز ایجاد مدل‌های مختلف دسته‌بندی همانند بخش ۲-۴ انجام گردید. با این تفاوت که مدل‌هایی که در بخش ۲-۴ ایجاد شدند بر اساس ویژگی‌های T، N و M بودند، اما مدل‌هایی که در این گام ایجاد شدند بر اساس ویژگی‌هایی است که توسط الگوریتم کلونی مورچگان انتخاب شده‌اند. این گام نیز در هر دو دیتاست SEER و محلی یکسان می‌باشد. جدول ۱، مجموعه ویژگی‌های به‌کار رفته در

[6] 10-Fold Cross Validation
[7] Accuracy
[8] Area Receiver Operating Characteristic (AROC)
[9] Feature Selection (FS)

سیستم TNM و ویژگی‌های انتخاب شده توسط الگوریتم کلونی مورچگان این پژوهش را نشان می‌دهد.

همان‌گونه که از جدول ۱ مشخص است الگوریتم کلونی مورچگان با استفاده از دیتاست SEER علاوه بر دو ویژگی T و N، ویژگی‌های دیگری مانند Histology، Birthplace، CS Extension، Regional Node Positive، Recode ICD-0-2 to 10 (Morphology and Primary Site)، CS Site-Specific Factor 2، CS Site-Specific Factor 3 و CS Site-Specific Factor 6 را نیز به عنوان ویژگی‌های مهم برای تعیین مرحله سرطان شخص انتخاب کرده است. با توجه به اینکه فیلد متاستاز در دیتاست SEER وجود داشته است، اما الگوریتم کلونی مورچگان آن ویژگی را انتخاب نکرده است.

همچنین در دیتاست محلی، این الگوریتم علاوه بر سه ویژگی T، N و M ویژگی‌های دیگری مانند درجه تمایز تومورهای سرطانی(Grade)، تهاجم عروقی (Vascular Invasion)، سن بیمار، گروه خونی شخص، تعداد فرزندان را نیز به عنوان ویژگی‌های مهم برای تعیین مرحله سرطان شخص انتخاب کرده است.

جدول ۱: ویژگی‌های مورد استفاده در مدل‌های پیشنهادی این پژوهش

سیستم TNM		ویژگی‌های مورد استفاده در سیستم پیشنهادی تعیین مرحله سرطان پستان
		$M_1, M_0, N_3, N_2, N_1, N_0, T_4, T_3, T_2, T_1, T_0$
ویژگی‌های استخراجی توسط الگوریتم کلونی مورچگان	دیتاست محلی	T، N، M، درجه تمایز تومورهای سرطانی(Grade)، تهاجم عروقی(Vascular Invasion)، سن بیمار، گروه خونی شخص، تعداد فرزندان
	دیتاست SEER	T, N, Histology, Birthplace, CS Extension, Regional Node Positive, Recode ICD-O-2 to 10 (Morphology and Primary Site), Site-Specific Factor 2, Site-Specific Factor 3, Site-Specific Factor 6

یافته‌ها

در این پژوهش، مدلی برای پیش‌بینی و همچنین تشخیص مرحله سرطان پستان معرفی شده است. دو روش استاندارد ایجاد مدل، یکی با استفاده از سیستم TNM و دیگری با استفاده از ویژگی‌هایی که توسط الگوریتم کلونی مورچگان انتخاب شده‌اند، پیشنهاد گردید. برای نشان دادن کارایی روش‌های پیشنهادی، در ابتدا از دیتاست SEER و سپس از دیتاست محلی استفاده شده است. اکثر طبقه‌بندهای معروف مانند درخت تصمیم، IBK، KNN، SMO، SVM و... (که اکثر محققان از آنها استفاده می‌کنند) بر روی داده‌ها اعمال و دو معیار "دقت" و "سطح زیر نمودار راک" هر طبقه‌بند محاسبه گردید و نهایتا، بهترین طبقه‌بند جهت ایجاد مدل پیشنهادی این پژوهش انتخاب گردید. جدول ۲ کلاس‌های مختلف طبقه‌بندها و اعضای موجود در هر کدام را نشان می‌دهد.

نتایج حاصل از آزمایش سیستم TNM: طبقه‌بندهای مختلفی بر روی متغیرهای سیستم TNM اعمال و دو شاخص "دقت" و "سطح زیر نمودار راک" آنها محاسبه گردید. جدول ۳، دقت و سطح زیر نمودار راک حاصل از استفاده از این مجموعه طبقه‌بندها با استفاده از سیستم TNM را برای دیتاست SEER و دیتاست محلی ایجاد شده، نمایش می‌دهد. در این جدول، بهترین دقت‌های بدست آمده برای طبقه‌بند گروه‌های شش‌گانه جدول ۲ (از هر گروه فقط یک طبقه‌بند) با خانه‌های رنگی مشخص شده‌اند. مشخص شده‌اند. لازم به ذکر است که تمامی طبقه‌بندهای معرفی شده در گروه‌های شش گانه جدول ۲ برای آزمایش‌های این پژوهش، به‌کار گرفته شدند. لیکن بعضی از آنها نتایج ضعیفی تولید کرده که در نتیجه از درج آن نتایج در جداول آتی خودداری گردید. به عنوام مثال، در گروه طبقه‌بندهای Bayes، طبقه‌بند Naïve

[DOR: 20.1001.1.17359406.1397.11.3.5.5]
DOI: 10.30699/acadpub.ijbd..11.3.55]

Bayes Multinomial Text نیز وجود دارد که استفاده از آن برای دیتاست SEER منجر به تولید دقتی برابر با ۳۲/۹۸٪ و سطح زیر نمودار راک ۰/۴۹۴ گردید که این نتایج در مقایسه با نتایج حاصل از سه طبقه‌بند دیگر این گروه، بسیار ضعیف‌تر بوده که در نتیجه منجر به حذف این طبقه‌بند از جدول نتایج شده است. در سایر گروه‌های جدول ۲ نیز طبقه‌بندهایی وجود دارند که از ذکر نتایج ضعیف آنها خودداری شده است.

جدول ۲: کلاس‌های مختلف طبقه‌بندها و اعضای موجود در هر کلاس

گروه طبقه‌بند	اعضای گروه
Bayes	BayesNet, Naïve Bayes, Naïve Bayes Updatable, Naïve Bayes Multinomial Text
Functions	Multi Layer Perceptron, Simple Logistic, Logistic, SMO
Lazy	KStar, LWL, IBK
Meta	Ada Boost M1, Attribute Selected Classifier, Bagging, Classification Via Regression, CV Parameter Selection, Filtered Classifier, Iterative Classifier Optimizer, Logit Boost, Multi Class Classifier, Multi Class Classifier Updatable, Multi Scheme, Random Committee, Randomizable Filtered Classifier, Random Subspace, Stacking, Vote, Weighted Instances Handler Wrapper
Rules	Decision Table, JRip, PART, OneR, ZeroR
Trees	Decision Stump, Hoeffding Tree, Random Tree, Random Forest, LMT, J48, REPTree

همان‌گونه که از جدول ۳ مشخص است، برای دیتاست SEER، طبقه‌بند Logistic با دقتی برابر با ۹۹/۹۳٪ و سطح زیر نمودار راک ۱ و برای دیتاست محلی، طبقه‌بندهای Multi class classifier و Logistic با دقتی برابر با ۹۹/۹۱٪ و سطح زیر نمودار راک ۱، بالاترین کارایی و دقت را در مجموعه طبقه‌بندهای موجود برای تشخیص مرحله سرطان پستان داشته‌اند.

نتایج حاصل از آزمایش سیستم با استفاده از الگوریتم کلونی مورچگان: در این بخش نیز طبقه‌بندهای مختلفی بر روی متغیرهایی که با استفاده از الگوریتم کلونی مورچگان انتخاب شدند، اعمال و دو شاخص "دقت" و "سطح زیر نمودار راک" آنها محاسبه گردید. جدول ۴، دقت و سطح زیر نمودار راک حاصل از استفاده از این مجموعه طبقه‌بندها با استفاده از الگوریتم کلونی مورچگان را برای دیتاست SEER و دیتاست محلی ایجاد شده، نمایش می‌دهد. در این جدول بهترین دقت‌های بدست آمده برای طبقه‌بند گروه‌های شش‌گانه جدول ۲ (از هر گروه فقط یک طبقه‌بند) با خانه‌های رنگی مشخص شده‌اند.

نکته: برای ساخت مدل در زمان آموزش و آزمایش سیستم TNM در دیتاست محلی، جایگزینی داده‌های مفقود با مقدار میانگین، باعث شده است که در برخی از الگوریتم‌ها مانند Logistic، SimpleLogistic و LMT مقدار دقت برابر با ۱۰۰٪ بدست آید. به منظور انجام ارزیابی منطقی‌تر، طبقه‌بندی کننده‌هایی که دقت آنها ۱۰۰٪ نبودند، در مقایسه‌ها مورد استفاده قرار گرفته‌اند.

همان‌گونه که از جدول ۴ مشخص است، برای دیتاست SEER، طبقه‌بند Random Forest با دقتی برابر با ۹۹/۴۳٪ و سطح زیر نمودار راک ۱ و برای دیتاست محلی، طبقه‌بند JRip با دقتی برابر با ۹۸/۹۵٪ با سطح زیر نمودار راک ۰/۹۹۳، بالاترین کارایی و دقت را در مجموعه طبقه‌بندهای موجود برای تشخیص مرحله سرطان پستان داشته‌اند.

گزارش خطا: میزان و رخداد خطاها در حوزه پزشکی به‌خصوص در زمینه تشخیص مرحله سرطان، بسیار حایز اهمیت است. از آنجا که نوع و طول دوره درمان افراد بسته به مراحل مختلف بیماری، متفاوت است، لذا اگر مرحله سرطان یک بیمار دیر یا به اشتباه در مرحله دیگری تشخیص داده شود، می‌تواند عواقب جبران ناپذیری داشته باشد. در چنین شرایطی شانس زنده ماندن بیمار نیز پایین خواهد آمد. لذا در این پژوهش پس از محاسبه معیارهای دقت و سطح زیر نمودار راک هر طبقه‌بند، خطاهای تشخیص مرحله سرطان، مورد بررسی و ارزیابی قرار گرفت.

جدول ۳: دقت و سطح زیر نمودار راک طبقه‌بندهای مختلف با استفاده از سیستم TNM

الگوریتم	طبقه‌بندهای گروه	دیتاست SEER		دیتاست محلی	
		دقت (٪)	سطح زیر نمودار راک	دقت (٪)	سطح زیر نمودار راک
Bayes	BayesNet	۹۳/۵۹	۰/۹۹۲	۹۴/۶۸	۰/۹۹۲
	NaiveBayes	۹۴/۵۹	۰/۹۹۴	۹۴/۶۰	۰/۹۹۳
	NaiveBayesUpdateable	۹۴/۵۹	۰/۹۹۴	۹۴/۶۰	۰/۹۹۳
Functions	Multi Layer Perceptron	۹۸/۳۶	۰/۹۸۹	۹۹/۰۴	۰/۹۹۵
	Simple Logistic	۹۹/۷۹	۱/۰۰	۹۹/۶۵	۱/۰۰
	Logistic	۹۹/۹۳	۱/۰۰	۹۹/۹۱	۱/۰۰
	SMO	۹۹/۴۳	۱/۰۰	۹۹/۰۴	۰/۹۹۷
Lazy	IBK	۹۹/۷۲	۰/۹۹۹	۹۹/۳۰	۰/۹۹۹
	LWL	۸۶/۵۴	۰/۹۸۷	۸۶/۴۱	۰/۹۹۱
	KStar	۹۶/۸۷	۰/۹۹۹	۹۵/۶۴	۱/۰۰
Meta	Attribute Selected Classifier	۹۸/۳۶	۰/۹۹۷	۸۴/۰۶	۰/۹۶۵
	Bagging	۹۷/۵۸	۰/۹۹۶	۹۷/۶۵	۰/۹۹۵
	Classification Via Regression	۹۷/۵۱	۱/۰۰	۹۲/۹۴	۰/۹۷۴
	Filtered Classifier	۹۸/۹۳	۰/۹۹۸	۹۸/۷۸	۰/۹۹۵
	Iterative Classifier Optimizer	۹۶/۳۰	۰/۹۹۹	۹۸/۰۸	۰/۹۹۹
	Logit Boost	۹۶/۳۰	۰/۹۹۹	۹۸/۰۸	۰/۹۹۹
	Multi Class Classifier	۹۷/۱۵	۱/۰۰	۹۹/۹۱	۱/۰۰
	Multi Class Classifier Updatable	۹۲/۲۴	۰/۹۹۹	۹۳/۸۲	۱/۰۰
	Random Committee	۹۹/۷۲	۰/۹۹۹	۹۹/۶۵	۱/۰۰
	Randomizable Filtered Classifier	۹۸/۹۳	۰/۹۹۹	۹۹/۰۴	۰/۹۹۹
	Random Subspace	۹۴/۵۲	۰/۹۹۶	۹۴/۸۶	۰/۹۹۸
Rules	Decision Table	۹۸/۶۵	۱/۰۰	۹۸/۶۱	۰/۹۹۹
	JRip	۹۸/۰۸	۰/۹۹۷	۹۹/۳۹	۰/۹۹۹
	PART	۹۹/۳۶	۰/۹۹۹	۹۹/۰۴	۰/۹۹۷
	OneR	۵۶/۰۵	۰/۷۰۳	۳۸/۸۵	۰/۶۱۴
Trees	J48	۹۸/۹۳	۰/۹۸۷	۹۸/۷۸	۰/۹۹۵
	LMT	۹۹/۷۹	۱/۰۰	۹۹/۶۵	۱/۰۰
	Random Tree	۹۹/۶۴	۰/۹۹۸	۹۹/۵۶	۰/۹۹۷
	Random Forest	۹۹/۵۷	۱/۰۰	۹۹/۵۶	۱/۰۰

از آنجا که بروز خطا در طبقه‌بندها و مدل‌های پیش‌بینی، اجتناب ناپذیر است، لذا بایستی خطاهای موجود در سیستم شناخته و بررسی شوند. در این تحقیق برای پرداختن به این مسئله مهم، از ماتریس برخورد[10] استفاده شده است. این ماتریس برای تمام مدل‌های دسته‌بندی، تولید شد. اما در ادامه تنها چهار ماتریس برخورد مربوط به طبقه‌بندهایی که دقت بالایی داشته‌اند، نمایش داده شده‌اند.

در مدل تشخیص مرحله سرطان، که توسط الگوریتم Logistic و با استفاده از داده‌های بین‌المللی SEER ایجاد گردید، ۱۴۰۳ نمونه به درستی و یک نمونه به اشتباه تشخیص داده شد. همان‌گونه که در جدول ۵ نشان داده شده است،

[10] Confusion Matrix

[DOR: 20.1001.1.17359406.1397.11.3.5.5]
DOI: 10.30699/acadpub.ijbd..11.3.55]

جدول ۴: دقت و سطح زیر نمودار راک طبقه‌بندهای مختلف با استفاده از الگوریتم کلونی مورچگان

		دیتاست SEER		دیتاست محلی	
الگوریتم	طبقه‌بندهای گروه	دقت (٪)	سطح زیر نمودار راک	دقت (٪)	سطح زیر نمودار راک
Bayes	BayesNet	۹۴/۳۷	۰/۹۹۸	۹۱/۱۲	۰/۹۸۵
	NaiveBayes	۶۷/۵۲	۰/۹۳۹	۸۵/۲۸	۰/۹۶۹
	NaiveBayesUpdateable	۶۷/۵۲	۰/۹۳۹	۸۵/۲۸	۰/۹۶۹
Functions	Multi Layer Perceptron	۹۵/۰۹	۰/۹۸۵	۸۸/۷۶	۰/۹۶۸
	Simple Logistic	۶۵/۶۰	۰/۹۳۳	۹۷/۰۴	۰/۹۹۳
	Logistic	۶۷/۸۱	۰/۹۳۴	۹۵/۹۱	۰/۹۸۵
	SMO	۶۰/۳۳	۰/۸۳۸	۸۹/۷۲	۰/۹۶۹
Lazy	IBK	۹۸/۹۳	۰/۹۹۸	۸۴/۴۱	۰/۹۲۷
	LWL	۶۳/۸۲	۰/۹۶۳	۶۴/۰۲	۰/۹۴۷
	KStar	۹۵/۱۶	۰/۹۹۹	۸۹/۹۰	۰/۹۷۶
Meta	Attribute Selected Classifier	۹۹/۱۵	۰/۹۹۸	۹۸/۵۲	۰/۹۸۷
	Bagging	۹۸/۰۱	۰/۹۹۹	۹۴/۶۰	۰/۹۸۲
	Classification Via Regression	۹۹/۰۰	۰/۹۹۹	۹۷/۱۳	۰/۹۹۲
	Filtered Classifier	۹۷/۷۹	۰/۹۹۶	۹۷/۷۴	۰/۹۸۳
	Iterative Classifier Optimizer	۹۸/۰۱	۰/۹۹۹	۹۶/۵۲	۰/۹۹۳
	Logic Boost	۹۸/۰۱	۰/۹۹۹	۹۶/۵۲	۰/۹۹۳
	Multi Class Classifier	۵۹/۶۲	۰/۹۲۵	۸۹/۹۰	۰/۹۷۹
	Multi Class Classifier Updatable	۵۱/۹۹	۰/۷۷۳	۸۴/۷۶	۰/۹۴۵
	Random Committee	۹۹/۳۶	۰/۹۹۸	۹۵/۷۳	۰/۹۹۳
	Randomizable Filtered Classifier	۹۷/۷۹	۰/۹۹۶	۸۴/۵۸	۰/۹۲۸
	Random Subspace	۹۸/۸۶	۱/۰۰	۹۲/۸۶	۰/۹۸۸
Rules	Decision Table	۹۵/۴۴	۰/۹۹۷	۹۶/۵۲	۰/۹۸۶
	JRip	۹۸/۷۹	۰/۹۹۷	۹۸/۹۵	۰/۹۹۳
	PART	۹۹/۲۲	۰/۹۹۹	۹۸/۲۶	۰/۹۸۹
	OneR	۷۱/۶۵	۰/۸۱۴	۶۷/۶۸	۰/۸۰۱
Trees	J48	۹۹/۱۵	۰/۹۹۸	۹۸/۴۳	۰/۹۸۷
	LMT	۹۵/۷۹	۰/۹۹۸	۹۶/۹۵	۰/۹۸۹
	Random Tree	۹۸/۷۲	۰/۹۹۴	۹۱/۷۲	۰/۹۵۰
	Random Forest	۹۹/۴۳	۱/۰۰	۹۵/۷۳	۰/۹۹۵

شخصی در مرحله چهار بیماری قرار داشته است، در صورتی‌که به اشتباه مرحله سرطان وی، مرحله صفر تشخیص داده شده است. این خطا در تشخیص باعث می‌شود شخصی که نیاز به درمان پیچیده‌ای دارد و باید تحت درمان داروهای شیمیایی، پرتودرمانی و جراحی قرار گیرد، به اشتباه سالم تشخیص داده شود. البته لازم به ذکر است که با توجه به دقت بسیار زیاد سیستم پیشنهادی در تعیین صحیح تمامی نمونه‌ها بجز این یک مورد،

لذا این احتمال نیز وجود دارد که اصولا اطلاعات مربوط به این بیمار در دیتاست، از ابتدا اشتباه ثبت شده باشد که منجر به این خطای شناسایی نیز شده است.

جدول ۵: ماتریس برخورد مربوط به الگوریتم Logistic برای دیتاست SEER با استفاده از سیستم TNM

A	B	C	D	E	F	G	H	Classified as
۴۶۳	۰	۰	۰	۰	۰	۰	۰	**A: stage1A**
۰	۳۳۰	۰	۰	۰	۰	۰	۰	**B: stage0**
۰	۰	۲۷۶	۰	۰	۰	۰	۰	**C: stage2A**
۰	۰	۰	۱۰۶	۰	۰	۰	۰	**D: stage2B**
۰	۰	۰	۰	۴۵	۰	۰	۰	**E: stage3B**
۰	۰	۰	۰	۰	۹۲	۰	۰	**F: stage3A**
۰	۱	۰	۰	۰	۰	۵۷	۰	**G: stage4**
۰	۰	۰	۰	۰	۰	۰	۳۴	**H: stage3C**

جدول ۶: ماتریس برخورد مربوط به الگوریتم RandomForest برای دیتاست SEER با استفاده از الگوریتم کلونی مورچگان

A	B	C	D	E	F	G	H	Classified as
۴۶۳	۰	۰	۰	۰	۰	۰	۰	**A: stage1A**
۰	۳۲۸	۰	۰	۰	۰	۲	۰	**B: stage0**
۰	۰	۲۷۶	۰	۰	۰	۰	۰	**C: stage2A**
۰	۰	۰	۱۰۶	۰	۰	۰	۰	**D: stage2B**
۰	۰	۰	۰	۴۵	۰	۰	۰	**E: stage3B**
۰	۰	۰	۰	۰	۹۲	۰	۰	**F: stage3A**
۰	۲	۰	۱	۰	۰	۵۴	۱	**G: stage4**
۰	۰	۰	۰	۰	۰	۰	۳۴	**H : stage3C**

جدول ۷: ماتریس برخورد مربوط به الگوریتم MultiClassClassifier برای دیتاست محلی با استفاده از سیستم TNM

A	B	C	D	E	F	G	H	Classified as
۵۷	۰	۰	۰	۰	۰	۰	۰	**A: stage4**
۰	۲۱۹	۰	۰	۰	۰	۰	۰	**B: stage2A**
۰	۰	۲۴۰	۰	۰	۰	۰	۰	**C: stage3A**
۰	۰	۰	۱۷۰	۰	۰	۰	۰	**D: stage2B**
۰	۰	۰	۰	۲۰۶	۰	۰	۰	**E: stage3C**
۰	۰	۰	۰	۰	۱۳۰	۰	۰	**F: stage1A**
۰	۰	۰	۰	۰	۰	۱۷	۰	**G: stage3B**
۰	۰	۱	۰	۰	۰	۰	۱۰۸	**H: stage0**

جدول ۸ : ماتریس برخورد مربوط به الگوریتم JRip برای دیتاست محلی با استفاده از الگوریتم کلونی مورچگان

A	B	C	D	E	F	G	Classified as
۵۶	۱	۳	۱	۳	۱	۳	**A : stage4**
۰	۲۱۹	۰	۲	۰	۰	۰	**B : stage2A**
۰	۰	۳۳۲	۰	۰	۰	۰	**C : stage3A**
۰	۰	۰	۱۷۲	۰	۰	۰	**D : stage2B**
۰	۰	۰	۰	۲۰۸	۰	۰	**E : stage3C**
۰	۰	۰	۰	۰	۱۳۰	۰	**F : stage1A**
۰	۰	۰	۰	۰	۰	۱۷	**G : stage3B**

[DOR: 20.1001.1.17359406.1397.11.3.5.5]
DOI: 10.30699/acadpub.ijbd..11.3.55]

در مدلی که توسط الگوریتم MultiClassClassifier و با استفاده از دیتاست محلی این پژوهش ایجاد گردید، ۱۱۴۷ نمونه به درستی و یک نمونه به اشتباه در مرحله دیگری تشخیص داده شده است. همان‌گونه که در جدول ۷ نشان داده شده است، یک نفر در stage0 قرار داشته که مدل به اشتباه آن را در stage3A تشخیص داده است. یعنی شخصی که سالم بوده و نیاز به هیچ درمانی نداشته است در مرحله سوم سرطان تشخیص داده شده است. بدیهی است که شروع غلط درمان چنین فردی باعث می‌شود که آسیب‌های جسمی، روحی و روانی شدیدی به وی وارد گردد. البته لازم به ذکر است که با توجه به دقت بسیار زیاد سیستم پیشنهادی در تعیین صحیح تمامی مونه‌ها بجز این یک مورد، لذا این احتمال نیز وجود دارد که اصولا اطلاعات مربوط به این بیمار در دیتاست، از ابتدا اشتباه ثبت شده باشد که منجر به این خطای شناسایی نیز شده است.

در مدل دیگری که توسط الگوریتم JRip و با استفاده از داده‌های محلی این پژوهش ایجاد گردید، ۱۱۳۴ نمونه به درستی و ۱۴ نمونه به اشتباه در مراحل دیگر تشخیص داده شده‌اند. همان‌گونه که در جدول ۸ نشان داده شده است، دو نفر در stage2A قرار داشته‌اند در صورتی که مدل به اشتباه آنها را در stage2B تشخیص داده است. همچنین ۱۲ نفر دیگر در مرحله چهار (مرحله بسیار پیشرفته) قرار داشته‌اند اما مدل پیش‌بینی به اشتباه، یکی از آنها را در مرحله stage2A، سه نفر از آنها را در مرحله stage3A، یکی دیگر از آنها را در مرحله stage2B، سه بیمار دیگر را در مرحله stage3C و به‌ترتیب یک و سه بیمار دیگر را در stage1A و stage3B قرار داده است. در این حالت هیچ نمونه‌ای در stage0 تشخیص داده نشده است بنابراین ماتریس برخورد این الگوریتم نیز فاقد ردیف stage0 می‌باشد.

بحث

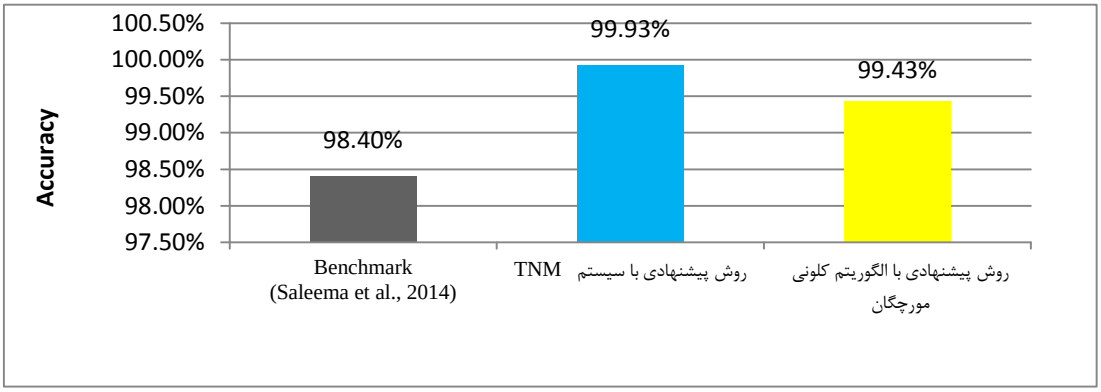
از آنجا که در سال‌های اخیر، تشخیص بیماری‌های مختلف، به خصوص تشخیص سرطان، مراحل این بیماری و پیامدهای آن، به یک چالش بزرگ در حوزه سلامت تبدیل شده است و حتی متداول‌ترین تکنیک‌های تشخیصی، مانند ماموگرافی، نیز نمی‌توانند قابلیت تشخیص بالایی داشته باشند، لذا معرفی تکنیک‌های تشخیصی موثرتر، بسیار حایز اهمیت است. به همین دلیل امروزه توجه قابل ملاحظه‌ای به مدل‌های آماری، جهت دسته‌بندی و طبقه‌بندی داده‌های پزشکی شده است. برای پیشبرد اهداف این پژوهش و نشان دادن کارایی روش پیشنهادی و مقایسه آن با پژوهش‌های صورت گرفته در سطح بین‌المللی، علاوه بر دیتاست محلی، از دیتاست بین‌المللی SEER، با هدف گسترش مدل پیش‌بینی و تشخیص مرحله سرطان پستان استفاده گردید. سیستم تشخیصی پیشنهادی این پژوهش قادر است در فرآیند تشخیص سرطان پستان، دقت را افزایش، زمان و خطای احتمالی کارشناسان و هزینه‌های درمان را کاهش و جزئیات را دقیق‌تر معرفی نماید.

بر اساس پژوهشی که توسط سلیما و همکاران وی (۳۱) جهت تشخیص مرحله سرطان پستان و با استفاده از دیتاست SEER صورت گرفت، الگوریتم درخت تصمیم از بالاترین دقت برخوردار بود که با استفاده از نمونه‌برداری طبقه‌بندی شده متعادل، به دقتی برابر با ۹۸/۴۰٪ دست یافتند. اما در روش پیشنهادی این پژوهش، با استفاده از سیستم TNM و الگوریتم کلونی مورچگان، به ترتیب دقت‌های ۹۹/۹۳٪ و ۹۹/۴۳٪ بدست آمد (شکل ۱). در پژوهش حاضر، شاخص سطح زیر نمودار راک هر طبقه‌بند نیز مورد ارزیابی و بررسی قرار گرفته است و حداکثر مقدار بدست آمده برای سطح زیر نمودار راک برابر با ۱ و حداقل مقدار بدست آمده برابر با ۰/۴۹۴ بوده است. در حالی‌که در پژوهش سلیما و همکاران (۳۱)، به این مهم پرداخته نشده است.

همان‌گونه که از شکل ۱ مشخص است، روش پیشنهادی این پژوهش در دیتاست SEER، با دقتی بیشتر و به‌صورت موثر، مرحله سرطان پستان را تشخیص می‌دهد.

برخی از دیتاست‌های بین‌المللی با وجود داشتن حجم قابل توجهی از اطلاعات، قابل استفاده برای تمام کشورها نبوده و واقعیت‌های پزشکی مربوط به کشورهای مختلف را نشان نمی‌دهند. چون هر کشوری از شرایط محیطی و پزشکی متفاوتی برخوردار است (۶). به منظور ارزیابی دقیق‌تر مدل پیشنهادی، علاوه بر دیتاست بین‌المللی SEER، از یک دیتاست محلی و واقعی که مختص پژوهش حاضر است نیز استفاده گردید تا بدین وسیله بتوان مشکلات، واقعیت‌ها و محدودیت‌های موجود در ایران را بررسی کرد.

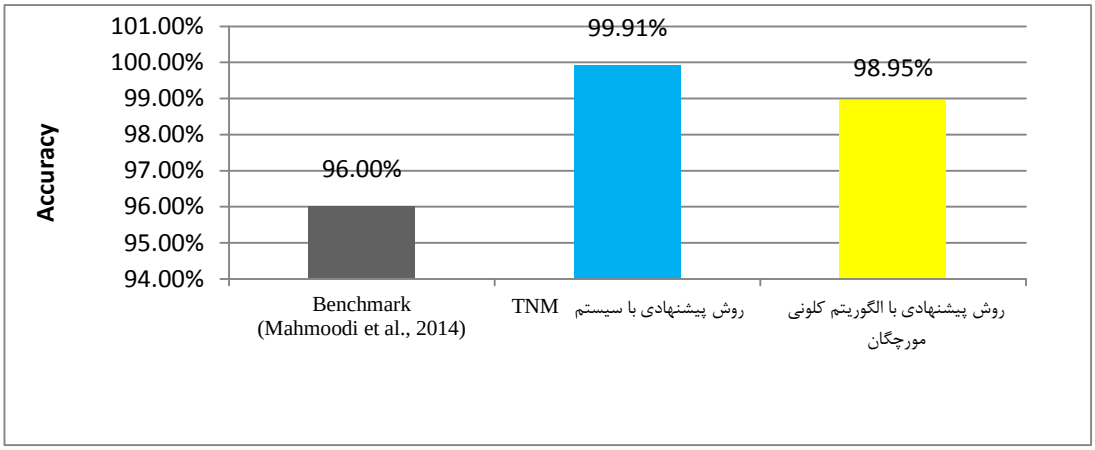
ذکر این نکته بسیار مهم ضروری است که برای یک مقایسه عادلانه، باید تمام شرایط، از جمله داده‌های هر دو پژوهش، با هم یکسان باشند. در بخش استفاده از دیتاست محلی نیز نیاز به یک Benchmark است.



شکل ۱: مقایسه دقت روش پیشنهادی پژوهش با یافته‌های سلیما و همکاران (۳۱) برای دیتاست SEER

لیکن در پژوهش‌های محلی موجود، متاسفانه محققین دیتاست اختصاصی خود را به‌طور عمومی در اختیار سایرین قرار نمی‌دهند. در نتیجه، هر پژوهش منحصرا یافته‌های خود را اعلام و فقط نتایج تحقیق‌های سایرین را با کار خود مقایسه می‌کند (۶، ۳۲-۳۴). در پژوهش حاضر علی‌رغم تلاش‌های مکرر برای بدست آوردن دیتاست سایر محققین، این امر محقق نگردید. نهایتا در یکی از مقایسه‌های انجام شده، نتایج این پژوهش با نتایج حاصل از تحقیق محمودی و همکاران وی (۳۳) مقایسه گردید، علی‌رغم آن‌که این دو تحقیق از داده‌های یکسانی برخوردار نیستند. در واقع یکی دیگر از دلایل پیاده‌سازی روش پیشنهادی این پژوهش با استفاده از دیتاست بین‌المللی SEER، همین موضوع بوده است تا بتوان به طور واقعی، کارایی سیستم تشخیص مرحله سرطان را در شرایط عادلانه و منصفانه ارزیابی نمود.

بر اساس پژوهشی که توسط محمودی و همکاران وی (۳۳) جهت تشخیص مرحله سرطان پستان و با استفاده از داده‌های جمع‌آوری شده از بیمارستان ولی‌عصر بیرجند صورت گرفته است، الگوریتم نزدیک‌ترین همسایه با دقتی برابر با ۹۶٪ بیشترین کارایی را داشته است. اما در روش پیشنهادی این پژوهش و با استفاده از داده‌هایی که از مراکز سرطان شیراز جمع‌آوری شد و با به‌کارگیری سیستم TNM و الگوریتم کلونی مورچگان به ترتیب، دقت‌های ۹۹/۹۱٪ و ۹۸/۹۵٪ بدست آمد. در این پژوهش، شاخص سطح زیر نمودار راک هر طبقه‌بند نیز، برای داده‌های محلی مورد ارزیابی و بررسی قرار گرفته است. در حالی‌که در پژوهش محمودی و همکاران وی، این معیار کلیدی مورد بررسی قرار نگرفته است. حداکثر مقدار بدست آمده برای سطح زیر نمودار راک برابر با ۱ و حداقل مقدار بدست آمده برابر با ۰/۴۹۵ بوده است. نتایج مقایسه دقت روش پیشنهادی پژوهش با یافته‌های محمودی و همکاران وی (۳۳) با استفاده از یک دیتاست محلی در شکل ۲ نشان داده شده است. همان‌گونه که در شکل ۲ مشاهده می‌شود، روش پیشنهادی این پژوهش در دیتاست محلی نیز، با دقت بالایی مرحله سرطان پستان را تشخیص داده است.



شکل ۲: مقایسه دقت روش پیشنهادی پژوهش با یافته‌های محمودی و همکاران (۳۳) برای دیتاست محلی

نتیجه‌گیری

پیش‌بینی زودهنگام و تشخیص مرحله بیماران مبتلا به سرطان پستان و پیشگیری از آن، با توجه به شیوع بالای آن در سراسر دنیا، به‌عنوان یکی از بهترین رویکردها در جهت کنترل این بیماری، بسیار حایز اهمیت است. هدف از این تحقیق، گسترش مدلی برای تشخیص و پیش‌بینی مرحله سرطان پستان بوده، به‌طوری‌که قادر باشد این بیماری را با بالاترین دقت و کمترین هزینه تشخیص دهد. در این مطالعه برای انتخاب مهم‌ترین و موثرترین ویژگی‌ها، از الگوریتم کلونی مورچگان استفاده گردید.

در این پژوهش و بر اساس نتایج بدست آمده از دیتاست بین‌المللی SEER، با استفاده از سیستم TNM و الگوریتم کلونی مورچگان، به ترتیب دقت‌های ۹۹/۹۳٪ و ۹۹/۴۳٪ توسط الگوریتم‌های Logistic و Random Forest بدست آمد. همچنین با استفاده از داده‌هایی که از مراکز سرطان شیراز جمع‌آوری شد، با به‌کارگیری سیستم TNM و الگوریتم کلونی مورچگان به ترتیب دقت‌های ۹۹/۹۱٪ و ۹۸/۹۵٪ توسط الگوریتم‌های MultiClassClassifier و JRip حاصل شد. همچنین در این تحقیق کمترین خطا مربوط به طبقه‌بند Logistic در داده‌های SEER و طبقه‌بند MultiClassClassifier در دیتاست محلی بوده است. یافته‌های این پژوهش نشان می‌دهند که علاوه بر ویژگی‌های مورد استفاده در سیستم رایج TNM، عوامل جدید دیگر همانند تهاجم عروقی، سن بیمار، گروه خونی، تعدا فرزندان، محل تولد، بافت شناسی سلولی، نوع بافت درگیر و Site-Specific Factorهای شماره ۲، ۳ و ۶ نیز می‌توانند توسط متخصصین سرطان شناسی در تعیین مرحله بیماری بیماران مبتلا به سرطان پستان نیز استفاده شوند. نهایتا آن‌که مدل پیش‌بینی و تشخیصی مرحله سرطان پستان پیشنهادی این پژوهش، می‌تواند برای سرطان‌های دیگر مانند روده، کبد و ریه نیز به‌کار رود.

تقدیر و تشکر

این مقاله مستخرج از پایان نامه کارشناسی ارشد سرکار خانم سعیده ناصری نوروزانی در دانشگاه آزاد اسلامی شیراز است.

تعارض منافع

نویسندگان اعلام می‌دارند که هیچ تعارض منافعی در پژوهش حاضر وجود ندارد.

References

1. Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services 2012; 2(1):17.
2. Padmapriya B, Velmurugan T. A survey on breast cancer analysis using data mining techniques. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) 2014; 1234-7.
3. Rastghalam R, Pourghassem H. Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images. Pattern Recognition 2016; 51:176-86.
4. Huang ML, Hung YH, Lee WM, Li RK, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. Journal of medical systems 2012; 36(2):407-14.
5. Asadabadi A, Bahrampour A, Haghdoost AA. Prediction of Breast Cancer Survival by Logistic Regression and Artificial Neural Network Models. Iranian Journal of Epidemiology 2014; 10(3):1-8.
6. García-Laencina PJ, Abreu PH, Abreu MH, Afonoso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in biology and medicine 2015; 59:125-33.
7. Shirvani H, Alhani F, Montazeri A. The effect of family-centered empowerment model on the functional scales quality of

life in women with breast cancer undergoing chemotherapy. Iranian Journal of Breast Disease 2017; 10(1):62-72.

8. Salama GI, Abdelhalim M, Zeid MA. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 2012; 32(569):2.

9. Mandal SK. Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. International Journal of Engineering and Computer Science 2017; 6(2):20388-91.

10. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications 2014; 41(4): 1476-82.

11. Senturk ZK, Kara R. Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering 2014; 4(1):35.

12. Brunicardi F, Andersen D, Billiar T, Dunn D, Hunter J, Matthews J, Pollock R. Schwartz's principles of surgery, 10e. McGraw-hill 2014.

13. Askarpour S, Ghafoori S-H. Survey data mining algorithms in the field of breast cancer. 3rd E-conference on Recent Research in Science and Technology, 2015.

14. Tahergorabi Z, Moodi M, Mesbahzadeh B. Breast Cancer: A preventable disease. Journal of Birjand University of Medical Sciences 2014; 21(2):126-41.

15. Ranjkesh M, Fathi Azar F, Ghatreh Samani F, Tarzamni MK, Vali Khani E. Evaluation of adjunctive sonography results in screening of women with mammographically dense breasts for early diagnosis of breast cancer. Iranian Journal of Breast Disease 2017; 10(1):7-19.

16. Jedy-Agba E, McCormack V, Adebamowo C, dos-Santos-Silva I. Stage at diagnosis of breast cancer in sub-Saharan Africa: a systematic review and meta-analysis. The Lancet Global Health 2016; 4(12):923-5.

17. Sheikhpour R, Agha Sarram M, Zare Mirakabad MR, Sheikhpour R. Breast Cancer Detection Using Two-Step Reduction of Features Extracted From Fine Needle Aspirate and Data Mining Algorithms. Iranian Journal of Breast Disease 2015; 7(4):43-51.

18. Copeland MM. American joint committee on cancer staging and end results reporting. Objectives and progress. Cancer 1965; 18(12):1637-40.

19. Sobin LH, Fleming ID. TNM classification of malignant tumors, (1997). Cancer: Interdisciplinary International Journal of the American Cancer Society 1997; 80(9):1803-4.

20. Olfatbakhsh A, Haghighat S, Khani M, Beheshtian T, Alavi N, Sari F, Hosseinpour P. Evaluation of Ultrasound Accuracy in Axillary Lymph Node Involvement in Breast Cancer Patients. Iranian Journal of Breast Disease 2017; 10(2):7-15.

21. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians 2017; 67(2):93-9.

22. Copeland MM. Clinical staging system for cancer and end results reporting. CA: a cancer journal for clinicians 1961; 11(2): 42-7.

23. Benson JR. The TNM staging system and breast cancer. The lancet oncology 2003; 4(1):56-60.

24. The American Cancer Society medical and editorial content team. Stages of Breast Cancer. Available At: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-staging, 2018.

25. Oskouei RJ, Kor NM, Maleki SA. Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. American journal of cancer research 2017; 7(3):610.

26. Moghaddasi H, Hoseini A, Asadi F, Jahanbakhsh M. Application of data

mining in health. Health Information Management 2012; 9(2):297-304.

27. Li X, Jiang W. Method for generating multiple risky barcodes of complex diseases using ant colony algorithm. Theoretical Biology and Medical Modelling 2017; 14(4):1-12.

28. Sweetlin JD, Nehemiah HK, Kannan A. Computer aided diagnosis of pulmonary hamartoma from CT scan images using ant colony optimization based feature selection. Alexandria engineering journal 2018; 57(3):1557-67.

29. Sun Y, Shang J, Liu JX, Li S, Zheng CH. epiACO-a method for identifying epistasis based on ant Colony optimization algorithm. BioData Mining 2017; 10(1): 23.

30. Kanan HR, Faez K, Taheri SM. Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In Industrial Conference on Data Mining 2007; 63-76.

31. Saleema JS, Bhagawathi N, Monica S, Shenoy PD, Venugopal KR, Patnaik LM. Cancer prognosis prediction using balanced stratified sampling. International Journal on Soft Computing, Artificial Intelligence and Applications 2014; 3(1):9-18.

32. Kiani B, Atashi A. A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics 2014; 1(1):26-31.

33. Mahmoodi MS, Mahmoodi SA, Haghighi F, Mahmoodi SM. Determining the stage of breast cancer by data mining algorithms. Iranian Journal of Breast Disease 2014; 7(2):36-44.

34. Ghasem Ahmad L. Using data mining techniques for prediction breast cancer recurrence. Iranian Journal of Breast Disease 2013; 5(4):23-34.