JOURNAL OF SOUTHERN COMMUNICATION ENGINEERING
ISLAMIC AZAD UNIVERSITY BUSHEHR BRANCH

E-ISSN: 2980-9231
https://jce.bushehr.iau.ir

https://doi.org/10.30495/jce.2023.1993480.1227

Vol. 14/ No. 53/Autumn 2024

Research Article

Classification of Breast Cancer Masses using Non-Linear Quadratic Support Vector Machine and Comparison with Self-Organizing Neural Network

Soodeh Bakhshandeh, Assistant Professor [1*] | Seyede Monire Atyabi, MSc[2] | Sahar Saberi, Assistant Professor[3]

[1]Department of Computer Engineering, East Tehran Branch, Islamic Azad University, Tehran, Iran, Soodeh.Bakhshandeh@gmail.com, Soodeh.bakhshandeh@iau.ac.ir

[2]Department of Computer Engineering. South Tehran Branch Islamic Azad University. Tehran. Iran, donya.atyabi2017@gmail.com

[3]Department of Computer Engineering. East Tehran Branch, Islamic Azad University, Tehran, Iran, saharsaberi@gmail.com

Correspondence
Soodeh Bakhshandeh, Assistant Professor of Computer Engineering, East Tehran Branch, Islamic Azad University, Tehran, Iran
Soodeh.Bakhshandeh@gmail.com

Received: 1 October 2023
Revised: 12 November 2023
Accepted: 5 December 2023

Abstract

Breast cancer is the second most common cancer after lung cancer and the fifth leading cause of death in women. In less developed countries, breast cancer is the most important cause of death. In this disease, the cells of the breast tissue change and divide into multiple cells and cause a lump. If breast cancer is in the early stages, treatment is possible. There are many treatment methods such as surgery to remove the defective area, drug therapy, radiation therapy, chemotherapy, hormone therapy, and immunotherapy. These treatments have the potential to save lives when administered in the early stages. From the above explanations, it can be seen that early detection of breast cancer is very important and in this research, an attempt has been made to identify suspected cancer data with the quadratic support vector machine method and based on the features extracted from valid and numerous MRI images. Let's classify so that the process of diagnosing the disease in the early stages is easier and faster. The results showed that 356 out of 357 malignant data and 202 out of 211 benign data were correctly classified. The classification accuracy of malignant data was 99.7% and the classification accuracy of benign data was 97.5%, and finally the overall classification accuracy was 98.2%, which indicates the optimal performance of this method in breast cancer data classification.

Keywords: Breast Cancer, Wisconsin Data, Support Vector Machine, Quadratic Kernel.

Highlights

- Developing a method to diagnose breast cancer with the aim of diagnosing the disease in the early stages.
- Applying support vector machine with quadratic kernel with the aim of reducing classification time.
- Using the graphic feature selection method (SU-CFAM) with good speed and performance.
- Achieving 98.2% accuracy without using SU-CFAM, and 99.1% using it.

Citation: S. Bakhshandeh, S. Monire Atyabi, and S. Saberi, "Classification of Breast Cancer Masses using Non-Linear Quadratic Support Vector Machine and Comparison with Self-Organizing Neural Network," *Journal of Southern Communication Engineering*, vol. 14, no. 53, pp. 1–14, 2024, doi:10.30495/jce.2023.1993480.1227, [in Persian].

مقاله پژوهشی

طبقه‌بندی توده‌های سرطانیِ سینه با استفاده از ماشین بردار پشتیبانِ غیرخطیِ کوادراتیک و مقایسه با شبکه عصبی خودسازمان‌ده

سوده بخشنده * [۱] | سیده منیره اطیابی [۲] | سحر صابری [۳]

[۱] گروه مهندسی کامپیوتر، واحد تهران شرق، دانشگاه آزاد اسلامی، تهران، ایران، Soodeh.bakhshandeh@iau.ac.ir

[۲] گروه مهندسی کامپیوتر، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران، donya.atyabi2017@gmail.com

[۳] گروه مهندسی کامپیوتر، واحد تهران شرق، دانشگاه آزاد اسلامی، تهران، ایران، saharsaberi@gmail.com

نویسنده مسئول

*سوده بخشنده، استادیار گروه مهندسی کامپیوتر، واحد تهران شرق، دانشگاه آزاد اسلامی، تهران، ایران، Soodeh.Bakhshandeh@gmail.com

تاریخ دریافت: ۹ مهر ۱۴۰۲
تاریخ بازنگری: ۲۱ آبان ۱۴۰۲
تاریخ پذیرش: ۱۴ آذر ۱۴۰۲

https://doi.org/10.30495/jce.2023.1993480.1227

چکیده:

سرطان سینه بعد از سرطان ریه دومین سرطان شایع و پنجمین دلیل اصلی مرگ و میر در زنان است. تشخیص زودهنگام این سرطان بسیار حائز اهمیت بوده و حتی در صورت تشخیص به موقع این نوع از سرطان، نجات جان افراد نیز امکان‌پذیر است. با در نظر گرفتن این مسئله، در پژوهش روبرو تلاش شده است تا با بهره‌گیری از روش ماشین بردار پشتیبان کوادراتیک و بر اساس ویژگی‌های استخراج شده از تصاویر MRI معتبر، نسبت به طبقه-بندی داده‌های مشکوک به سرطان اقدام گردد تا روند تشخیص بیماری در مراحل اولیه، راحت‌تر و سریع‌تر صورت پذیرد. در این روش به دلیل ماهیت حجم کم محاسبات و بالا بودن سرعت آن در روند آموزش و نهایت آزمایش، ماشین بردار پشتیبان کوادراتیک، انتخاب شده است. در راستای قوی‌تر شدن روش مربوطه، از روش انتخاب ویژگی SU-CFAM که یک روش انتخاب ویژگی مبتنی بر گراف می‌باشد، بهره گرفته شده است. نتایج روش با بهره‌گیری از فاز انتخاب ویژگی و بدون آن مقایسه شد. نتایج نشان داد دقت روش بدون بهره‌گیری از SU-CFAM، ۹۸/۲٪ و با بهره‌گیری از آن به ۹۹/۱٪ رسید که نشان‌دهنده عملکرد مطلوب این روش در طبقه‌بندی داده‌های سرطان سینه است.

کلیدواژه‌ها: داده‌های ویسکانسین، سرطان سینه، کرنل کوادراتیک، ماشین بردار پشتیبان

۱- مقدمه

سرطان یک بیماری کشنده است که در سال اخیر حدود ده میلیون مرگ و ۱۹/۳ میلیون مورد سرطان گزارش شده است. سرطان سینه بعد از سرطان ریه دومین سرطان شایع و پنجمین دلیل اصلی مرگ و میر در زنان است. در کشورهای کمتر توسعه‌یافته سرطان سینه مهم‌ترین عامل مرگ و میر است. در این بیماری، سلول‌های بافت سینه تغییر می‌کنند و به سلول‌های متعدد تقسیم می‌شوند و باعث ایجاد توده می‌شوند. سرطان در مجاری یا لوبول‌هایی که به نوک سینه‌ها متصل هستند شروع می‌شود. بیشتر توده‌های سینه خوش‌خیم و غیر سرطانی هستند و باعث ایجاد فیبروم، حساسیت، ضخیم شدن ناحیه یا توده می‌شوند. اغلب، تومورهای سینه در اندازه کوچک هیچ علامتی ندارند و به راحتی قابل درمان هستند توده بدون درد نشانه سلول‌های غیر طبیعی است. سابقه خانوادگی، عوامل تولید مثل، ویژگی‌های فردی، وزن اضافی بدن، رژیم غذایی، الکل، تنباکو، عوامل محیطی و سایر عوامل خطر، مانند کار در نوبت شب، همگی از مسائل مربوط به سرطان سینه هستند. سرطان سینه در مرحله اولیه به کندی گسترش می‌یابد اما با گذشت زمان بر سایر قسمت‌های بدن تأثیر می‌گذارد. آزمایشات زیادی برای تشخیص

تومورهای سینه توصیه می‌شود، از جمله ماموگرافی، تصویربرداری تشدید مغناطیسی[1] و سونوگرافی. ماموگرافی یک آزمایش مقرون‌به‌صرفه و کم اشعه است که برای تشخیص زودهنگام تومور پستان پیشنهاد می‌شود. MRI یک آزمایش جایگزین است که برای تائید وجود تومور استفاده می‌شود. در طول آزمایش MRI ممکن است واکنش آلرژیک به رنگ کنتراست رخ دهد. این یک پیامد ناخواسته آزمایش MRI است. در مراحل اولیه، آزمایش توصیه شده ماموگرافی است. اگر سرطان سینه در مراحل اولیه باشد ، درمان امکان‌پذیر است. روش‌های درمانی زیادی مانند جراحی برای برداشتن ناحیه معیوب، دارودرمانی، پرتودرمانی، شیمی‌درمانی، هورمون‌درمانی و ایمونوتراپی وجود دارد. این درمان‌ها، زمانی که در مراحل اولیه انجام شوند، پتانسیل نجات جان افراد را دارند. میزان بقا در صورت تشخیص در مرحله اولیه، در کشورهای توسعه یافته ۹۰ درصد، در آفریقای جنوبی ۴۰ درصد و در هند ۶۶ درصد است. کشورهای کم‌درآمد منابع کمتری دارند، بنابراین روش‌های تشخیص زودهنگام و درمان‌ها می‌تواند برای نجات جان زنان مفید باشد[۱].

از توضیحات فوق چنین برمی‌آید که تشخیص زودهنگام سرطان پستان بسیار حائز اهمیت است و در این پژوهش نیز تلاش شده است تا با روش‌های یادگیری ماشین مانند ماشین بردار پشتیبان[2] و بر اساس ویژگی‌های استخراج شده از تصاویر MRI معتبر، داده‌های مشکوک به سرطان را طبقه‌بندی کنیم تا روند تشخیص بیماری در مراحل اولیه، راحت‌تر و سریع‌تر صورت پذیرد. در روش ارائه شده با هدف کاهش زمان طبقه‌بندی، از ماشین بردار پشتیبان با کرنل کودراتیک[3] استفاده نموده است. این کرنل در مقایسه با اغلب کرنل‌های دیگر در این حوزه، دارای ماهیت سرعت بالا در کنار دقت مناسب می‌باشد. از سوی دیگر با توجه به اهمیت بهره‌گیری از روش‌های انتخاب ویژگی در فاز پیش‌پردازش، از روش CU-CFAM، استفاده شده است. این روش انتخاب ویژگی، به دلیل ماهیت گرافی آن دارای سرعت و عملکرد مناسبی می‌باشد و به علاوه امکان بهره‌گیری از آن در حالت با ناظر و بدون ناظر، وجود دارد.

در ادامه و در بخش ۲، به بررسی کارها و پژوهش‌های مرتبط در این حوزه می‌پردازیم. در بخش ۳، روش پیشنهادی مورد بررسی قرار می‌گیرد. در بخش ۴ به بررسی نتایج حاصل از این روش در مقایسه با دیگر روش‌ها پرداخته می‌شود و در نهایت در بخش ۵، به نتیجه‌گیری و جمع‌بندی خواهیم پرداخت.

۲- نگاهی به کارها و پژوهش‌های مرتبط

زرویی و ایلدری، در پژوهشی به بررسی سیستم‌های تصمیم‌گیری مبتنی بر یادگیری ماشین و پردازش تصویر برای تصویربرداری سرطان سینه پرداختند. مقاله آنان به صورت مروری بر ادبیات ساختاریافته با استفاده از تکنیک‌های یادگیری ماشین و پردازش تصویر برای تصویربرداری سرطان سینه انجام شد. مجموعه‌ای از ۵۳۰ مقاله منتشر شده بین سال‌های ۲۰۰۰ تا اوت ۲۰۱۹ بر اساس ده معیار انتخاب و تجزیه و تحلیل شدند. معیارهای مورد بررسی شامل سال و کانال انتشار، نوع تجربی، نوع تحقیق، کار پزشکی، تکنیک‌های یادگیری ماشین، مجموعه داده‌های مورد استفاده، روش‌های اعتبارسنجی، معیارهای عملکرد و تکنیک‌های پردازش تصویر که شامل پیش‌پردازش تصویر، تقسیم‌بندی، استخراج ویژگی و انتخاب ویژگی، می‌باشد. نتایج نشان داد که تکنیک‌های یادگیری عمیق تا حد زیادی برای انجام طبقه‌بندی استفاده می‌شوند. اکثر مطالعات انتخاب شده از ماموگرافی به عنوان روش‌های تصویربرداری به جای تصویربرداری اولتراسوند یا MRI، استفاده نمودند. در مورد تکنیک‌های پردازش تصویر، در اکثر مطالعات انتخاب‌شده، در مرحله پیش‌پردازش ،تصاویر ورودی را با کاهش نویز و عادی‌سازی رنگ‌ها تغییر داده و در برخی از آن‌ها از تقسیم‌بندی برای استخراج منطقه مورد نظر با روش آستانه‌گذاری استفاده می‌کنند [۲].

اِکیچی و جازال، در پژوهشی به بررسی تشخیص سرطان سینه با استفاده از ترموگرافی و شبکه های عصبی کانولوشنال[4]، پرداختند. در پژوهش آنان، یک الگوریتم جدید برای استخراج ویژگی‌های مشخصه پستان بر اساس داده‌های زیستی، تجزیه و تحلیل تصویر و آمار تصویر پیشنهاد شده است. این ویژگی‌ها از تصاویر حرارتی گرفته شده توسط یک دوربین حرارتی استخراج

[1] Magnetic Resonance Imaging (MRI)
[2] Support Vector Machine (SVM)
[3] Quadratic Kernel
[4] Convolutional Neural Network

شده‌اند و برای طبقه‌بندی تصاویر پستان با استفاده از شبکه‌های عصبی کانولوشنال بهینه‌سازی شده توسط الگوریتم بیز[1]، استفاده شده است. با استفاده از الگوریتم پیشنهادی آنان، نرخ دقت ۹۸/۹۵٪ برای تصاویر حرارتی در مجموعه داده متعلق به ۱۴۰ نفر به دست آمد [۳].

سادوخان و همکاران، در پژوهشی به تشخیص سرطان سینه با استفاده از پردازش تصویر و یادگیری ماشین پرداختند. آنان روش کامپیوتری برای تشخیص سرطان در مراحل اولیه خود را در مدت زمان بسیار کوتاه ارائه دادند. به عبارتی از یادگیری ماشینی برای آموزش مدلی با استفاده از ویژگی‌های پیش‌بینی شده هسته سلول‌ها استفاده نمودند. طی پژوهش آنان، یک مطالعه مقایسه‌ای از دو الگوریتم مختلف K-نزدیکترین همسایگی [2]و SVM انجام شد که در آن دقت هر طبقه‌بندی اندازه‌گیری شده بود. پس از این، آنان یک تصویر دیجیتالی از آسپیره بافت سینه با استفاده از پردازش تصویر تجزیه و تحلیل کردند تا ویژگی‌های هسته سلول‌ها را دریابند. سپس مقادیر ویژگی را در مدل آموزش دیده خود اعمال نمودند تا متوجه شوند که آیا تومور توسعه یافته خوش‌خیم است یا بدخیم [۴].

ساهنی و میتال، در پژوهشی به بررسی تشخیص سرطان سینه با استفاده از تکنیک‌های پردازش تصویر پرداختند. در کار پیشنهادی، تصاویر دریافتی از ماموگرافی و MRI، به عنوان دو روش در تشخیص تومور، به عنوان ورودی روش مورد استفاده قرار گرفته است. در یک روش، قسمت مرتبط با تومور از تصویر حاصل با روش‌های تقسیم‌بندی مختلف مانند تشخیص لبه و روش آستانه جدا می‌شود. به علاوه عملگرهای مختلفی بر روی تصویر حاصل اعمال شده است و از نظر کمی توسط آنتروپی پارامتر اندازه‌گیری، عملکرد روش تائید شده است [۵].

اَدِل و همکاران، در پژوهشی به تشخیص سرطان سینه با استفاده از پردازش تصویر و یادگیری ماشینی برای تصاویر الاستوگرافی پرداختند. در این روش، تکنیک‌های پردازش تصویر برای استخراج ویژگی بر روی تصاویر به دست آمده اعمال می‌شود. در ادامه، روش‌های پیش‌پردازش داده‌ها و تحلیل مؤلفه‌های اصلی به‌عنوان یک تکنیک کاهش ابعاد برای مجموعه داده اعمال می‌شوند. اعتبارسنجی مدل با اعتبارسنجی متقاطع K-fold انجام می‌شود تا از تعمیم الگوریتم اطمینان حاصل شود. سپس دقت، ماتریس سردرگمی و تلفات لجستیک برای الگوریتم مورد استفاده ارزیابی می‌شوند. در این روش، حداکثر دقت طبقه‌بندی در هنگام استفاده از SVM با هسته تابع پایه شعاعی ۹۴/۱۲٪ است [۶].

گاردِزی و همکاران، در پژوهشی به تشخیص سرطان پستان با استفاده از داده‌های ماموگرافی پرداختند. این بررسی با هدف بررسی ادبیات سنتی یادگیری ماشین و یادگیری عمیق با کاربرد خاص برای تشخیص سرطان پستان انجام شد. این بررسی همچنین بینشی مختصر در مورد برخی از شبکه‌های یادگیری عمیق معروف ارائه می‌دهد. در پژوهش آنان، مروری بر تکنیک‌های یادگیری ماشین و یادگیری عمیق با کاربرد خاص برای سرطان پستان ارائه شد [۷].

وارلامیس و همکاران، در پژوهشی به "استفاده از تکنیک‌های داده‌کاوی و روش‌های تجزیه و تحلیل داده‌ها برای اندازه‌گیری داده‌های مرگ و میر و عوارض سرطان در یک ثبت سرطان منطقه‌ای: مورد جزیره کرت، یونان" پرداختند. آنان معتقدند که بررسی آمار سرطان بسیار مهم است زیرا این امر مستلزم برنامه‌ریزی طولانی مدت، یادگیری احتمالی و مشاهده مداوم هر بیمار سرطانی است. هدف این مطالعه نشان دادن چگونگی استفاده از فرآیندهای داده‌کاوی برای بهبود نتایج تجزیه و تحلیل آماری از داده‌های ثبت سرطان بود [۸].

تان و همکاران، در پژوهشی به ارتباط بین تغییرات در ویژگی‌های تصویر ماموگرافی و خطر ابتلا به سرطان پستان کوتاه مدت پرداختند. آنان بر اساس بررسی کمی تمایزات ویژگی‌های تصویر ماموگرافی دوطرفه در مجموعه تصاویر ماموگرافی دیجیتال، یک مدل محاسباتی جدید برای پیش‌بینی خطر ابتلا به سرطان پستان در آینده نزدیک توسعه و آزمایش کردند [۹].

یان و همکاران، پیشنهاد دادند از شبکه‌های عصبی خود رمزنگار به عنوان یک دسته‌بندی کننده برای تشخیص مناطق سرطانی پروستات در تصاویر MRI استفاده گردد. در این روش پیشنهادی، روش کمینه‌سازی انرژی برای شناسایی محل نگاشت، با نظر گرفتن ارتباط میان پیکسل‌های مجاور معرفی شد [۱۰].

[1] Bayesian Algorithm
[2] K-Nearest Neighbors

ژائو و همکاران، در پژوهشی به غربالگری خودکار سلول‌های دهانه رحم با استفاده از پردازش تصویر بلوک، پرداختند. در این مطالعه، یک رویکرد تحلیلی به کمک کامپیوتر برای شناسایی وجود سلول‌های مشکوک در یک تصویر سلولی کامل دهانه رحم پیشنهاد شده است. تفاوت اصلی روش آنان با الگوریتم معمولی در این است که تصویر به جای سلول‌های قطعه‌بندی شده به بلوک‌هایی با اندازه معین تقسیم می‌شود که می‌تواند پیچیدگی محاسباتی را تا حد زیادی کاهش دهد. از طریق تجزیه و تحلیل داده‌ها، برخی از ویژگی‌های بافت و رنگ هیستوگرام تفاوت‌های قابل توجهی را بین بلوک‌های با و بدون سلول‌های مشکوک نشان می‌دهد. بنابراین این ویژگی‌ها را می‌توان به عنوان ورودی طبقه‌بندی کننده SVM استفاده کرد [۱۱].

در [۱۲]، سیواکامی و ساراواثی، در پژوهش خود به بررسی استخراج داده‌های بزرگ: پیش‌بینی سرطان سینه با استفاده از مدل ترکیبی DT-SVM [۱] پرداختند. این مطالعه با استفاده از مجموعه داده‌های سرطان پستان ویسکانسین [۲] انجام شد. در روش ارائه شده در این مقاله دقت به دست آمده ۹۱٪ با ضریب خطای ۲/۵۸٪ بوده است. در این مقاله به مقایسه الگوریتم ارائه شده با دیگر روش‌ها شامل سایر الگوریتم‌های طبقه‌بندی نیز مانند IBL[3] ، SMO [۴] و Naïve Bayes نیز پرداخته شده است. برای IBL، دقت به‌دست‌آمده ۸۵/۲۳ درصد با نرخ خطای ۱۲/۶۳ درصد بود. برای SMO، دقت ۷۲/۵۶٪ با نرخ خطا ۵/۹۶٪ بود. برای Naïve Bayes، دقت به دست آمده ۸۹/۴۸٪ با نرخ خطا ۹/۸۹٪ بود. بنابراین این مطالعه مقایسه‌ای نشان داد که DT-SVM بهتر از هر الگوریتم طبقه‌بندی دیگری عمل می‌کند.

۳- روش پیشنهادی

۳-۱- فاز پیش‌پردازش

دو راهکار عمده برای کاهش ابعاد [۵] مجموعه داده‌ای ارائه شده است : استخراج ویژگی [۶] و انتخاب ویژگی [۷]. در استخراج ویژگی، فضای اولیه ویژگی‌ها به یک فضای کوچک‌تر نگاشت می‌شود. درواقع، در این راهکار، با ترکیب ویژگی‌های موجود، تعداد کمتری ویژگی ایجاد می‌شود به طوری که این ویژگی‌ها دارای تمام (یا بخش اعظمی از) اطلاعات موجود در ویژگی‌های اولیه باشند. از طرف دیگر، در انتخاب ویژگی، یک زیرمجموعه از ویژگی‌های اولیه انتخاب می‌شود. انتخاب ویژگی یک تکنیک مهم و پراستفاده در پیش‌پردازش داده‌ها محسوب می‌شود که موجب افزایش سرعت الگوریتم‌های یادگیری ماشین و بهبود عملکرد طبقه‌بندی کننده می‌شود.

در این مقاله از روش نقشه عدم قطعیت متقارن کلاس-ویژگی انجمنی برای انتخاب ویژگی [۸] استفاده شده است. در این روش، در ابتدای کار نمایش گرافی از ویژگی‌ها استخراج می‌گردد. بدین منظور با بهره‌گیری از یک پارامتر β و مفهوم SU [۹] گراف وزنی به شکل زیر استخراج می‌گردد:

$$w_{ij}=\begin{cases}\beta\dfrac{\sum_{t\in\{i,j\}}SU(G_t,C)}{2}-(1-\beta)SU(G_i,G_j), & 0\le\beta\le1, i\neq j\\ 1, & otherwise\end{cases} \quad (۱)$$

که به تفکیک $SU(G_i,G_j)$ مقدار همبستگی-F میان دو ویژگی G_i و G_j و $SU(G_i,C)$ و $SU(G_j,C)$ مقدار وابستگی-T میان دو ویژگی G_i و G_j و کلاس C می‌باشد. برای نرمال‌سازی وزن‌ها، از روش Softmax استفاده می‌گردد و وزن نهایی براساس رابطه ۲، استخراج می‌گردد:

$$\hat{w}_{ij}=\frac{1}{1+\exp\left(-\dfrac{w_{ij}-\bar{w}}{\sigma}\right)} \quad (۲)$$

[1] Decision Tree - Support Vector Machine
[2] Breast Cancer Wisconsin (Diagnostic)
[3] Instance-Based Learning
[4] Sequential Minimal Optimization
[5] Dimensionality reduction
[6] Feature extraction
[7] Feature selection
[8] Symmetric Uncertainty Class-Feature Association Map (SU-CFAM)
[9] Symmetric Uncertainty

که در آن به ترتیب، w_{ij} وزن یال بین گره F_i و F_j و $\overline{w}$ و σ میانگین و واریانس تمامی وزن‌های یال‌ها در گراف می‌باشند. پس از اعمال رابطه ۲ بر روی ویژگی‌ها، گرافی حاصل می‌گردد که شباهت میان ژن‌ها و شباهت میان ژن و برچسب کلاس را در برخواهد داشت.

پس از ایجاد گراف در مرحله قبل، الگوریتم Louvain بر روی آن اعمال گردیده و در نهایت خوشه‌های مختلفی از گره‌ها (ژن‌ها)، بر اساس وزن یال‌ها، حاصل می‌گردد. با این کار تمام گره‌های داخل یک خوشه بیشترین میزان شباهت را با هم دارا بوده و گره‌های خارج از آن خوشه کمترین شباهت را با گره‌های یک خوشه دارا می‌باشند. در این مرحله خوشه‌های ایجاد شده را زیرگراف می‌نامیم که ورودی مرحله بعد تلقی می‌گردند.

در مرحله انتخاب بهترین ویژگی در هر خوشه، در ابتدا ماتریس‌های مجاورت که ماتریس‌های دودویی می‌باشند ساخته می‌شود. برای تبدیل ماتریس‌های شباهت به ماتریس مجاورت از حد آستانه γ استفاده می‌گردد و هر وزن در گراف مربوطه با این حد آستانه مقایسه می‌گردد،

$$a_{ij} = \begin{cases} 1, & if\ \hat{w}_{ij} > \gamma\ and\ i \neq j \\ 0, & otherwise \end{cases} \quad (۳)$$

که در آن $\hat{w}_{ij}$ وزن نرمال شده است که دربردارنده شباهت میان ویژگی‌ها و ویژگی و کلاس می‌باشد. اعمال رابطه ۳ به زیرگراف‌های حاصل از مرحله قبل، منجر به ایجاد زیرگراف‌های مجاورت می‌شود که به آن نقشه‌های کلاس-ویژگی انجمنی[1] گفته می‌شود. در ادامه و برای انتخاب بهترین ویژگی‌ها، تمامی CFAMهای موجود در مجموعه مستقل بیشینه انتخاب می‌گردند [۱۳].

۳-۲- طبقه‌بندی

در یادگیری ماشینی، SVM، که شبکه بردار پشتیبان نیز نامیده می‌شود، یک تکنیک یادگیری نظارت شده است که با یافتن ابر صفحه‌ای که حاشیه بین دو کلاس را به حداکثر می‌رساند، طبقه‌بندی را انجام می‌دهد.

مراحل اصلی الگوریتم SVM به شرح زیر ارائه شده است:

- یک ابر صفحه بهینه را تعریف کنید: حاشیه را به حداکثر برسانید.
- تعریف فوق را برای مسائل غیرخطی قابل‌تفکیک بسط دهید: برای طبقه‌بندی اشتباه یک جریمه(پنالتی) تعریف کنید.
- نگاشت داده‌ها به فضای با ابعاد بالا که در آن طبقه‌بندی با سطوح تصمیم خطی آسان‌تر است: مسئله را دوباره فرمول‌بندی کنید تا داده‌ها به طور ضمنی به این فضا نگاشت شوند.

برای جدا کردن دو گروه، ساده‌ترین راه با یک خط مستقیم (۱ بعد)، صفحه مسطح (۲ بعد) یا یک ابر صفحه N بعدی است. با این حال، در برخی موارد، یک منطقه غیرخطی می‌تواند داده‌ها را به طور موثرتری از هم جدا کند. بنابراین، در جایی که نمی‌توان از یک ابرصفحه خطی برای جدا کردن دو گروه استفاده کرد، SVM از دو راه حل استفاده می‌کند، یکی "تابع کرنل" و دیگری "نگاشت داده‌ها به فضایی با ابعاد بالا" تا بدین وسیله داده‌ها را برای انجام جداسازی خطی ممکن کند.

۳-۳- نگاشت داده‌ها به فضایی با ابعاد بالا

برای نگاشت داده‌ها در فضای جدید، باید تابع $\phi(x)$ را پیدا کنیم. فرمول SVM به‌صورت زیر می‌شود:

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\varepsilon_i \\ &s.t. y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \varepsilon_i, i = 1,...,m \\ &\varepsilon_i \geq 0, i = 1,...,m \end{aligned} \quad (۴)$$

[1] Class-Feature Association Maps (CFAM)

که در آن داده‌های آزمایش به صورت زوج مرتب (x_i , y_i) ، ω بردار نرمال عمود بر ابرصفحه، ε_i متغیر کمکی، می‌باشد. در این حالت داده‌ها به صورت $\phi(x)$ ظاهر می‌شوند و وزن‌های ω اکنون وزن‌هایی در فضای جدید هستند. علاوه بر این، اگر $\phi(x)$ ابعاد بسیار بالایی داشته باشد، نگاشت بسیار پیچیده خواهد شد.

برای این حالت و در مسائل دو کلاسه SVM مسئله لاگرانژ به صورت زیر تعریف می‌شود،

$$\begin{aligned}
&\max_{\alpha} J(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\mathrm{K}(x^{(i)},x^{(j)})\\
&s.t.\,\mathrm{K}(x^{(i)},x^{(j)})=\left\langle\phi(x^{(i)}),\phi(x^{(j)})\right\rangle\\
&0\le\alpha_i\le C, i=1,\ldots,m\\
&\sum_{i=1}^{m}\alpha_i y^{(i)}=0
\end{aligned} \qquad (۵)$$

که در آن α همان ضرایب لاگرانژ هستند. حال مساله به یک مساله بهینه‌سازی درجه دوم تبدیل شده و می‌شود آن را با راه حل‌های مربوط به مسائل درجه دوم محاسبه نمود و مقادیر ضرایب لاگرانژ را با آن محاسبه کرد. مقدار ضرایب لاگرانژ برای اعضای عضو بردارهای پشتیبان دارای مقدار و برای باقی اعضا معادل صفر خواهد بود. تفاوتی که در اینجا با حالت خطی مشاهده می‌شود، وجود تابع K است (تابع کرنل) که ضرب داخلی داده‌ها در فضای نگاشت شده است. یعنی ما اول داده‌ها را با تابع کرنل نگاشت داده و پس از آن ضرب داخلی آن را محاسبه می‌کنیم. مساله اصلی در این حالت تعیین معیاری برای تشخیص کلاس داده‌ها است. بر این اساس و با حل مساله لاگرانژ بیان شده در فرمول ۵، به رابطه زیر می‌رسیم که در اصل تعیین کننده کلاس داده‌ها خواهد بود،

$$\begin{aligned}
&b=y^{(k)}-\sum_{i,j=1}^{m}\alpha_i y^{(i)}\mathrm{K}(x^{(i)},x^{(k)})\\
&f(x)=sign(\sum_{i,j=1}^{m}\alpha_i y^{(i)}\mathrm{K}(x^{(i)},x^{(k)})+b)
\end{aligned} \qquad (۶)$$

همان‌طور که در رابطه فوق می‌بینیم دیگر خبری از ω نیست. فقط کافی است b و مقدار نهایی را با استفاده از ضرب داخلی داده‌ها در فضای نگاشت شده به دست آوریم.

۳-۴- تابع کرنل

زمانی که نگاشت داده‌ها در فضایی با ابعاد بالا بسیار پیچیده است، به عنوان یک راه حل جایگزین، از تابع کرنل استفاده می‌کنیم، که به صورت زیر تعریف می شود،

$$\mathrm{K}(x,x_i)=(\phi(x_i)^T.\phi(x_j)) \qquad (۷)$$

با راه حل تابع کرنل ، محاسبه ابرصفحه جداکننده به یافتن $\phi(x)$ نیاز ندارد. بنابراین، می‌توان از هر تابع کرنلی برای توصیف مجدد در فضای با ابعاد بزرگ‌تر استفاده کرد. توابع کرنل مختلفی ارائه شده است، از جمله گوسین، چند جمله‌ای، کوادراتیک و RBF[1]. توجه داشته باشیم که هر تابع کرنل مزایا و معایب خود را دارد. بنابراین انتخاب نوع کرنل و پارامترهای آن به عهده کاربر و بر اساس مسئله پیش رو خواهد بود.

در پژوهش حاضر از کرنل کوادراتیک استفاده شده است. تابع کرنل کوادراتیک (درجه دوم) نسبت به تابع کرنل گاوسی، محاسبات کمتری دارد و می‌تواند به عنوان یک راه حل جایگزین برای زمانی که استفاده از هسته گاوسی بسیار پیچیده می‌شود استفاده شود. تابع کوادراتیک را می‌توان به صورت زیر نوشت،

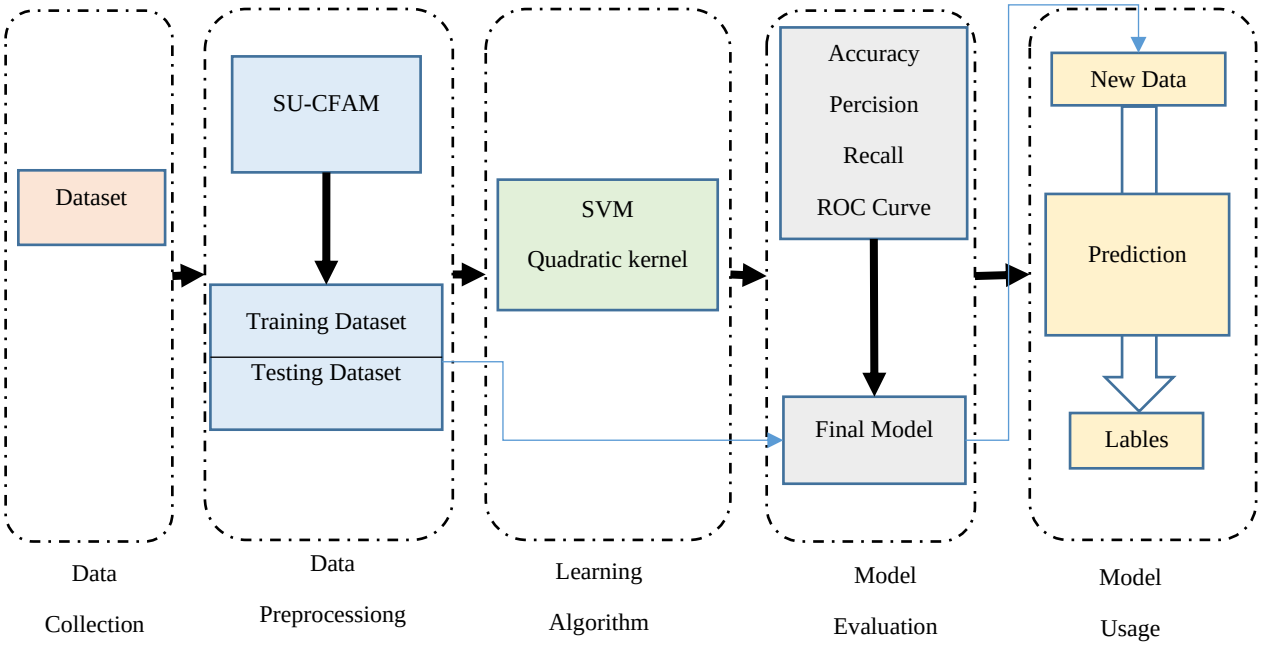
$$\mathrm{K}(x,x_i)=((x*x_i)+1)^2 \qquad (۸)$$

کرنل درجه دوم برای تبدیل داده‌های ورودی به فضایی با ابعاد بالاتر استفاده می‌شود که اجازه می‌دهد تا مرزهای تصمیم‌گیری غیرخطی را ایجاد کند. حاصل‌ضرب نقطه بین دو بردار ورودی و همچنین مربع حاصل‌ضرب نقطه را محاسبه می‌کند. این بدان معنی است که نه تنها ترکیب خطی متغیرهای ورودی بلکه تعامل آن‌ها را نیز در نظر می‌گیرد. که در آن با اعمال مجموعه

[1] Radial Basis Function

داده‌های آموزشی به مدل و استفاده از فرمول فوق به عنوان تابع کرنل، سعی در طبقه‌بندی داده‌ها با ضریب دقت بالایی را خواهیم داشت [۱۴].

در این مقاله ماشین بردار پشتیبان انتخابی از کرنل کودراتیک بهره برده و از سوی دیگر با توجه به اهمیت بهره‌گیری از روش‌های انتخاب ویژگی در فاز پیش‌پردازش، از روش CU-CFAM، در فاز پیش‌پردازش، استفاده شده است. این روش انتخاب ویژگی، به دلیل ماهیت گرافی آن دارای سرعت و عملکرد مناسبی می‌باشد و به علاوه امکان بهره‌گیری از آن در حالت با ناظر و بدون ناظر، وجود دارد. دیاگرام روش پیشنهادی در شکل ۱ ارائه شده است.



شکل ۱: بلوک دیاگرام روش پیشنهادی
Figure 1. Diagram of the proposed method

۴- نتایج

در این بخش با استفاده از کدنویسی در محیط اسکریپت نرم‌افزار متلب[1]، نسبت به پیاده‌سازی و بهره‌گیری از روش SU-CFAM در فاز پیش‌پردازش و روش SVM کوادراتیک در فاز طبقه‌بندی، اقدام نموده‌ایم. مجموعه داده مورد استفاده در این مقاله، مجموعه داده مربوط به سرطان سینه دانشگاه ویسکانسین می‌باشد که جزو معتبرترین مجموعه‌های داده این حوزه می‌باشد که در بخش بعد، جزییات آن شرح داده شده است.

با اعمال روش مذکور، نتایج مربوط به دقت تشخیص سیستم آموزش داده شده توسط این روش را مورد بررسی قرار می‌دهیم. در داده‌های فراخوانی شده، بافت‌های سرطانی با عناوین خوش‌خیم(M) و بدخیم(B) علامت‌گذاری شده‌اند و ویژگی‌های مختلف آن‌ها نیز ثبت شده که به عنوان داده‌های آموزشی و تستی به SVM کوادراتیک طراحی شده داده می‌شود تا عمل طبقه‌بندی روی آن‌ها صورت پذیرد.

۴-۱- جامعه آماری و مشخصات نمونه

در مقاله حاضر، داده‌های مربوط به سرطان سینه از دانشگاه ویسکانسین گرفته شده و یکی از معتبرترین و کامل‌ترین داده‌ها در این حوزه می‌باشد و معروف به داده‌ی WDBC[2] است [۱۵]. بافت‌های سرطانی با عناوین خوش‌خیم و بدخیم علامت‌گذاری

[1] Matlab R2022b
[2] Breast Cancer Wisconsin (Diagnostic)

شده‌اند و ویژگی‌های مختلف آن‌ها نیز ثبت شده و به‌عنوان داده‌های آموزشی و تستی به SVM طراحی شده داده می‌شود تا عمل طبقه‌بندی روی آن‌ها صورت پذیرد.
مشخصات نمونه‌های استفاده شده به شرح جدول ۱ می‌باشد.

جدول ۱: جزییات پایگاه داده WDBC
Table 1. WDBC Dataset Details

تعداد کل نمونه: ۵۶۹
تعداد ویژگی‌ها: ۳۲ ویژگی شامل شناسه، لاس (خوش‌خیم یا بدخیم) و ۳۰ ویژگی ورودی **این ۳۰ ویژگی مذکور شامل ۳ مقدارِ میانگین، خطای استاندارد و بدترین یا بزرگ‌ترین (میانگین سه مقدار بزرگ) که برای هرکدام از ده ویژگی ذیل (که برای هر هسته سلول محاسبه شده است) به دست آمده:** • **شعاع (میانگین فواصل مرکز تا نقاط پیرامون)** • **بافت (انحراف استاندارد مقادیر مقیاس خاکستری)** • **محیط** • **مساحت** • **میزان صاف بودن (تغییر موضعی در طول شعاع)** • **فشردگی** • **تقعر (شدت بخش‌های مقعر کانتور)** • **نقاط مقعر (تعداد قسمت‌های مقعر کانتور)** • **تقارن** • **بعد شکستآر (فراکتال)**

۴-۲- معیارهای ارزیابی

با بهره‌گیری از مقادیر TP به عنوان مثبت صحیح[1]، FP مثبت اشتباه[2]، TN منفی صحیح[3]، FN منفی اشتباه[4] و Total تعداد کل نمونه‌ها، معیارهای ارزیابی استفاده شده در این مقاله بدین شرح می‌باشند:

- **دقت (Accuracy)**

متداول‌ترین معیار کارایی برای الگوریتم‌های طبقه‌بندی است که می‌توان آن را در قالب تعداد پیش‌بینی‌های صحیح صورت گرفته به نسبت همه پیش‌بینی‌های صورت گرفته، تعریف کرد.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (۹)$$

- **صحت (Precision)**

وقتی که مدل نتیجه را مثبت پیش‌بینی می‌کند، این معیار ارزیابی می‌کند که تا چه اندازه نتیجه حاصله درست می‌باشد. زمانی که ارزش FP بالا باشد، معیار صحت، معیار مناسبی خواهد بود.

$$Percision = \frac{TP}{TP + FP} \quad (۱۰)$$

- **Recall:**

زمانی که ارزش FN بالا باشد، معیارRecall ، معیار مناسبی خواهد بود.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (۱۱)$$

[1] True Positive
[2] False Positive
[3] True Negative
[4] False Negative

- **منحنی ROC[1]:**

یکی از روش‌های بررسی و ارزیابی عملکرد دسته‌بندی دودویی، نمودار مشخصه عملکرد است. کارایی الگوریتم‌های دسته‌بندهای دودویی معمولاً به وسیله شاخص‌هایی مثل حساسیت[2] و صراحت[3] سنجیده می‌شوند. اما در نمودار ROC هر دوی این شاخص‌ها ترکیب شده و به صورت یک منحنی نمایش داده می‌شوند. در این قسمت به جزییات نتایج حاصل از روش و مقایسه آن با دیگر روش‌ها می‌پردازیم.

در گام اول، روش ارائه شده را بدون بهره‌گیری از روش انتخاب ویژگی CU-CFAM و با استفاده از آن، بررسی می‌کنیم. در جدول ۲ و ۳، به بررسی تعداد موارد درست و نادرست تشخیص و دقت طبقه‌بندی با استفاده از روش SVM کوادراتیک، بدون بهره‌گیری از SU-CFAM و با استفاده از آن می‌پردازیم. همان‌گونه که مشاهده می‌فرمایید روش ارائه شده در زمان عدم استفاده از SU-CFAM به دقت ۹۸/۲٪ و با بهره‌گیری از SU-CFAM در فاز پیش‌پردازش، به دقت ۹۹/۱٪ رسیده است که دقت بسیار مناسبی برای تشخیص این بیماری می‌باشد.

جدول۲: تعداد موارد درست و نادرست تشخیص (روش SVM کوادراتیک، بدون بهره‌گیری از SU-CFAM و با استفاده از آن)

Table2. The numbers of true and fals detection (Quadratic SVM, without and with using SU-CFAM)

روش	نوع تشخیص	بدخیم	خوش‌خیم	درصد از داده‌های بدخیم	درصد از داده‌های خوش‌خیم
بدون SU-CFAM	درست	۳۵۶ مورد	۲۰۲ مورد	۶۲/۷	۳۵/۶
	نادرست	۱ مورد	۹ مورد	۰/۲	۱/۶
با SU-CFAM	درست	۳۵۷ مورد	۲۰۶ مورد	۶۲/۹	۳۶/۴
	نادرست	۰ مورد	۵ مورد	۰	۰/۸

جدول۳: دقت طبقه‌بندی (روش SVM کوادراتیک، بدون بهره‌گیری از SU-CFAM و با استفاده از آن)

Table 3. The accuracy of the detection (Quadratic SVM, without and with using SU-CFAM)

روش	درصد طبقه‌بندی درست نهایی	درصد طبقه‌بندی نادرست نهایی
بدون SU-CFAM	۹۸/۲	۱/۸
با SU-CFAM	۹۹/۱	۰/۹

در جدول ۴، مقایسه‌ای میان بهره‌گیری از توابع مختلف در ماشین بردار پشتیبان مورد استفاده در این مقاله، انجام گرفته است. در این مقایسه برای تمامی روش‌های ارائه شده، در دو حالت بدون بهره‌گیری از SU-CFAM و با استفاده از آن، مقادیر حاصل از دقت، Precision و Recall ارائه شده است. مشاهده می‌شود که در این مقایسه هم روش پیشنهادی به دقت مناسبی در مقایسه با دیگر روش‌ها رسیده است.

جدول۴: مقایسه استفاده از SVM با کرنل‌های مختلف بدون بهره‌گیری از SU-CFAM و با استفاده از آن بر روی سه پارامتر از دقت، Precision و Recall

Table 4. Comparision of different kernels of SVM without and with using SU-CFAM on three parameters (Accuracy, Precision and Recall)

SVM با توابع کرنل متفاوت		دقت روش	Precision	Recall
کوادراتیک	بدون SU-CFAM	**۹۸/۲**	۹۷/۶	**۹۶/۵**
	با SU-CFAM	**۹۹/۱**	**۹۸/۱**	**۹۷/۳**
گاوسی	بدون SU-CFAM	۹۶/۰	۹۳/۲	۹۴/۹
	با SU-CFAM	۹۷/۱	۹۳/۴	۹۵/۱
پایه شعاعی گاوسی (RBF)	بدون SU-CFAM	۹۷/۶	۹۴/۵	۹۶/۴
	با SU-CFAM	۹۶/۸	۹۵/۲	۹۶/۳
تانژانت هایپربولیک (tanh)	بدون SU-CFAM	۹۸/۱	**۹۷/۸**	۹۵/۱
	با SU-CFAM	۹۹/۰	۹۶/۸	۹۷/۱

[1] Receiver Operating Characteristic (ROC)
[2] Sensitivity
[3] Specificity

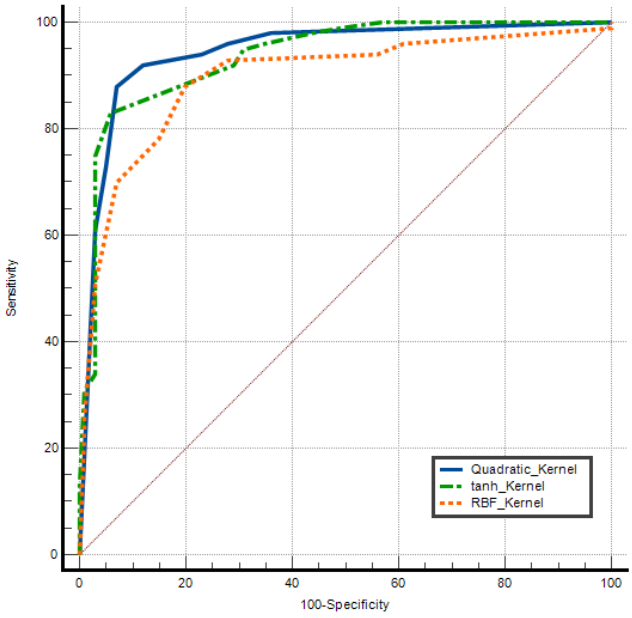
جدول ۵ به مقایسه میان دو روش انتخاب ویژگی PCA[1] و SU-CFAM در فاز انتخاب ویژگی، می‌پردازد. تحلیل مؤلفه‌های اصلی تبدیلی در فضای برداری است، که تحلیل مجموعه داده‌های بزرگ با تعداد زیادی بعد یا ویژگی، افزایش تفسیرپذیری داده‌ها با حفظ حداکثر مقدار اطلاعات و تجسم داده‌های چندبعدی را فراهم می‌کند. تحلیل مؤلفه‌های اصلی در واقع، یک تکنیک آماری برای کاهش ابعاد یک مجموعه داده است. این کار با تبدیل خطی داده‌ها به یک سیستم مختصات جدید انجام می‌شود که (بیشتر) تغییرات در داده‌ها را می‌توان با ابعاد کمتری نسبت به داده‌های اولیه توصیف کرد. بسیاری از مطالعات از دو مؤلفه اصلی اول برای نمایش داده‌ها در دو بعد و شناسایی بصری خوشه‌های نقاط داده نزدیک به هم استفاده می‌کنند.

جدول۵: مقایسه استفاده از SVM با کرنل‌های مختلف با بهره‌گیری از PCA و SU-CFAM و بر روی سه پارامتر از دقت، Precision و Recall

Table 5. Comparision of different kernels of SVM with using PCA and SU-CFAM on three parameters (Accuracy, Precision and Recall)

SVM با توابع کرنل متفاوت		دقت روش	Precision	Recall
کوادراتیک	روش PCA	**۹۹/۳**	۹۸/۰	**۹۷/۳**
	روش SU-CFAM	**۹۹/۵**	**۹۹/۲**	**۹۸/۵**
گاوسی	روش PCA	۹۶/۳	۹۴/۵	۹۶/۵
	روش SU-CFAM	۹۷/۵	۹۵/۴	۹۵/۷
پایه شعاعی گاوسی	روش PCA	۹۸/۳	۹۴/۲	۹۶/۴
(RBF)	روش SU-CFAM	۹۷/۴	۹۶/۹	۹۷/۲
تانژانت هایپربولیک	روش PCA	۹۸/۴	۹۷/۲	۹۴/۵
(tanh)	روش SU-CFAM	۹۸/۱	**۹۸/۱**	۹۶/۹

شکل ۲، نمودار ROC حاصل از اعمال روش ارائه شده با سه کرنل کوادراتیک، RBF و tanh و با بهره‌گیری از روش انتخاب ویژگی SU-CFAM را نشان می‌دهد. این نمودار نیز نشان‌دهنده برتری استفاده از کرنل کودراتیک در مقایسه با دیگر کرنل‌ها می‌باشد.



شکل ۲: نمودار ROCحاصل از اعمال روش ارائه شده با سه کرنل کوادراتیک، RBF و tanh و با بهره‌گیری از روش انتخاب ویژگی SU-CFAM

Figure 2. ROC curve of the proposed method with three kernels (Quadratic, RBF and Tanh) with using SU-CFAM

[1] Principal Component Analysis (PCA)

در مقایسه‌ای دیگر، نتایج حاصل از روش نگاشت خودسازمان‌ده[1] با روش ارائه شده در این مقاله، مورد بررسی قرار گرفت. در هر دو روش از روش انتخاب ویژگی SU-CFAM استفاده شده است. نتایج مقایسه‌ای این دو روش در جدول ۶ و ۷ آورده شده است. با دقت در اعداد جداول فوق و مقایسه با ارقام مربوط به خروجی روش SVM کوادراتیک مشخص می‌شود که درصد دقت نهایی طبقه‌بندی با استفاده از روش این پژوهش به میزان حدوداً ۱۴ درصد بیشتر از روش SOM بوده که این تفاوت فاحش، بیشتر به دلیل طبقه‌بندی موارد خوش‌خیم ایجاد شده است که البته می‌تواند مربوط به تعداد دیتاهای کمتر موارد خوش‌خیم نسبت به دیتاهای موارد بدخیم بوده باشد و همین قضیه نشان می‌دهد روش SVM کوادراتیک پیاده‌سازی شده حتی در مواردی که دیتاهای کمتری در دسترس می‌باشد نیز از دقت بالایی در طبقه‌بندی برخوردار است.

جدول۶: تعداد موارد درست و نادرست تشخیص (روش SVM کوادراتیک و روش SOM)

Table 6. The numbers of true and fals detection (Quadratic SVM, SOM)

روش	نوع تشخیص	بدخیم	خوش‌خیم	درصد از داده‌های بدخیم	درصد از داده‌های خوش‌خیم
پژوهش حاضر	درست	۳۵۷ مورد	۲۰۶ مورد	۶۲/۹	۳۶/۴
	نادرست	۰ مورد	۵ مورد	۰	۰/۸
SOM	درست	۳۵۶ مورد	۱۲۹ مورد	۶۲/۷	۲۲/۷
	نادرست	۱ مورد	۸۲ مورد	۰/۲	۱۴/۴

جدول۷: دقت طبقه بندی (روش SVM کوادراتیک و روش SOM)

Table 7. The Accuracy (Quadratic SVM, SOM)

روش	درصد طبقه بندی درست نهایی	درصد طبقه بندی نادرست نهایی
پژوهش حاضر	۹۹/۱	۰/۹
SOM	۸۵/۴	۱۴/۶

در نهایت در جدول ۸، مقایسه‌ای میان روش ارائه شده و دیگر روش‌های ارائه شده در حوزه تشخیص بیماری سرطان سینه بر روی داده‌های مربوط به سرطان سینه از دانشگاه ویسکانسین انجام گرفته است. نگوین و همکاران [۱۶] با استفاده از روش استخراج ویژگی و رأی‌گیری جمعی به‌عنوان یکی از الگوریتم‌های یادگیری ماشین در حوزه طبقه‌بندی داده‌های سرطان سینه به دقت ۹۷/۵۰ درصد رسیدند.

جدول۸: دقت روش‌های طبقه‌بندی در حوزه سرطان سینه

Table8. Accuracy of the other methods in the breast cancer detection field

پژوهشگر	سال	روش	درصد دقت طبقه‌بندی
عثمان و همکاران	۲۰۲۰	یادگیری جمعی	۹۷/۰۰
دامیترو و همکاران	۲۰۲۰	الگوریتم بیضین	۷۴/۲۴
نگوین و همکاران	۲۰۱۹	استخراج ویژگی و رأی‌گیری جمعی	۹۷/۵۰
کائوشیک و همکاران	۲۰۱۶	شبکه عصبی پرسپترون	۸۳/۵۰
مرت و همکاران	۲۰۱۵	RBF-NN	۸۸/۰۰
مرت و همکاران	۲۰۱۵	KNN	۹۱/۰۹
مرت و همکاران	۲۰۱۵	ANN	۹۷/۵۳
مون و همکاران	۲۰۱۳	یادگیری عمیق	۹۳/۴
پژوهش حاضر		SVM کوادراتیک – بدون SU-CFAM	**۹۸/۲**
		SVM کوادراتیک – با SU-CFAM	**۹۹/۱**

در همین حوزه با الگوریتمی دیگر به نام یادگیری جمعی با استفاده از تابع پایه شعاعی (RBF)، عثمان و همکاران [۱۷] به دقت ۹۷/۰۰ درصد دست یافتند.

[1] Self-organizing map (SOM)

همچنین با الگوریتم بیضین[1] دامیترو و همکاران [۱۸] به دقت ۷۴/۲۴ درصد و با الگوریتم یادگیری جمعی با استفاده از روش چندلایه پرسپترون (MLP) ، کائوشیک و همکاران [۱۹] به دقت ۸۳/۵۰ درصد رسیدند. در مطالعه مرت و همکاران [۲۰] نیز که از داده‌های ویسکانسین برای طبقه‌بندی سرطان سینه استفاده کرده‌اند، عملکرد روش‌های KNN، ANN، RBF-NN بررسی شد. روش ANN با دقت ۹۷/۵۳ درصد و روش KNN با دقت ۹۱/۰۰ درصد و نهایتاً روش RBF-NN با دقت ۸۸ درصد، طبقه‌بندی را انجام دادند. همچنین مون و همکاران [۲۱] با استفاده از روش یادگیری عمیق سعی در طبقه‌بندی داده‌های سرطان سینه داشته‌اند که دقت آن روش را به ۹۳/۴۰ درصد رساندند.

همان‌طور که از درصد دقت روش‌های مورد بررسی در پژوهش‌های دیگر مشخص است، این روش‌ها در بهترین حالت در حدود ۹۷/۵۰درصد، و در بدترین حالت در حدود ۷۴/۲۴ درصد، از دقت طبقه‌بندی برخوردار بودند و روش اجرا شده در این پژوهش دارای ۹۸/۲۰ درصد دقت طبقه‌بندی بود و این تمایز نشان از عملکرد مطلوب روش SVM کوادراتیک دارد که علاوه بر غیرخطی بودن دارای محاسبات ساده‌تری نسبت به روش‌های فوق‌الذکر نیز می‌باشد.

۵- بحث و نتیجه‌گیری

سرطان سینه به عنوان عامل اصلی مرگ و میر زنان در کشورهای توسعه یافته تبدیل شده است. مؤثرترین راه برای کاهش مرگ و میر ناشی از سرطان سینه، تشخیص زودهنگام آن است. این موضوع نیاز به یک روش تشخیصی دقیق و قابل اعتماد دارد که به پزشکان اجازه می‌دهد تومورهای خوش‌خیم پستان را از بدخیم تشخیص دهند، بدون اینکه مستقیماً به بیوپسی جراحی مراجعه کنند. هدف از این پیش‌بینی‌ها این است که بیماران را به یک گروه "خوش‌خیم" غیر سرطانی یا یک گروه "بدخیم" که سرطانی هستند تخصیص دهند. ذکر این نکته ضروری است که سلول‌های سرطانی ممکن است به سایر قسمت‌های بدن گسترش یابند. در مورد سرطان سینه، آن‌ها از طریق لنف یا خون منتقل می‌شوند. در سناریوی اخیر، سرطان پستان در مرحله پیشرفته خود در نظر گرفته می‌شود و معمولاً مداخله جراحی به نام بیوپسی مورد نیاز است. به شدت توصیه می‌شود که سرطان سینه قبل از پیشرفت در مراحل پیشرفته مهار شود. در این مقاله با در نظر گرفتن اهمیت تشخیص این سرطان، به ارائه راهکاری با حداقل محاسبات لازم پرداختیم. راهکار ارائه شده در فاز پیش‌پردازش داده‌ها، با بهره‌گیری از یک روش مبتنی بر گراف، سعی در حداقل نمودن ویژگی‌های مؤثر داشت. در ادامه و با بهره‌مندی از ماشین بردار پشتیبان کودراتیک، نسبت به طبقه‌بندی داده‌های ارائه شده در دو گروه خوش‌خیم و بدخیم، اقدام نمودیم. ماشین بردار پشتیبان کودراتیک، به دلیل ماهیت محاسبات پایین، کمک شایانی در راستای کاهش زمان پردازش و کاهش منابع مورد استفاده می‌نماید. دقت روش ارائه شده در مقایسات انجام شده با دیگر روش‌ها، نشان از موفقیت روش ارائه شده داشت. پیشنهاد می‌گردد در ادامه مسیر این مقاله، مقایساتی با دیگر پایگاه داده‌های موجود در این حوزه انجام گیرد. از سوی دیگر، با توجه به انتخاب روش SU-CFAM در فاز انتخاب ویژگی، که یک روش با ماهیت گرافی است و امکان تعمیم به داده‌ها با تعداد ویژگی بسیار بالا را دارد، پیشنهاد می‌گردد مجموعه داده‌های انتخابی در پژوهش‌های آینده داده‌هایی مشابه داده‌های میکروآرایه باشد.

مراجع

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin*, vol. 71, no. 3, 2021, doi: 10.3322/caac.21660.

[2] H. Zerouaoui and A. Idri, "Reviewing machine learning and image processing based decision-making systems for breast cancer imaging," *Journal of Medical Systems*, vol. 45, no. 8, pp. 1-20, 2021, doi: 10.1007/s10916-020-01689-1.

[1] Bayesian classification

[3] S. Ekici and H. Jawzal “Breast cancer diagnosis using thermography and convolutional neural networks,” *Medical hypotheses*, vol. 137, p. 109542, 2020, doi: 10.1016/j.mehy.2019.109542.

[4] S. Sadhukhan, N. Upadhyay and P. Chakraborty, “Breast cancer diagnosis using image processing and machine learning,”. in *Emerging Technology in Modelling and Graphics* , vol. 937, pp. 113-127, 2020, doi: 10.1007/978-981-13-7403-6_12.

[5] P. Sahni and N. Mittal “Breast cancer detection using image processing techniques,” in *Advances in interdisciplinary engineering*, pp. 813-823, 2019, doi: 10.1007/978-981-13-6577-5_79.

[6] M. Adel, A. Kotb, O. Farag, M. S. Darweesh and H. Mostafa, "Breast Cancer Diagnosis Using Image Processing and Machine Learning for Elastography Images," *International Conference on Modern Circuits and Systems Technologies (MOCAST)*, Thessaloniki, Greece, 2019, pp. 1-4, doi: 10.1109/MOCAST.2019.8741846.

[7] S. J. S. Gardezi, A. Elazab, B. Lei and T. Wang “Breast cancer detection and diagnosis using mammographic data: Systematic review,” *Journal of medical Internet research*, vol. 21, no. 7, p. e14464, 2019, doi: 10.2196/14464.

[8] I. Varlamis, I. Apostolakis, D. Sifaki-Pistolla, N. Dey, V. Georgoulias and C. Lionis “Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece,” *Computer methods and programs in biomedicine*, vol. 145, pp. 73-83, 2017, doi: 10.1016/j.cmpb.2017.04.011.

[9] M. Tan, B. Zheng, J. K. Leader and D. Gur, "Association Between Changes in Mammographic Image Features and Risk for Near-Term Breast Cancer Development," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 7, pp. 1719-1728, July 2016, doi: 10.1109/TMI.2016.2527619.

[10] K. Yan *et al.*, "Comprehensive autoencoder for prostate recognition on MR images," *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, Prague, Czech Republic, 2016, pp. 1190-1194, doi: 10.1109/ISBI.2016.7493479.

[11] M. Zhao, A. Wu, J. Song, X. Sun and N. Dong, “Automatic screening of cervical cells using block image processing,” *Biomedical engineering online*, vol. 15, no. 1, pp. 1-20, 2016, doi: 10.1186/s12938-016-0131-z.

[12] K. Sivakami and N. Saraswathi, “Mining big data: breast cancer prediction using DT-SVM hybrid model,” *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, vol. 1, no. 5, pp. 418-429, 2015.

[13] S. Bakhshandeh, R. Azmi and M. Teshnehlab, “Symmetric uncertainty class-feature association map for feature selection in microarray dataset,”. *Int. J. Mach. Learn. and Cyber*., vol. 11, pp. 15–32, 2020, doi: 10.1007/s13042-019-00932-7.

[14] A. Shmilovici, “Support Vector Machines,” in *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 2005, pp. 257-276, doi: 10.1007/0-387-25465-X_12.

[15] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

[16] Q. H. Nguyen et al., "Breast Cancer Prediction using Feature Selection and Ensemble Voting," *International Conference on System Science and Engineering (ICSSE)*, Dong Hoi, Vietnam, 2019, pp. 250-254, doi: 10.1109/ICSSE.2019.8823106.

[17] A. H. Osman and H. M. A. Aljahdali, "An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model," in *IEEE Access*, vol. 8, pp. 39165-39174, 2020, doi: 10.1109/ACCESS.2020.2976149.

[18] D. Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification," *Annals of the University of Craiova*, vol. 36, no. 2, 2009.

[19] D. Kaushik and K. Kaur, "Application of Data Mining for high accuracy prediction of breast tissue biopsy results," *Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC)*, Moscow, Russia, 2016, pp. 40-45, doi: 10.1109/DIPDMWC.2016.7529361.

[20] A. Mert, N. Kılıç, E. Bilgili and A. Akan, "Breast cancer detection with reduced feature set," *Computational and Mathematical Methods in Medicine*, 2015, doi: 10.1155/2015/265138.

[21] W. K. Moon *et al.*, "Classification of breast tumors using elastographic and B-mode features: comparison of automatic selection of representative slice and physician-selected slice of images," in *Ultrasound in medicineand biology*, vol. 39, no. 7, pp. 1147-1157, 2013, doi: 10.1016/j.ultrasmedbio.2013.01.017.