University of Qom

Engineering Management and Soft Computing

JEMSC.QOM
Online ISSN: 2538-2675

Breast Cancer Detection Using Ensemble Classifiers for Accuracy Improvement

Mahboubeh Shamsi[1]✉, Mohadaseh Karimian[2] and Marziyeh Karimian[3]

1. Corresponding author, Assistant Prof. faculty of Electrical and Computer, Qom University of Technology, Qom, Irany. Email: shamsi@qut.ac.ir
2. Msc. of Computer Engineering, Faculty of Electrical and Computer Engineering, Shahab Danesh University, Qom, Iran. Email: m.karimian90@gmail.com
3. Msc. of Computer Engineering, Faculty of Electrical and Computer Engineering, Shahab Danesh University, Qom, Iran. Email: m.karimian64@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 2022 December 25
Received in revised form 2023 February 24
Accepted 2023 March 5
Published online 2023 March 16

Keywords:
Accuracy Improvement,
Data Mining,
Ensemble Classifiers,
Feature Selection,
Sampling.

ABSTRACT

Early diagnosis of breast cancer plays a crucial role in treating the patient. Nowadays, data mining algorithms can provide intelligent methods in the health and treatment system that accurately detect breast cancer. The purpose of this study is breast cancer detection using ensemble classifier based on WBC and WDBC prepared databasesa. Our proposed model in the WBC database (reducing features by cfs+ optimizing samples using Resample+ ensemble classifier using data mining algorithms (kstar + random forest + Naïve Bayes and Bayes network)) has the best detection accuracy (100%), implementation time (0 seconds) and without any errors and on the WDBC database (reducing features by cfs+ optimizing samples using Resample+ ensemble classifier using data mining algorithms (IBK algorithm+ Naïve Bayes, Bayes network and kstar)) has an accuracy of 99/29, the implementation time is 0 seconds, and the mean absolute error is 0/007. The results of this study show that according to the ensemble classifier methods using data mining algorithms on the prepared database, new systems can be designed to help physicians that facilitate treatment processes.

Cite this article: Shamsi, M., Karimian, M. & Karimian, M. (2022). Breast Cancer Detection Using Ensemble Classifiers for Accuracy Improvement. *Enginiiring Management and Soft Computing, 8* (2). 92-109. **DOI**: https://doi.org/10.22091/jemsc.2020.5425.1129

DOI: https://doi.org/10.22091/jemsc.2020.5425.1129

Publisher: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸

مدیریت مهندسی و رایانش نرم

دانشگاه قم

تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت

محبوبه شمسی[۱]✉، محدثه کریمیان[۲] و مرضیه کریمیان[۳]

۱. نویسنده مسئول، استادیار دانشکده برق و کامپیوتر، دانشگاه صنعتی قم، قم، ایران. رایانامه: shamsi@qut.ac.ir

۲. کارشناسی ارشد مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهاب دانش، قم، ایران. رایانامه: m.karimian90@gmail.com

۳. کارشناسی ارشد مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهاب دانش، قم، ایران. رایانامه: m.karimian64@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۱/۱۰/۰۴

تاریخ بازنگری: ۱۴۰۱/۱۲/۰۵

تاریخ پذیرش: ۱۴۰۱/۱۲/۱۴

تاریخ انتشار: ۱۴۰۱/۱۲/۲۵

کلیدواژه‌ها:
انتخاب ویژگی،
بهبود دقت،
داده‌کاوی،
طبقه‌بندهای ترکیبی، نمونه‌گیری.

چکیده

تشخیص زودهنگام سرطان سینه نقش بسیار کلیدی در درمان بیمار ایفا می‌کند. امروزه الگوریتم‌های داده‌کاوی می‌توانند روش‌های هوشمندی در نظام سلامت ارائه دهند که با دقت بالایی سرطان سینه را تشخیص دهند. هدف از انجام این مطالعه، تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی بر روی پایگاه داده‌ی آماده‌سازی شده‌ی WBC و WDBC می‌باشد. مدل پیشنهادی ما در پایگاه داده‌ی WBC (کاهش ویژگی‌ها با CFS + بهینه کردن نمونه‌ها با روش Resample + طبقه بند ترکیبی (kstar+ جنگل تصادفی+ شبکه‌ی بیز و بیزین ساده))، دارای بهترین دقت تشخیص (۱۰۰٪)، زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی می‌باشد و در پایگاه داده‌ی WDBC (کاهش ویژگی‌ها با+ CFS بهینه کردن نمونه‌ها با روش Resample+ طبقه‌بند ترکیبی (الگوریتم IBK+ شبکه‌ی بیز، بیزین ساده و kstar))، دارای دقت ۹۹/۲۹٪، زمان پیاده‌سازی ۰ ثانیه و میانگین خطای مطلق ۰/۰۰۷ می‌باشد. نتایج این مطالعه نشان می‌دهد که با توجه به روش‌های طبقه‌بند ترکیبی بر روی پایگاه‌داده‌ی آماده‌سازی شده می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرآیندهای تشخیصی و درمانی شوند.

استناد: شمسی، محبوبه؛ کریمیان، محدثه و کریمیان، مرضیه. (۱۴۰۱). «تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت». *مدیریت مهندسی و رایانش نرم، دوره ۸* (۲). صص: ۱۰۹-۹۲. https://doi.org/10.22091/jemsc.2020.5425.1129

ناشر: دانشگاه قم

۱) مقدمه

سرطان سینه یکی از شایع‌ترین علت مرگ‌ومیر در زنان محسوب می‌شود و در سال‌های اخیر رشد قابل‌ملاحظه‌ای در تعداد افراد مبتلا به این بیماری گزارش شده است (سیگل، میلر وجمال، ۲۰۱۰). محققان میزان بالای مرگ‌ومیر زنان بر اثر سرطان سینه را ناشی از تشخیص دیرهنگام این بیماری می‌دانند. با تشخیص دقیق و به‌موقع سرطان سینه و پیشرفت‌های به‌دست آمده در درمان، می‌توان با ارائه‌ی درمان مؤثر و راهکارهای ویژه، به افزایش بقا، کاهش مرگ و ارتقا کیفیت زندگی بیماران کمک شایانی نمود (جوشی و مهتا، ۲۰۱۵). یکی از کاربردهای مهم داده‌کاوی مربوط به حوزه‌ی پزشکی و تشخیص بیماری‌ها است. امروزه به دلیل گسترش دانش در حوزه‌ی پزشکی و پیچیدگی تصمیمات مرتبط با تشخیص و درمان، توجه متخصصین به استفاده از ابزارهای هوشمند و تکنولوژی‌های کامپیوتر در امور پزشکی جلب شده است. (گبنا، کریستوفرو یتوند، ۲۰۱۷).

در این مقاله، با بررسی تکنیک‌های متفاوت یادگیری ماشین[۱] و طبقه‌بندهای ترکیبی، سیستمی کارآمد برای تشخیص سرطان سینه طراحی می‌گردد که با دقت بالایی خوش‌خیم یا بدخیم بودن تومورهای سینه را تشخیص می‌دهد. در بخش روش‌شناسی پژوهش، در ابتدا مشخصات پایگاه داده‌ها بیان می‌شوند. سپس مراحل پیش‌پردازش داده‌ها[۲] برای آماده‌سازی پایگاه داده‌ها شرح داده می‌شوند. آنگاه طبقه‌بندهای استفاده شده و فرآیندهای ساخت طبقه‌بند ترکیبی توضیح داده می‌شوند. سپس ارزیابی مدل با استفاده از مقیاس‌های کارایی بیان می‌شود. در بخش یافته‌های پژوهش، مدل پیشنهادی ارائه می‌شود و به مقایسه‌ی نتایج روش خود با نتایج روش‌های دیگران (مقالات مرتبط) پرداخته می‌شود و در بخش آخر نتیجه‌گیری و پیشنهادات بیان می‌شود. در این مقاله ابزار WEKA[۳] برای آزمایشات مورد استفاده قرار گرفته است.

۲) پیشینهٔ پژوهش

تاکنون تحقیقات زیادی در رابطه با تشخیص سرطان سینه با کمک الگوریتم‌های داده‌کاوی انجام شده است. ایلباز تأثیر طبقه‌بندهای متفاوت را در بهبود پیش‌بینی و تشخیص بیماری‌ها برای متخصصان پزشکی ارائه داد. در این مقاله برای تشخیص سریع و به‌موقع سرطان سینه، روش ترکیبی KNN+C4.5+SVM+MLP+BN[4] با کاهش ابعاد داده Rough Set[5] در پایگاه داده‌ی WBC پیشنهاد شد (ایلباز، ۲۰۱۵).

جیاسینگ و وِلُچامی با استفاده از الگوریتم‌های انتخاب ویژگی در داده‌کاوی، گامی حیاتی در افزایش دقت طبقه بندها در تشخیص سریع و به‌موقع سرطان سینه برداشتند. در این مقاله، روش طبقه‌بند Random Forest با استفاده از الگوریتم Modified Bat برای انتخاب ویژگی در پایگاه داده‌ی WDBC پیشنهاد شد (جیاسینگ و وِلُچامی، ۲۰۱۷).

جُشی و مِحتا با بهره‌گیری از تکنیک‌های متنوع و در دسترس یادگیری ماشین، در دو مقاله به تشخیص سرطان سینه پرداختند. در یک مقاله با مقایسه‌ی نتایج طبقه‌بندهای شبکه عصبی همچون MLP، RBF، perceptron و شبکه‌ی عصبی

---

[1]. Machine learningalgorithms
[2]. Data preprocessing
[3]. WEKA (Waikato Environment for Knowledge Analysis). http://www.cs.waikato.ac.nz/~ml/weka/ .
[4]. Bayes Net
[5]. Data dimensionality reduction

عمیق[۶]، الگوریتم MLP با کاهش ویژگی با استفاده از الگوریتم طبقه‌بندی آنالیز گسسته‌ی خطی[۷]، بهترین طبقه‌بند در پایگاه داده‌ی WDBC شناخته شد (جُشی و محِتا، ۲۰۱۸ب) و در مقاله‌ی دیگر با ارائه‌ی روش پیشنهادی الگوریتم K نزدیک‌ترین همسایه[۸] با کاهش ویژگی توسط آنالیز گسسته‌ی خطی در پایگاه داده‌ی WDBC به پیش‌بینی سرطان سینه پرداختند (جُشی و محِتا، ۲۰۱۸الف).

رُهان و همکاران با بررسی یک مدل طبقه‌بند ترکیبی به طبقه‌بندی انواع تومورهای خوش‌خیم و بدخیم از مجموعه داده‌های WBC پرداختند. آن‌ها با ترکیب دو الگوریتم یادگیری ماشین با ناظر ازجمله جنگل تصادفی (RF) و AdaBoost بهبود عملکرد پیش‌بینی مدل را نشان دادند (روهان، سدیک، اسلام و یوسف، ۲۰۱۹).

آراچ و همکاران، با مقایسه‌ی دقت بین طبقه‌بندهای مختلف BN، RF، SMO، J48، IBK، MLP و ترکیب آن‌ها در دو پایگاه داده، نشان دادند که طبقه‌بند ترکیبی موجب بهبود دقت نتایج می‌شود. طبقه‌بند ترکیبی BN+RF+SMO+IBK به همراه انتخاب ویژگی PCA در پایگاه داده‌ی WBC و SMO+BN، SMO+RF، SMO+MLP، SMO+IBK در پایگاه داده‌ی WDBC، بهترین نتایج را نسبت به سایر طبقه‌بندهای ترکیبی داشتند (آراچ و بادن، ۲۰۱۹).

آوینیاش و همکاران، عملکرد طبقه‌بند ترکیبی SVM+SMO را در پایگاه داده‌ی WBC ارزیابی کردند (آوینیاش، بیژوی و جیاراج، ۲۰۲۰).

۳) روش‌شناسی پژوهش

در این پژوهش، ابتدا مشخصات پایگاه داده‌ها توضیح داده می‌شوند. سپس مراحل پیش‌پردازش داده‌ها برای آماده‌سازی پایگاه داده‌ها بیان می‌شوند. آنگاه طبقه‌بندهای استفاده شده و تمام فرآیندهای ساخت طبقه‌بند ترکیبی شرح داده می‌شوند. سپس ارزیابی مدل با استفاده از مقیاس‌های کارایی بیان می‌شود.

۴) مشخصات پایگاه داده‌ها

در این مقاله از دو پایگاه داده‌ی سرطان سینه WBC و WDBC برگرفته از مخزن UCI استفاده شده است.

پایگاه داده‌ی WBC توسط دکتر ویلیام وُلبِرگ در سال ۱۹۹۲، ساخته شد و از دانشگاه ویسکانسین–مدیسون، ایالات‌متحده آمریکا ارائه شده است. پایگاه داده‌ی WBC شامل ۶۹۹ نمونه می‌باشد؛ هر نمونه شامل ۱۱ ویژگی است. ۱۰ ویژگی، اطلاعات خصیصه‌ها را نشان می‌دهند درحالی‌که ۱ ویژگی شامل اطلاعات کلاس است. ویژگی کلاس دو مقدار ۲=خوش‌خیم[۹] و ۴=بدخیم[۱۰] را نمایش می‌دهد. ۱۶ مقدار ازدست‌رفته[۱۱] در این پایگاه داده وجود دارد که با "؟" نشان داده می‌شوند. توزیع کلاس شامل ۴۵۸ نمونه‌ی خوش‌خیم (٪۶۵/۵) و ۲۴۱ نمونه‌ی بدخیم (٪۳۴/۵) است و یک پایگاه داده‌ی ناهماهنگ[۱۲] است. (نیومن، هتیچ، بلیک، مرز و آها، ۱۹۹۸).

---

6. Deep neural network
7. Linear Discriminant Analysis (LDA)
8. K-Nearest Neighbor (KNN)
9. Benign
10. Malignant
11. Missing values
12. Unbalanced

پایگاه داده‌ی WDBC به کمک دکتر ویلیام وُلبِرگ، نیک استریت و اُلوی مانگاساریَن در سال ۱۹۹۵، به وجود آمد و توسط دانشگاه ویسکانسین-مدیسون ایالات‌متحده آمریکا ارائه شد. پایگاه داده‌ی WDBC شامل ۵۶۹ نمونه است؛ هر نمونه دارای ۳۲ ویژگی است. ویژگی اول شامل شماره شناسه‌ی پرونده‌ی بیمار می‌باشد، ویژگی آخر، ویژگی کلاس می‌باشد و ۳۰ ویژگی باقی‌مانده از تصویر دیجیتالی آزمایش آسپیراسیون سوزنی توده‌ی پستان به‌دست آمده است که این ویژگی‌ها، خصوصیات هسته‌ی سلول در تصویر را بیان می‌کنند. هریک از نمونه‌های پایگاه داده‌ی WDBC با یک برچسب خوش‌خیم یا بدخیم مشخص می‌گردند. از ۵۶۹ نمونه‌ی مذکور، ۳۵۷ نمونه دارای برچسب خوش‌خیم و ۲۱۲ نمونه دارای برچسب بدخیم هستند. این پایگاه داده مقادیر ازدست‌رفته ندارد و یک پایگاه داده‌ی ناهماهنگ است (نیومن، هتیچ، بلیک، مرز و آها، ۱۹۹۸).

۵) پیش‌پردازش پایگاه داده

مرحله‌ی آماده‌سازی داده‌ها با عنوان پیش‌پردازش داده‌ها شناخته می‌شود و جهت بهبود کیفیت داده‌های واقعی برای داده‌کاوی لازم است (هان، پی و کامبر، ۲۰۱۱). برای آماده‌سازی پایگاه داده، دو مرحله‌ی پاک‌سازی داده[۱۳] و کاهش ابعاد داده[۱۴] را به ترتیب بر روی پایگاه داده‌ها انجام می‌دهیم که در ادامه این دو مرحله را شرح می‌دهیم.

۶) پاک‌سازی داده

در این مرحله حذف یا جایگزینی مقادیر ازدست‌رفته انجام می‌شود. روش استفاده‌شده در این مقاله، جایگزینی با میانگین (MS)[۱۵] می‌باشد. بدین‌صورت که داده‌های ازدست‌رفته در هر ویژگی را با میانگین مقادیر همان ویژگی جایگزین می‌کند. در اکثر تحقیقاتی که مسئله‌ی داده‌های ازدست‌رفته گزارش و شناسایی شده است، از روش MS برای حل این مشکل استفاده کرده‌اند (پنگ، هارول، لیو و احمان، ۲۰۰۶). بنابراین در این تحقیق نیز، از روش MS برای مقادیر ازدست‌رفته استفاده می‌شود.

۷) کاهش ابعاد داده

برای کاهش ابعاد داده از سه مرحله‌ی کاهش ویژگی‌ها، بهینه کردن نمونه‌ها و تبدیل داده استفاده می‌کنیم که در ادامه این سه مرحله توضیح داده می‌شود.

۷-۱) الف) کاهش ویژگی‌ها (انتخاب ویژگی‌های مناسب)

به دلیل ناهمگنی[۱۶] پایگاه داده‌های پزشکی، افزونگی‌های زیادی در میان داده‌های با ابعاد بالا وجود دارد. هال در مقاله‌ی خود اثبات کرده است که "دقت طبقه‌بندی با استفاده از کاهش مجموعه ویژگی‌ها برابر یا بهتر از دقت طبقه‌بندی با استفاده از همه‌ی ویژگی‌ها است" (هال، ۱۹۹۹). انتخاب یک مجموعه ویژگی که خصوصیات مفید بیشتری را می‌گیرد؛ همیشه نمی‌تواند کارایی طبقه‌بندی را بهبود دهد؛ اما پیچیدگی زمانی و مکانی را کاهش می‌دهد. در این مقاله، روش انتخاب

[13]. Data cleaning
[14]. Data dimensionality reduction
[15]. Mean substitution
[16]. Heterogeneity

ویژگی تحت وکا بر روی مجموعه داده‌ی آموزشی به کار برده می‌شود. کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر همبستگی (CFS)[17] برای مشخص کردن متریک‌های مرتبط با ویژگی کلاس استفاده شده است؛ زیرا CFS مبتنی بر متد همبستگی است و کارایی بهتری دارد. از فواید آن می‌توان، بهبود دقت تشخیص، کاهش در زمان اجرا و ابعاد بازگشتی آن را نام برد (هال، ۹۹۹ ؛ میکالک و کواسنیکا، ۲۰۰۶).

در پایگاه داده‌ی WBC، بعد از اِعمال CFS، بهترین ویژگی‌ها که در تشخیص سرطان سینه مؤثر هستند به‌دست می‌آید که شامل همان ۹ ویژگی و ویژگی کلاس، می‌باشد و تنها، ویژگی شماره شناسه‌ی بیمار حذف می‌شود.

در پایگاه داده‌ی WDBC، بعد از اِعمال CFS، بهترین ویژگی‌ها، شامل ۱ ویژگی Worst Concave Points و ویژگی کلاس (Diagnosis) می‌باشد و سایر ویژگی‌ها که از اهمیت کم‌تری برخوردار هستند، حذف می‌شوند.

۷-۲) ب) بهینه کردن نمونه‌ها

ممکن است مدلی که در پایگاه داده‌های ناهماهنگ ایجاد می‌شود، بهینه نباشد؛ زیرا بیشتر الگوریتم‌های داده‌کاوی، پایگاه داده را هماهنگ در نظر می‌گیرند. بنابراین، با استفاده از تکنیک‌های نمونه‌گیری، داده‌ها را قبل از استفاده از الگوریتم‌ها، هماهنگ می‌کنیم. تکنیک‌های نمونه‌گیری یا هماهنگ کردن داده‌ها به دو دسته تقسیم می‌شوند:

۱) نمونه‌گیری بیش‌ازحد[18]: این گروه از الگوریتم‌ها با افزایش داده‌های کلاس کمینه (بدخیم)، داده‌ها را هماهنگ می‌کند.

۲) نمونه‌گیری کمتر از حد[19]: این گروه با حذف نمونه‌های کلاس اکثریت (خوش‌خیم)، داده‌ها را هماهنگ می‌کند.

روش Resample یک روش نمونه‌گیری تصادفی است که هر دو نمونه‌گیری بیش‌ازحد و کم‌تر از حد را انجام می‌دهد (چاولا، جاپکوویچ وکوتچ، ۲۰۰۴) و مانند SMOTE یک نمونه‌گیری با ناظر[20] است. Resample برای تولید یک نمونه‌ی تصادفی می‌تواند، داده‌های نمونه را با جایگزینی و بدون جایگزینی انجام دهد. در هنگام استفاده از این فیلتر پایگاه داده باید دارای ویژگی کلاس اسمی باشد. ازآنجاکه روش Resample، هر دو نمونه‌گیری بیش از حد و کم‌تر از حد را، با جایگزینی و بدون جایگزینی، انجام می‌دهد، هم‌چنین باعث بهبود دقت طبقه‌بندها می‌شود و قابلیت اطمینان در برابر نویز را بالا می‌برد (راچمن، خودرا و ویدانتورو، ۲۰۱۷)؛ در این مقاله از این روش برای نمونه‌گیری استفاده می‌کنیم.

۷-۳) ج) تبدیل داده

فعالیت‌هایی مانند نرمال‌سازی، گسسته سازی[21] و استانداردسازی داده‌ها، همچنین تبدیل ویژگی اسمی به یک رشته[22]، دودویی[23]، تبدیل ویژگی عددی به اسمی[24]، دودویی[25] و ... در این حوزه جای می‌گیرند.

---

17. Correlation Feature Selection
18. Over-sampling
19. Under-sampling
20. Supervised
21. Discretization
22. Nominal To String
23. Nominal To Binary
24. Numeric To Nominal
25. Numeric To Binary

نرمال‌سازی مهم‌ترین مرحله‌ی تبدیل داده در نظر گرفته می‌شود و مجموعه‌ی داده‌ها توسط نرمال‌سازی تبدیل می‌شوند. نرمال‌سازی مینیمم-ماکزیمم بر فاصله‌ی میان نقاط حداقل و حداکثر تأکید دارد و مقیاس‌سازی را بر اساس تفاوت این دو مقدار انجام می‌دهد، که بازه‌ی مقادیر ویژگی‌ها به [۰ ,۱] مقیاس می‌شوند. مقادیر جدید طبق رابطه‌ی ۱ محاسبه می‌شوند:

رابطه (۱)

$$X^* = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

که در آن $\min(x_i)$ و $\max(x_i)$ به ترتیب مینیمم و ماکزیمم مقدار متغیر x_i است.

ازآنجاکه اکثر مقالات موجود در زمینه‌ی تشخیص سرطان سینه از نرمال‌سازی مینیمم-ماکزیمم (عبدالله،العنزیو الشهران، ۲۰۱۸؛ کراوسیک، ۲۰۱۵) و تبدیل ویژگی عددی به اسمی (آلیامی و همکاران، ۲۰۱۷؛ گوپتا و شالینی، ۲۰۱۸) استفاده کرده‌اند؛ در این مقاله نیز از روش‌های نرمال‌سازی مینیمم-ماکزیمم و تبدیل ویژگی عددی به اسمی استفاده می‌شود.

پایگاه داده‌ی اصلی و پایگاه داده‌ی آماده‌سازی شده به ترتیب در جداول ۱ و ۲ نشان داده شده‌است.

جدول ۱. پایگاه داده‌ی اصلی

پایگاه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	داده‌های از دست رفته	تعداد کلاس‌ها	نوع ویژگی‌ها
WBC	۶۹۹	۱۱	۱۶	۲	(Integer)Numeric
WDBC	۵۶۹	۳۲	–	۲	(Real)Numeric

جدول ۲. پایگاه داده‌ی آماده‌سازی شده

پایگاه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	داده‌های از دست رفته	تعداد کلاس‌ها	نوع ویژگی‌ها
WBC	۶۹۹ (نرمال‌سازی و نمونه گیری شده)	۱۰ (نرمال‌سازی شده)	----	۲	Nominal
WDBC	۵۶۹ (نرمال‌سازی و نمونه‌گیری شده)	۲ (نرمال‌سازی شده)	----	۲	Nominal

۸) الگوریتم‌های طبقه‌بندی

در این مقاله، با توجه به طبقه‌بندهای پرکاربرد در زمینه‌ی پزشکی و تشخیص سرطان سینه، از طبقه‌بندهای جنگل تصادفی، K*، IBK[۲۶]، شبکه‌ی بیز، بیزین ساده[۲۷] و طبقه‌بند ترکیبی استفاده شده است. که در ادامه توضیح مختصری راجع به هر روش (طبقه‌بند) داده می‌شود.

[26]. Instance Based for K-Nearest neighbor
[27]. Naïve Bayes (NB)

۸–۱) جنگل تصادفی

جنگل تصادفی توسط بریمن (نتون و اولسون، ۲۰۰۰) پیشنهاد شد و جزء روش‌های یادگیری طبقه‌بند ترکیبی می‌باشد. در این روش همه‌ی طبقه‌بندها، درخت تصمیم است؛ بنابراین مجموعه‌ی همه‌ی طبقه‌بندها تشکیل جنگل را می‌دهد. هر درخت تصمیم با استفاده از انتخاب تصادفی ویژگی‌ها در هر گره انشعاب را تشخیص می‌دهد. در بین طبقه‌بندی هر درخت رأی می‌دهد و پرطرفدارترین رأی انتخاب می‌شود. جنگل تصادفی به تعداد ویژگی‌های انتخاب انشعاب حساس نیست؛ زیرا جنگل تصادفی از ویژگی‌های کمی در هر انشعاب استفاده می‌کند؛ پس در پایگاه‌داده‌های بزرگ مناسب است و از Bagging و Boosting (طبقه‌بندهای ترکیبی) سریع‌تر است.

۸–۲) الگوریتم IBK

روش K نزدیک‌ترین همسایه (KNN) یک تکنیک دسته‌بندی است که تصمیم‌گیری در مورد این‌که یک نمونه‌ی جدید در کدام کلاس قرار گیرد با بررسی تعدادی (k) از شبیه‌ترین نمونه‌ها یا همسایه‌ها انجام می‌شود. این روش برای یافتن شباهت بین نمونه‌ها نیاز به یک معیار فاصله نظیر فاصله‌ی اقلیدسی یا فاصله‌ی منهتن دارد (هان، پی و کامبر، ۲۰۱۱). IBK یک طبقه‌بندی KNN است که از اندازه‌ی متریک استفاده می‌کند. تابع فاصله‌ی استفاده‌شده، یک پارامتر از روش جستجو است که شامل اقلیدس، چبیشوف، منهتن و مساحت Minkowski می‌باشد.

۸–۳) الگوریتم K* (KStar)

این الگوریتم یک طبقه‌بندی مبتنی بر نمونه است. تفاوت این طبقه‌بندی با مابقی یادگیرنده‌های مبتنی بر نمونه در استفاده از توابع فاصله است و از توابع فاصله بر اساس آنتروپی استفاده می‌کند.

روشی که برای محاسبه‌ی فاصله‌ی بین دو نمونه استفاده می‌شود از نظریه‌ی اطلاعات استفاده می‌کند. فاصله‌ی بین دو نمونه به معنای پیچیدگی انتقال یک نمونه به دیگری است. در ابتدا مجموعه‌ی متناهی از اطلاعاتی که نمونه‌ها را به نمونه‌های دیگر نگاشت می‌دهد تعریف می‌شود. یک برنامه که یک نمونه‌ی (a) را به دیگری (b) انتقال می‌دهد یک دنباله‌ی متناهی از انتقال‌هایی است که از a شروع می‌شود و به b ختم می‌شود (کلاری، تریگس، ۱۹۹۵).

۸–۴) بیزین ساده

این روش یکی از کارآمدترین الگوریتم‌های یادگیری می‌باشد. بر اساس فرض، استقلال قوی، یک ساختار ساده است که ارتباط بین متغیر مستقل و متغیر وابسته را آنالیز می‌کند تا یک احتمال شرطی را برای هر ارتباط استنتاج کند. با استفاده از قضیه‌ی بیز داریم:

رابطه ۲)
$$P(H|X) = P(X|H) \times P(H)/P(X)$$

در معادله‌ی ۲، X یک رکورد داده است. H تعدادی فرض است و به رکوردی از X اشاره می‌کند که به یک کلاس خاص C تعلق دارد. $P(H|X)$ یک احتمال پسین شرطی H روی X است. $P(H)$ نیز یک احتمال پیشین است. احتمال پسین

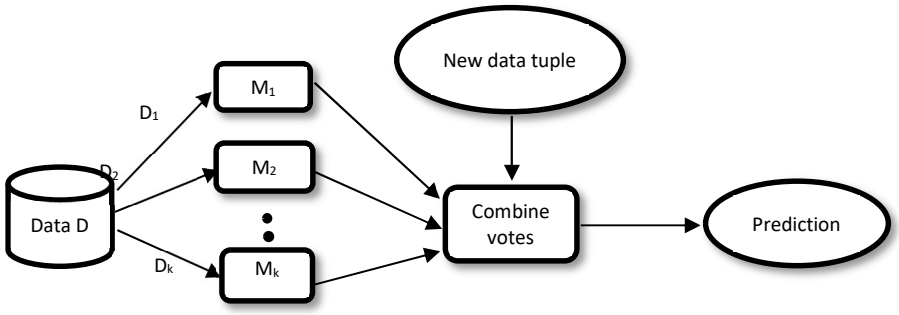
$P(H|X)$، مبتنی بر اطلاعات بیشتری مانند دانش پیش‌زمینه‌ای از احتمال پیشین $P(H)$ که مستقل از X است، می‌باشد. به‌طور مشابه $P(H|X)$ یک احتمال پسین شرطی X بر روی H است (ژانگ و سو، ۲۰۰۸).

۸-۵) شبکه‌ی بیز

شبکه‌ی بیز (جنسنس، ۱۹۹۶) یک مدل گرافیکی احتمالی است که مجموعه‌ای از متغیرهای تصادفی و وابستگی شرطی را از طریق گراف غیر چرخشی جهت‌دار[۲۸] نشان می‌دهد. به عنوان مثال، یک شبکه‌ی بیز می‌تواند رابطه‌ی احتمالی بین بیماری و علائم آن را نشان دهد. با گرفتن علائم بیماری، شبکه می‌تواند احتمال وجود انواع بیماری‌ها را بیان کند.

۸-۶) طبقه‌بند ترکیبی

Bagging، Boosting و جنگل‌تصادفی از طبقه‌بندهای ترکیبی هستند. در روش طبقه‌بند ترکیبی، یک دسته kتایی از مدل‌های یادگیری را باهم ترکیب می‌کند و یک مدل بهبودیافته‌ی مرکب از آن‌ها به دست می‌آورد. از یک پایگاه‌داده، k مجموعه‌ی آموزشی ایجاد می‌شود که مدل دسته‌بند M_i را می‌سازد. سپس با گرفتن یک داده‌ی جدید، هر طبقه‌بند رأی خود را می‌دهد و طبقه‌بند ترکیبی رأی‌ها را باهم ترکیب می‌کند و برچسب کلاس پیشگویی شده را برمی‌گرداند (وزنیاک، گرانا و کورچادو، ۲۰۱۴). دقت طبقه‌بند ترکیبی بیشتر از یک طبقه‌بند است؛ زیرا برفرض اگر طبقه‌بند ترکیبی اکثریت آرا[۲۹] را برگرداند، با دادن یک داده‌ی چندتایی، همه‌ی طبقه‌بندها برچسب کلاس پیشگویی را برمی‌گردانند و خروجی، برچسب اکثریت می‌باشد. (کیتلر، هاتف، دوین و ماتاس، ۱۹۹۸). در شکل ۱ روش طبقه‌بند ترکیبی برای ساخت مدل نشان داده شده است (وست دی، منگیامیلی، رامپال و وست وی، ۲۰۰۵).



شکل ۱. روش طبقه‌بند ترکیبی برای ساخت مدل طبقه‌بندی

۹) ارزیابی مدل

در این پژوهش، با توجه به مقالات بررسی‌شده در جهت تشخیص سرطان سینه، برای ارزیابی مدل از مقیاس‌های کارایی که مهم‌ترین آن‌ها دقت، زمان پیاده‌سازی و میانگین خطای مطلق می‌باشد، استفاده شده است. در ادامه توضیح مختصری راجع به هرکدام داده می‌شود.

[28].Directed acyclic graph
[29].Majority Voting

۹-۱) دقت

دقت با نام نرخ طبقه‌بندی درست نیز شناخته شده است و نسبت تعداد سلول‌های سرطانی که درست پیش‌بینی شده‌اند به همه سلول‌ها است. این مقیاس به صورت رابطه‌ی ۳ محاسبه می‌شود (کورو و لیو، ۲۰۰۵).

رابطه ۳)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

۹-۲) زمان پیاده‌سازی

هرچه یک الگوریتم در مدت‌زمان کمتری پیاده‌سازی شود، کارایی و اثربخشی بهتری دارد. زمان پیاده‌سازی به اندازه‌ی دقت طبقه بندها مهم است (ایلباز، ۲۰۱۵).

۹-۳) میانگین خطای مطلق

میانگین خطای مطلق (MAE)، میانگین نمونه‌ی آزمایشی است که تفاوت‌های مطلق میان مشاهدات پیش‌بینی‌شده و مشاهدات واقعی را نشان می‌دهد که در آن، همه‌ی تفاوت‌های منحصربه‌فرد، وزن یکسان دارند. در این پژوهش، از روش MAE برای اندازه‌گیری خطا استفاده شده است (چوراسیا، پال، ۲۰۱۷).

۹-۴) سایر معیارهای ارزیابی کارایی مدل

روش‌های متعددی جهت مقایسه‌ی طبقه‌بندی وجود دارند. کارایی پیشگویی مدل‌ها برای دو کلاس بدخیم (عود سرطان) و خوش‌خیم (عدم عود سرطان) معمولاً با استفاده از ماتریس اغتشاش که در جدول ۳ نشان داده شده است بیان می‌گردد.

جدول ۳. ماتریس اغتشاش

	NO (Prediction)	Yes (Prediction)
NO (Actual)	True Negative (TN)	False Positive (FP)
Yes (Actual)	False Negative (FN)	True Positive (TP)

ستون‌ها، نتایج پیشگویی شده و ردیف‌ها، نتایج واقعی برچسب کلاس‌ها را نشان می‌دهند. سلول‌های بدخیم با برچسب YES و سلول‌های خوش‌خیم با برچسب NO نمایش داده شده‌اند. بنابراین عناصر قطری (TN,TP) در جدول ۳ پیش‌بینی درست و بقیه (FN,FP) پیش‌بینی اشتباه را بیان می‌کنند.

۹-۵) درستی

درستی[۳۰]، نسبت تعداد سلول‌های بدخیم را که درست پیش‌بینی شده است به همه سلول‌هایی که بدخیم پیش‌بینی شده است، نشان می‌دهد و به‌صورت رابطه ۴ محاسبه می‌شود.

رابطه ۴)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

هرچه درستی افزایش یابد تلاش کمتری در آزمون و بررسی هدر می‌رود (کورو و لیو، ۲۰۰۵).

[30]. Precision

۹-۶) نرخ کشف خطا

نرخ کشف خطا[۳۱] مقیاسی است که نسبت تعداد سلول‌های بدخیم را که درست پیش‌بینی شده‌اند به همه سلول‌هایی که درواقع بدخیم هستند نشان می‌دهد و به‌صورت رابطه‌ی ۵ محاسبه می‌شود. هرچه نرخ کشف خطا افزایش یابد سلول‌های بدخیم بیشتری کشف می‌شوند (کورو و لیو، ۲۰۰۵).

$$\text{Recall} = \text{PD} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ رابطه ۵)

۹-۷) مقیاس F

مقیاس F، میانگین هارمونیک[۳۲] درستی و نرخ کشف خطا است که تصویر کلی خوبی از کارایی پیش‌بینی را ارائه می‌دهد (ویتن و فرانک، ۲۰۰۵) و به‌صورت رابطه‌ی ۶ محاسبه می‌شود.

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$ رابطه ۶)

۹-۸) ناحیه زیر منحنی ROC

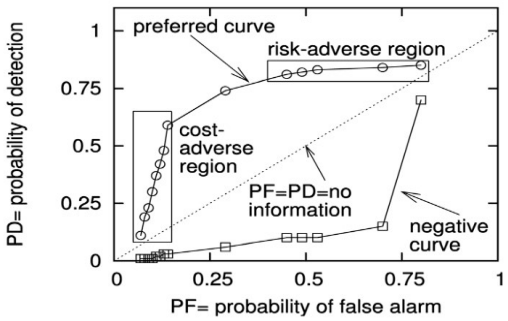
مشخصات عملیاتی گیرنده[۳۳](ROC) می‌تواند برای ارزیابی کارایی مدل‌های تشخیص و پیش‌بینی سرطان استفاده شود. منحنی ROC، از دو پارامتر نرخ کشف خطا (PD) در محور Y و احتمال هشدار اشتباه[۳۴] (PF) در محور X استفاده می‌کند و کارایی مدل را پیش‌بینی می‌کند. پارامترهای PD (Recall) و PF نیز از روی ماتریس اغتشاش محاسبه می‌شوند که به ترتیب در رابطه‌های ۵ و ۷ آورده شده‌اند.

$$\text{PF} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$ رابطه ۷)

منحنی ROC باید از نقاط (۰,۰) و (۱,۱) عبور کند (منزیز، گرین والد و فرانک، ۲۰۰۷). مهم‌ترین نواحی منحنی ROC در شکل ۲ نشان داده شده است. نقطه‌ی ایده‌آل منحنی در (۰,۱) است و هیچ خطایی در این نقطه وجود ندارد. خط گذرنده از نقاط (۰,۰) و (۱,۱) هیچ اطلاعاتی در اختیار قرار نمی‌دهد و به همین دلیل مقدار ناحیه‌ی زیر منحنی ROC (AUC[35])، باید بیشتر از ۰/۵ باشد. اگر نتیجه روی منحنی منفی[۳۶] قرار بگیرد نتیجه‌ی خوبی نیست و به معنی غیرقابل‌قبول بودن کارایی طبقه‌بندی است. منحنی مناسب[۳۷] در شکل ۲ نشان داده شده است. در ناحیه‌ی مخالف هزینه[۳۸] احتمال کشف خطا کم‌تر است و در زمانی که بودجه‌ی اعتباریابی و وارسی کم است، مناسب است. ناحیه‌ی مخالف خطر[۳۹] دارای هزینه‌ی زیادی است زیرا بااینکه احتمال کشف خطا زیاد است ولی احتمال هشدار خطا نیز زیاد است. برای سیستم‌های مأموریت بحرانی، ناحیه مخالف خطر و برای برنامه‌های تجارت ناحیه‌ی مخالف هزینه انتخاب می‌شود.

---

31. Recall
32. harmonic
33. Receiver Operating Characteristics
34. Probability Of False Alarm
35. Area Under Curve
36. Negative curve
37. Preferred curve
38. Cost-adverse
39. Risk-adverse



شکل۲. نواحی مختلف منحنی ROC (منزیز، گرین والد و فرانک، ۲۰۰۷).

۱۰) یافته‌های پژوهش

در این مقاله، برای مقایسه‌ی دقیق الگوریتم‌های طبقه‌بندی، از ۲ سناریو بر روی ۲ پایگاه داده‌ی سرطان سینه (برگرفته از UCI) استفاده شده است. ۱۵روش ساخت مدل (۵ الگوریتم طبقه‌بندی در سناریوی اول (S1) و ۱۰ الگوریتم طبقه‌بند ترکیبی در سناریوی دوم (S2)) در هر پایگاه داده بررسی شده است و ارزیابی با روش 10-fold cross validation (۱۰ بار ارزیابی) انجام شده است . اِعمال روش 10-fold cross validation بر روی ۲ پایگاه‌داده WBC وWDBC در جدول ۴ بررسی می‌شود.

جدول ٤. اعمال روش 10-fold cross validation برای هر پایگاه‌داده

پایگاه داده	تعداد نمونه‌ها	10-fold cross validation	
		مجموعه داده آموزشی	مجموعه داده آزمایشی
WBC	٦٩٩	٦٢٩	٧٠
WDBC	٥٦٩	٥١٢	٥٧

نتایج نیز بر اساس مقیاس‌های کارایی ازجمله دقت، زمان پیاده‌سازی و میانگین خطای مطلق بیان گردیده است. ۲ سناریو عبارتند از :

۱. سناریو اول (S1): انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه بندها به‌صورت جداگانه

۲. سناریو دوم (S2): انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه بندهای ترکیبی

در S1، پیاده‌سازی الگوریتم‌های طبقه‌بندی به‌صورت جداگانه بر روی پایگاه داده‌ها می‌باشد و در S2، طبقه‌بندهای ترکیبی بر روی پایگاه داده‌ها اعمال می‌شوند. در این پژوهش، روش ترکیب طبقه‌بندها بر اساس اکثریت آرا می‌باشد و ایده‌ی ترکیب طبقه‌بندها در هر پایگاه داده به این صورت می‌باشد که ابتدا بهترین طبقه‌بند انتخاب می‌شود و سپس با سایر طبقه بندها ترکیب می‌شود. در مرحله‌ی بعد بهترین ترکیب دوتایی طبقه‌بندها انتخاب می‌شود، آن‌گاه با سایر طبقه‌بندها ترکیب می‌شود. در مرحله‌ی بعد بهترین ترکیب سه‌تایی طبقه بندها انتخاب می‌شود و با سایر طبقه بندها ترکیب می‌شود. این کار، تا زمانی‌که همه‌ی طبقه‌بندها با هم ترکیب شوند، ادامه داده می‌شود و در مرحله‌ی آخر فقط یک طبقه‌بند ترکیبی وجود دارد. آن‌گاه با مقایسه‌ی همه‌ی طبقه‌بندهای ترکیبی بهترین نتایج انتخاب می‌شود.

در ادامه جداول و نمودارهای مربوط به هر پایگاه داده در هر سناریو ارائه می‌شود و توضیحات هر یک داده می‌شود.

۱۰-۱) سناریو اول: انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندها به‌صورت جداگانه

در پایگاه داده‌ی WBC، ۲ طبقه‌بند K* و RF دقت یکسانی دارند (٪۹۹/۸۵). همچنین زمان پیاده‌سازی آن‌ها نیز یکسان (۰ ثانیه) است. اما K* بهتر است، زیرا میانگین خطای مطلق آن کمتر است (به ترتیب ۰/۰۰۳ و ۰/۰۱).

در پایگاه داده‌ی WDBC، ۲ طبقه‌بند IBK و BN دارای دقت ٪۹۹/۲۹ و زمان پیاده‌سازی (۰ ثانیه) یکسانی دارند. اما میانگین خطای مطلق آن‌ها متفاوت است (به ترتیب ۰/۰۰۸ و ۰/۰۸). بنابراین IBK بهتر از BN است، زیرا میانگین خطای مطلق آن کمتر است.

مقایسه‌ی دقت طبقه‌بندها در ۲ پایگاه داده در جدول ۵ آورده شده است.

جدول ۵. مقایسه‌ی دقت طبقه‌بندها در ۲ پایگاه داده

طبقه‌بندها	RF	IBK	K*	NB	BN
WBC	۹۹/۸۵	۹۹/۷۱	۹۹/۸۵	۹۷/۱۳	۹۶/۹۹
WDBC	۹۹/۱۲	۹۹/۲۹	۷۹/۹۶	۹۵/۷۸	۹۹/۲۹

۱۰-۲) سناریو دوم: انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندهای ترکیبی

در پایگاه داده‌ی WBC، همان‌طور که در جدول ۶ نشان داده شده است، سه طبقه‌بند ترکیبی K*+RF+NB، K*+RF+BN و K*+RF+NB+IBK بالاترین دقت را دارند (٪۱۰۰). اما، زمان پیاده‌سازی طبقه‌بند ترکیبی K*+RF+NB+IBK، ۰/۰۲ ثانیه و دو طبقه‌بند ترکیبی دیگر ۰ ثانیه می‌باشد. بنابراین، دو طبقه‌بند ترکیبی K*+RF+NB و K*+RF+BN بهترین می‌باشند، زیرا دارای بهترین زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی (میانگین خطای مطلق برابر ۰ است) می‌باشند. در پایگاه داده‌ی WBC، طبقه‌بند ترکیبی دقت بیش‌تری نسبت به طبقه‌بند تنها دارد (به ترتیب ٪۱۰۰ و ٪۹۹/۸۵). بنابراین در WBC، طبقه‌بند ترکیبی بهترین می‌باشد.

در پایگاه داده‌ی WDBC، همان‌طور که در جدول ۷ نشان داده شده است، چهار طبقه‌بند ترکیبی IBK+BN، IBK+BN+NB، IBK+BN+K*، IBK+BN+RF بالاترین دقت را دارند (٪۹۹/۲۹). در حالی‌که زمان پیاده‌سازی آن‌ها متفاوت است. زمان پیاده‌سازی طبقه‌بند ترکیبی IBK+BN+RF، ۰/۰۲ ثانیه و سه طبقه بند ترکیبی دیگر دارای بهترین زمان پیاده‌سازی می‌باشند (۰ ثانیه). همچنین میانگین خطای مطلق این سه طبقه‌بند ترکیبی باهم برابر است (۰/۰۰۷). بنابراین سه طبقه‌بند ترکیبی IBK+BN، IBK+BN+NB و IBK+BN+K* بهترین می‌باشند. در سناریوی اول، طبقه‌بند ترکیبی در مقایسه با طبقه‌بند تنها باوجود یکسان بودن میزان دقت و زمان پیاده‌سازی، بهتر است زیرا میانگین خطای مطلق آن کمتر است (به ترتیب ۰/۰۰۷ و ۰/۰۰۸). بنابراین در این پایگاه داده (WDBC) نیز طبقه‌بند ترکیبی، بهترین است.

طبق نتایج به دست آمده در ۲ پایگاه داده، طبقه‌بند ترکیبی در مقایسه با طبقه‌بند تنها بهتر است و کارایی بالاتری دارد. بنابراین مدل پیشنهادی در این پژوهش، سناریوی دوم می‌باشد.

در جداول ۶ و ۷ مقایسه‌ی دقت طبقه‌بندهای ترکیبی برای هر پایگاه داده ارائه شده است.

جدول ٦. مقایسه‌ی دقت در پایگاه داده WBC

مرحله				
مرحله دوم	K*+RF	K*+NB	K*+BN	K*+IBK
	۹۹/۸۵	۹۸/۷۱	۹۸/۵٦	۹۹/۸۵
مرحله سوم	K*+RF+NB	K*+RF+ BN	K*+RF+IBK	
	۱۰۰	۱۰۰	۹۹/۸۵	
مرحله چهارم	K*+RF+NB +BN	K*+RF+ NB+IBK		
	۹۸/۷۱	۱۰۰		
مرحله پنجم	K*+RF+NB +IBK+BN			
	۹۹/۸۵			

جدول ۷. مقایسه‌ی دقت در پایگاه داده WDBC

مرحله				
مرحله دوم	IBK+BN	IBK+N B	IBK+K*	IBK+RF
	۹۹/۲۹	۹۷/۱۸	۹۰/۸٦	۹۹/۱۲
مرحله سوم	IBK+BN +NB	IBK+B N+K*	IBK+BN+RF	
	۹۹/۲۹	۹۹/۲۹	۹۹/۲۹	
مرحله چهارم	IBK+BN +NB+K*	IBK+BN+N B+RF		
	۹۷/۱۸	۹۹/۱۲		
مرحله پنجم	IBK+BN +NB+RF + K*			
	۹۹/۱۲			

۱۰–۳) مقایسه‌ی نمودارهای مقیاس‌های کارایی

در این قسمت نمودارهای مربوط به هر پایگاه داده در هر سناریو، توضیح داده می‌شود. در نمودارها، S1، بهترین طبقه‌بند در سناریوی اول (طبقه بند تنها) و S2 بهترین طبقه‌بند در سناریوی دوم (طبقه‌بند ترکیبی) می‌باشد.

در شکل ۳، دقت بهترین طبقه‌بند، در هر ۲ سناریو برای هر ۲ پایگاه داده نشان داده شده است. در WBC بهترین طبقه‌بند در سناریوی دوم (مدل پیشنهادی (طبقه بند ترکیبی))، بالاترین دقت را نسبت به بهترین طبقه‌بند در سناریوی اول (طبقه‌بند تنها) دارد و در WDBC دقت بهترین طبقه‌بند سناریوی دوم با بهترین طبقه‌بند سناریوی اول برابر است.

در شکل ۴، میانگین خطای مطلق بهترین طبقه‌بند در هر ۲ سناریو برای هر ۲ پایگاه داده نشان داده شده است که واضح است میانگین خطای مطلق بهترین طبقه‌بند سناریوی دوم از بهترین طبقه‌بند سناریوی اول کمتر است.

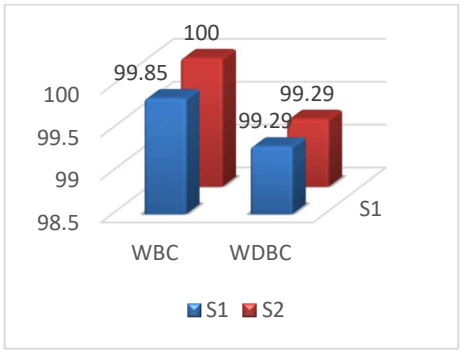
در شکل‌های ۵ و ۶، سایر معیارهای ارزیابی کارایی مدل، به ترتیب، برای دو پایگاه داده‌ی WBC و WDBC نشان داده شده است که در WBC بهترین طبقه‌بند سناریوی دوم بهتر از بهترین طبقه‌بند سناریوی اول است و کارایی بالاتری نسبت به آن دارد و در WDBC بهترین طبقه‌بند سناریوی اول و دوم باهم برابر هستند.

ازآنجاکه زمان پیاده‌سازی بهترین طبقه‌بند برای هر ۲ سناریو در هر ۲ پایگاه داده‌ی WBC و WDBC ۰ ثانیه است از رسم کردن این نمودار صرف‌نظر می‌کنیم.

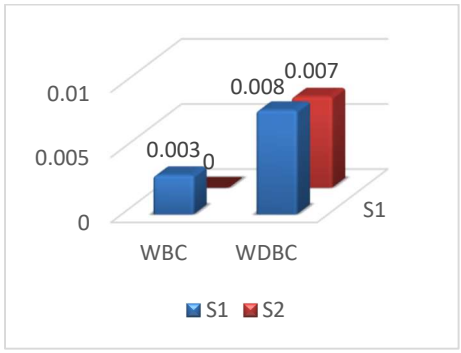
با توجه به نتایج به‌دست‌آمده از نمودارها، سناریوی دوم بهترین سناریو است و بالاترین کارایی را دارد.

۱۰-۴) مقایسه‌ی نتایج

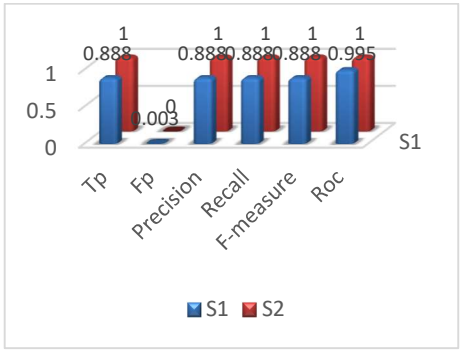
نتایج به‌دست‌آمده در این پژوهش، با سایر نتایج مقالات مرتبط بررسی شده است تا بهبود دقت و افزایش کارایی مدل پیشنهادی نسبت به سایر مقالات مشخص شود. مقایسه‌ی این نتایج در جدول ۸ ارائه شده است.



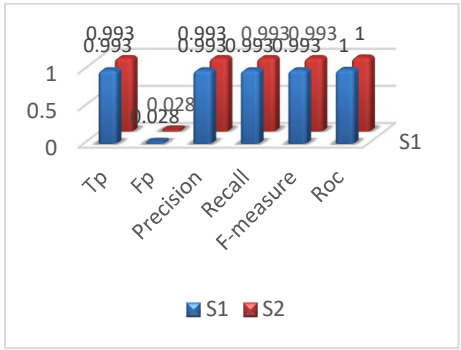
شکل ۳. ارزیابی دقت بهترین طبقه‌بند در هر سناریو در ۲ پایگاه داده



شکل ٤. ارزیابی میانگین خطای مطلق بهترین طبقه‌بند در هر سناریو در ۲ پایگاه داده



شکل ۵. ارزیابی سایر مقیاس‌های اندازه‌گیری کارایی مدل در پایگاه داده‌ی WBC



شکل ٦. ارزیابی سایر مقیاس‌های اندازه‌گیری کارایی مدل در پایگاه داده‌ی WDBC

جدول ۸. مقایسه نتایج خود با مقالات مرتبط

منابع	روش	دقت	پایگاه داده
(سلامه، عبدالحلیم و زید، ۲۰۱۲)	SMO+J48+NB+IBK	۹۷/۲۸	WBC
(ایلباز، ۲۰۱۵)	RS-KNN+C4.5+SVM+MLP+BN	۹۹/٤۱	WBC
(چوراسیا و پال، ۲۰۱۷ الف)	SVM+RBF	۹٦/۸٤	WBC
(الیس‌کاویک و سوباسی، ۲۰۱۷)	GA-Rotation Forest+SVM	۹۹/٤۸	WBC
(آلیامی و همکاران، ۲۰۱۷)	CC-SVM+MLP	۹٦/۷۱	WBC
(آدگک، چن، بانیسی و باریکزای، ۲۰۱۷)	AB+RF ،MBAB+RF	۹۷	WBC
(کومار، نیکیل و سومانگالی، ۲۰۱۷)	SVM+NB+J48	۹۷/۱۳	WBC
(هوانگ، چن، لین، کی و تسای، ۲۰۱۷)	GA-SVM+RBF+Boosting	۹۸/۲۸	WBC

منابع	روش	دقت	پایگاه داده
(عبدالله، العنزی و الشرحان، ۲۰۱۸)	SVM+FCM+GK	۹۹/۰٤	WBC
(آنی، خوزه، ویلسون و دیپا، ۲۰۱۸)	PCA-Rotation Forest+MLP	۹۸/۰۵	WBC
(روهان، سدیک، اسلام و یوسف، ۲۰۱۹)	AdaBoost+RF	۹۸/۵۷	WBC
(آراچ و بادن، ۲۰۱۹)	PCA-BN+RF+SMO+IBK	۹۷/۹۹	WBC
(آوینیاش، بیژوی و جیاراج، ۲۰۲۰).	SVM+SMO	۹٤/٦	WBC
(سلامه، عبدالحلیم و زید، ۲۰۱۲)	SMO ،SMO+IBK ،SMO+MLP	۹۷/۷۱	WDBC
(حازرا، ماندال و گوپتا، ۲۰۱٦)	PCA-NB+SVM	۹۵/۹۱	WDBC
(ماندال، ۲۰۱۷)	PCC-Logistic Regression	۹۸	WDBC
(جیاسینگ و وِلُچامی، ۲۰۱۷)	Modified Bat-Random Forest	۹٦/۸۵	WDBC
(نیلاشی، بن ابراهیم، احمدی و شاه‌مرادی، ۲۰۱۷)	PCA-CART+EM	۹۲/۸	WDBC
(خوریوال و میشرا، ۲۰۱۸)	Chi-Squared-MLP+Logistic	۹۸/۵۰	WDBC
(جُشی و مِحتا، ۲۰۱۸ ب)	LDA-MLP	۹۷/۰٦	WDBC
(جُشی و مِحتا، ۲۰۱۸ الف)	LDA-KNN	۹۷/۰٦	WDBC
(آراچ و بادن، ۲۰۱۹)	SMO+RF،SMO+BN،SMO+IBK،SMO+MLP	۹۷/۷۱	WDBC
روش پیشنهادی	Resample-CFS-K*+RF+BN Resample-CFS-K*+RF+NB	۱۰۰	WBC
روش پیشنهادی	Resample–CFS–IBK+BN Resample–CFS–IBK+BN+NB Resample–CFS–IBK+BN+K*	۹۹/۲۹	WDBC

۱۱) نتیجه‌گیری و پیشنهادها

در این مطالعه، ما با استفاده از روش طبقه‌بند ترکیبی (تجمعی)، کاهش ویژگی (با استفاده از روش انتخاب ویژگی مبتنی بر همبستگی (CFS))، نمونه‌گیری (با استفاده از روش نمونه‌گیری با ناظر Resample) و نرمال‌سازی داده‌ها، دقت شناسایی سیستم‌های سرطان سینه را افزایش دادیم. درواقع با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و باصرفه‌تری در نظام سلامت و درمان ارائه داد که با دقت بالایی قادر به تشخیص سرطان سینه باشند. استفاده از این سیستم‌ها می‌تواند موجب کاهش خطاهای احتمالی شود و دقت تشخیص سرطان سینه را بهبود بخشد. مدل پیشنهادی در پایگاه داده‌ی WBC، دارای بهترین دقت تشخیص (۱۰۰٪)، زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی می‌باشد و در پایگاه داده‌ی WDBC، دارای دقت ٪۹۹/۲۹ زمان پیاده‌سازی ۰ ثانیه و میانگین خطای مطلق ۰/۰۰۷ می‌باشد، که در هر دو پایگاه داده نسبت به سایر روش‌ها دارای بهترین نتایج است. همچنین نتایج مطالعات محققان دیگر بر روی پایگاه داده‌های WBC و WDBC در مقایسه با روش پیشنهادی، حاکی از برتری روش پیشنهادی این مطالعه است. نتایج این مطالعه نشان می‌دهد که با توجه به روش‌های طبقه‌بند ترکیبی با استفاده از الگوریتم‌های داده‌کاوی بر روی پایگاه‌داده‌ی آماده‌سازی شده می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرآیندهای تشخیصی و درمانی شوند.

منابع

Abdullah, M., Al-Anzi, F., & Al-Sharhan, S. (2018). Hybrid Multistage Fuzzy Clustering System for Medical Data Classification. Computing Sciences and Engineering (ICCSE), 2018 International Conference On, 1–6. IEEE. DOI: https://doi.org/10.1109/ICCSE1.2018.8374213

Adegoke, V. F., Chen, D., Banissi, E., & Barikzai, S. (2017). Prediction of breast cancer survivability using ensemble algorithms. Smart Systems and Technologies (SST), 2017 International Conference On, 223–231. IEEE. DOI: https://doi.org/10.1109/SST.2017.8188699

Alickovic, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4), 753–763. DOI: https://doi.org/10.1007/s00521-015-2103-9

Alyami, R., Alhajjaj, J., Alnajrani, B., Elaalami, I., Alqahtani, A., Aldhafferi, N., … Olatunji, S. O. (2017). Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines. Informatics, Health & Technology (ICIHT), International Conference On, 1–7. IEEE. DOI: https://doi.org/10.1109/ICIHT.2017.7899011

Ani, R., Jose, J., Wilson, M., & Deepa, O. S. (2018). Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems. In Progress in Advanced Computing and Intelligent Engineering (pp. 137–146). Springer. DOI: https://doi.org/10.1016/j.jisa.2023.103541

Arach, S., & Bouden, H. (2019). Performance Analysis on Three Breast Cancer Datasets using Ensemble Classifiers Techniques. Computer Science, 14(4), 935–952. DOI: https://doi.org/10.1016/j.eswa.2023.122641

Avinash, K., Bijoy, M. B., & Jayaraj, P. B. (2020). Early Detection of Breast Cancer Using Support Vector Machine With Sequential Minimal Optimization. In Advanced Computing and Intelligent Engineering (pp. 13–24). Springer DOI: https://doi.org/10.1007/978-981-15-1081-6_2

Chaurasia, V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10–22.

Chaurasia, V., & Pal, S. (2017b). Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease.

Chawla, N. V, Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 6(1), 1–6. DOI: https://doi.org/10.1145/1007730.1007733

Cleary, J. G., & Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. ICML, 108–114. DOI: https://doi.org/10.1016/B978-1-55860-377-6.50022-0

El-Baz, A. H. (2015). Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. Neural Computing and Applications, 26(2), 437–446 DOI: https://doi.org/10.1007/s00521-014-1731-9

Fenton, N. E., & Ohlsson, N. (2000). Quantitative analysis of faults and failures in a complex software system. Software Engineering, IEEE Transactions On, 26(8), 797–814. DOI: https://doi.org/10.1109/32.879815

Gbenga, D. E., Christopher, N., & Yetunde, D. C. (2017). Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. Nova, 6(1), 1–8 DOI: https://doi.org/10.20286/nova-jeas-060105

Gupta, P., & Shalini, L. (2018). Analysis of Machine Learning Techniques for Breast Cancer Prediction. International Journal Of Engineering And Computer Science, 7(05), 23891–23895. DOI: https://doi.org/10.31033/ijemr.11.1.12

Hall, M. A. (1999). Correlation-based feature selection for machine learning. DOI: https://doi.org/10.4236/ojbm.2021.92030

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier DOI: https://doi.org/10.4236/als.2019.74012

Hazra, A., Mandal, S. K., & Gupta, A. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, 145(2). DOI: https://doi.org/10.5120/ijca2016910595

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2017). SVM and SVM ensembles in breast cancer prediction. PloS One, 12(1), e0161501 DOI: https://doi.org/10.1371/journal.pone.0161501

Jensen, F. V. (1996). An introduction to Bayesian networks (Vol. 210). UCL press London. DOI: https://doi.org/10.1016/j.ifacol.2018.07.024

Joshi, A., & Mehta, A. (2018a). ANALYSIS OF K-NEAREST NEIGHBOR TECHNIQUE FOR BREAST CANCER DISEASE CLASSIFICATION. Machine Learning, 98, 13. DOI: https://doi.org/10.47611/jsrhs.v12i4.5577

Joshi, A., & Mehta, A. (2018b). BREAST CANCER DATA CLASSIFICATION USING NEURAL NETWORK AND DEEP NEURAL NETWORK TECHNIQUES. Int J Recent Sci Res, 9(4), 25788–25792. DOI: https://doi.org/10.1504/IJISDC.2020.10037864

Khuriwal, N., & Mishra, N. (2018). Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. 2018 IEEMA Engineer Infinite Conference (ETechNxT), 1–5. IEEE. DOI: https://doi.org/10.1109/ETECHNXT.2018.8385355

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226–239. DOI: https://doi.org/10.1109/34.667881

Koru, A. G., & Liu, H. (2005). Building effective defect-prediction models in practice. Software, IEEE, 22(6), 23–29. DOI: https://doi.org/10.1109/MS.2005.149

Krawczyk, B. (2015). One-class classifier ensemble pruning and weighting with firefly algorithm. Neurocomputing, 150, 490–500. DOI: https://doi.org/10.1016/j.neucom.2014.07.068

Kumar, U. K., Nikhil, M. B. S., & Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017 IEEE International Conference On, 108–114. IEEE DOI: https://doi.org/10.1109/ICSTM.2017.8089135

Mandal, S. K. (2017). 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, 6(2) DOI: https://doi.org/10.1088/1742-6596/1577/1/012051

Menzies, T., Greenwald, J., & Frank, A. (2007). Data mining static code attributes to learn defect predictors. Software Engineering, IEEE Transactions On, 33(1), 2–13. DOI: https://doi.org/10.1109/TSE.2007.256941

Michalak, K., & Kwasnicka, H. (2006). Correlation-based feature selection strategy in neural classification. Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference On, 1, 741–746. IEEE. DOI: https://doi.org/10.1109/ISDA.2006.128

Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J., & Aha, D. W. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA. 1998 of Conference, Http://Archive. Ics. Uci. Edu/Ml/Datasets. Html. DOI: https://doi.org/10.4236/me.2013.410068

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212–223. DOI: https://doi.org/10.1016/j.compchemeng.2017.06.011

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. Real Data Analysis, 3178 DOI: https://doi.org/10.1007/s42979-022-01249-z

Rachman, G. H., Khodra, M. L., & Widyantoro, D. H. (2017). Rhetorical Sentence Categorization for Scientific Paper Using Word2Vec Semantic 36Representation. Journal of Physics: Conference Series, 801(1), 12070. IOP Publishing DOI: https://doi.org/10.1088/1742-6596/801/1/012070

Rohan, T. I., Siddik, A. B., Islam, M., & Yusuf, M. S. U. (2019). A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost. 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 1–4. IEEE. DOI: https://doi.org/10.1109/IC4ME247184.2019.9036697

Salama, G. I., Abdelhalim, M., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569), 2

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. CA: A Cancer Journal for Clinicians, 67(1), 7–30. DOI: https://doi.org/10.3322/caac.21387

Teh, Y.-C., Tan, G.-H., Taib, N. A., Rahmat, K., Westerhout, C. J., Fadzli, F., … Yip, C.-H. (2015). Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. BMC Cancer, 15(1), 405 DOI: https://doi.org/10.1186/s12885-015-1419-2

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. European Journal of Operational Research, 162(2), 532–551 DOI: https://doi.org/10.1016/j.ejor.2003.10.013

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann ISBN:978-0-12-374856-0

Wozniak, M., Grana, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. Information Fusion, 16, 3–17. DOI: https://doi.org/10.1016/j.inffus.2013.04.006

Zhang, H., & Su, J. (2008). Naive Bayes for optimal ranking. Journal of Experimental & Theoretical Artificial Intelligence, 20(2), 79–93 DOI: https://doi.org/10.1080/09528130701476391