Iranian Journal of Breast Diseases. 2025; 18(3):112-137

Original Article

Breast Cancer Prediction Approach Based on Microarray Data Using Hybrid Gene Selection and Deep Learning

Masoumeh Motevalli[1], Madjid Khalilian[2]✉, Azam Bastanfard[3]

[1]Ph.D. Candidate, Department of Computer Engineering, Ka.C, Islamic Azad University, Karaj, Iran
[2]Department of Computer Engineering, Faculty of Artificial Intelligence, Ka.C, Islamic Azad University, Karaj, Iran
[3]Department of Computer Engineering, Ka.C, Islamic Azad University, Karaj, Iran

Received: 2024/11/11
Accepted: 2025/09/03

***Corresponding Author:**
Khalilian@kiau.ac.ir

Ethics Approval:
Not applicable

Abstract

Introduction: Breast cancer is the most common malignancy among women and the second leading cause of cancer mortality. Gene expression analysis using microarray data reveals molecular patterns associated with disease progression, aiding in diagnosis and treatment. However, the high dimensionality of such data poses significant challenges for machine learning methods.

Materials and Methods: This study presents a hybrid feature selection method that combines filter, wrapper, and deep learning approaches with the Giza Pyramids Construction algorithm (FWGPC) to manage high-dimensional microarray data. The approach enhances classification accuracy and identifies key genes linked to breast cancer.

Results: Experimental results demonstrate that the proposed method achieves classification accuracies of 99.96% and 96.1% on the BC-TCGA and GSE datasets, respectively. It outperforms many classification approaches in identifying new cases of breast cancer.

Conclusion: This study applies the FWGPC algorithm, which combines filter and wrapper approaches, for feature selection in breast cancer microarray data. The best outcomes are evaluated through deep learning, and classification accuracy is assessed. The approach enables key gene identification and improves classification performance.

Keywords: Breast Cancer, Deep Learning, Feature Selection, Giza Pyramids Construction Algorithm, Microarray Data

[DOI: 10.61882/ijbd.18.3.7]

Introduction

Breast cancer is one of the most common malignancies among women, and early diagnosis can significantly improve survival rates (1). Feature selection, by removing noisy and redundant data, enhances the accuracy of machine learning models and reduces computational complexity (2). A hybrid approach that combines filter and wrapper methods leverages the advantages of both strategies, thereby enhancing model performance (3). This study aims to develop a deep learning-based model with hybrid feature selection to identify key genes and provide an accurate tool for breast cancer diagnosis and classification (4).

Materials & Methods

This study proposes a novel hybrid feature selection framework based on the FWGPC algorithm, combining filter- and wrapper-based approaches. The initial solution population is evaluated using distinct fitness functions, and a competitive mechanism selects the best-performing solutions. Convolutional Neural Networks (CNNs) are employed to assess classification accuracy, and the optimal solution guides the generation of new populations to accelerate convergence. The method is validated on two breast cancer-related datasets, demonstrating improved accuracy, efficiency, and stability compared to existing classification techniques.

Results

The proposed hybrid method for breast cancer detection was evaluated on BC-TCGA and GSE DNA microarray datasets (Table 1). Results showed that the algorithm's accuracy steadily improved over iterations (Figure 1), with feature selection guided by the FWGPC algorithm selecting the most relevant genes for deep learning, enhancing classification quality and prediction of new samples (Table 2).

Table 1: Statistical information related to the datasets

Dataset	Number of instances	Number of genes	Number of normal instances	Rate of normal instances	Number of cancer instances	Rate of cancer instances
BC-TCGA	590	17,814	61	0.1034	529	0.8966
GSE	200	10,000	100	0.5	100	0.5

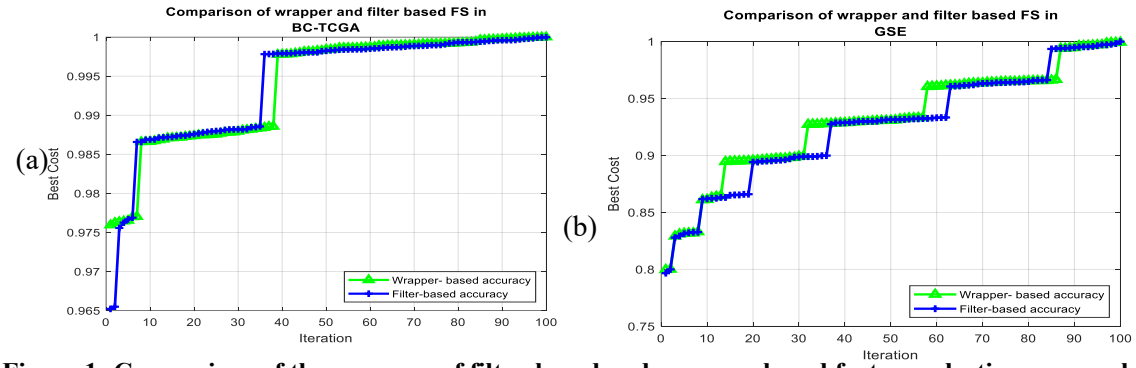
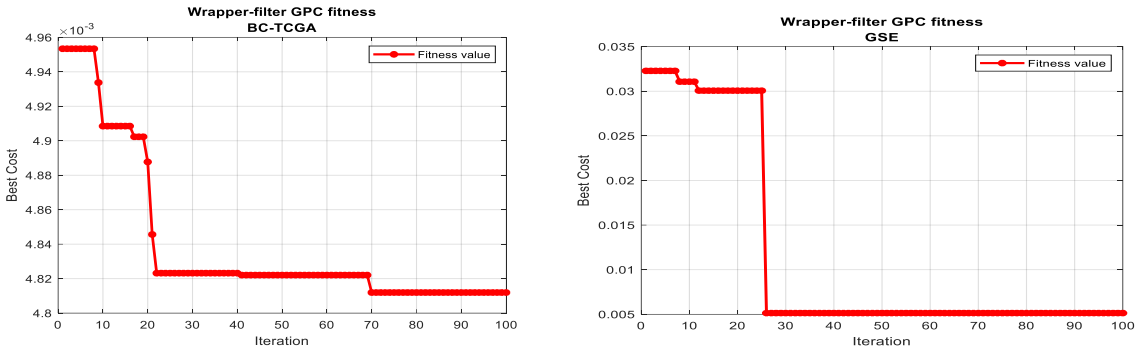


Figure 1: Comparison of the accuracy of filter-based and wrapper-based feature selection approaches based on deep learning (a) BC-TCA (b) GSE

Figure 2 shows that the proposed algorithm steadily converges toward the optimal set of genes. The fitness value stabilizes after approximately 100 iterations, indicating that further iterations do not significantly improve accuracy or reduce error, which highlights the method's efficiency and stability.

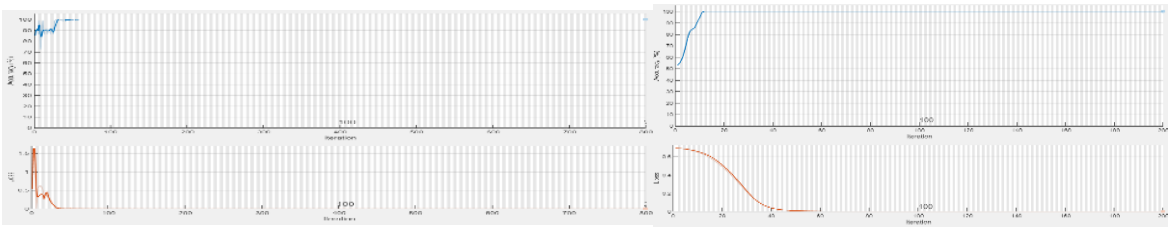
[DOI: 10.61882/ijbd.18.3.7]



(a) (b)

Figure 2: Comparison of the accuracy of filter-based and wrapper-based feature selection approaches based on deep learning (a) BC-TCA (b) GSE

The proposed method combines CNNs with KNN, Naive Bayes, Decision Trees, and Logistic Regression, utilizing genes selected by the GPC algorithm. This multi-method approach improves accuracy in predicting new samples, with selected genes serving as key features for classification (Figure 3).



(a) (b)

Figure 3: Convolutional neural network training process on the dataset (a) BC-TCA (b) GSE

The proposed method's performance was evaluated using a confusion matrix and four metrics: accuracy, sensitivity, specificity, and F1-score. Combining GPC-based gene selection with CNN, KNN, Decision Tree, Naive Bayes, and Logistic Regression significantly improved prediction metrics for breast cancer samples. Average results are shown in **Figure 4** and **Table 2**, highlighting the effectiveness of precise gene selection with the GPC algorithm.

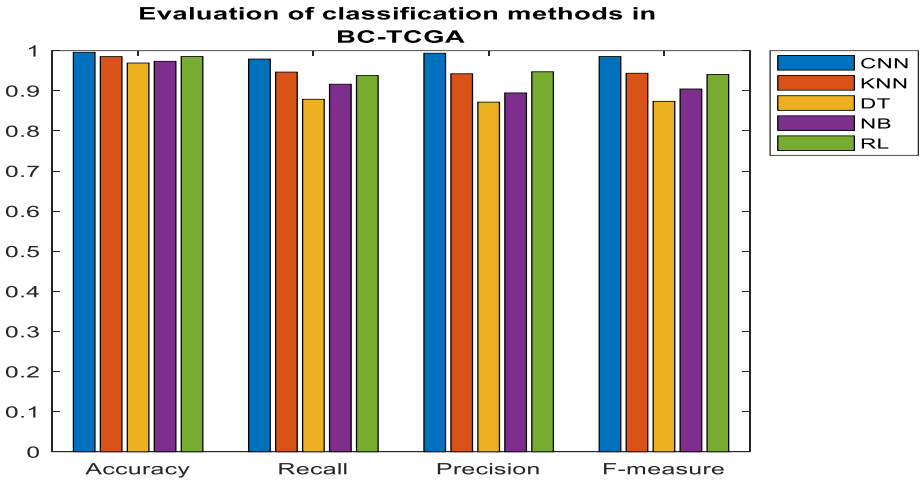


Figure 4: Bar chart comparing the average evaluation criteria for classification methods on the dataset (a) BC-TCA (b) GSE

[DOI: 10.61882/ijbd.18.3.7]

Table 2. Comparison of average values of evaluation criteria for classification methods on datasets

Dataset	Classifier	Accuracy	Recall	Precision	F-measure
BC-TCGA	CNN	0.9996	0.9792	0.9936	0.9854
	KNN	0.9851	0.9466	0.9423	0.9437
	DT	0.9692	0.8788	0.8717	0.8737
	NB	0.9735	0.9164	0.8946	0.9045
	LR	0.9853	0.9378	0.9473	0.9408
GSE	CNN	0.9610	0.9110	0.9555	0.9262
	KNN	0.7073	0.7073	0.8686	0.7806
	DT	0.8481	0.8481	0.8966	0.8564
	NB	0.7134	0.7134	0.8698	0.7813
	LR	0.8293	0.8293	0.8885	0.8470

Discussion

This study proposes a hybrid feature selection and deep learning approach for breast cancer diagnosis, combining filter and wrapper methods optimized via the FWGPC algorithm. CNNs are then used to extract complex patterns and reduce noise, resulting in significantly improved prediction accuracy compared to traditional classifiers.

The proposed method in this study for predicting breast cancer patients employs a hybrid feature selection technique based on filter and wrapper approaches, optimized using the FWGPC algorithm. Key features are extracted from the data and subsequently analyzed using a deep learning model based on CNNs for final prediction (5).

This approach improves accuracy, recall, and Area Under the Curve compared to traditional methods and previous models, enabling the identification of more optimal feature combinations (5). However, high computational complexity, the need for balanced datasets, and the limited extent of algorithm evaluation represent some of the challenges. Compared to prior studies, the combined use of deep learning and hybrid feature selection provides superior performance in predicting patients and identifying relevant genes (5).

Conclusion

This study presents a hybrid FWGPC-CNN approach that enhances breast cancer prediction accuracy, identifies key genes for personalized treatment, and offers potential as a clinical decision-support tool, while highlighting the need for further evaluation on real-world datasets.

References

1. Zhai J, Newton J, Copnell B. Posttraumatic growth experiences and its contextual factors in women with breast cancer: An integrative review. Health Care for Women International. 2019;40(5):554-80. doi:10.1080/07399332.2019.1578360.
2. Salehiniya H, Haghighat S, Parsaeian M, Majdzadeh R, Mansournia M, Nedjat S. Iranian breast cancer risk assessment study (IRBCRAS): a case control study protocol. WCRJ. 2018; 5:1-5. doi: 10.32113/wcrj_20183_1016.
3. Brunet J, Sabiston CM, Burke S. Surviving breast cancer: women's experiences with their changed bodies. Body image. 2013;10(3):344-51. doi:10.1016/j.bodyim.2013.02.002.
4. Kunikata H, Yoshinaga N, Nakajima K. Effect of cognitive behavioral group therapy for recovery of self-esteem on community-living individuals with mental illness: Non-randomized controlled trial. Psychiatry Clin Neurosci. 2016;70(10):457-68. doi: 10.1111/pcn.12418.
5. Kazemzadeh J, Rabeipour S, Rajabzadeh H. Investigating the status of sexual self-esteem and Its related factors in breast cancer survivors. Nursing and Midwifery Journal. 2022;20(1):85-93.
6. Izadi-Ajirlo A, Bahmani B, Ghanbari-Motlagh A. Effectiveness of cognitive behavioral group intervention on body image improving and increasing self-esteem in women with breast cancer after mastectomy. Archives of Rehabilitation. 2013;13(4):72-83.

[DOI: 10.61882/ijbd.18.3.7]

7. Shahsavari H, Matory P, Zare Z, Taleghani F, Kaji MA. Effect of self-care education on the quality of life in patients with breast cancer. Journal of education and health promotion. 2015;4(1):70. doi:10.4103/2277-9531.171782
8. Al Darweesh H, Hadi MA, Al Madani R, Al Mahsen Z. Reviving Nurses' Role as Health Educators; Breast Cancer in a Developing Country. 2016;01(01):001-5.
9. Imani E, Khodami A, Jamhiry R, HoseiniTeshnizi S. Comparison the effect of self-care training program in two methods of multi-media education and tele-nursing on social isolation in patients with COVID-19 in Bandar Abbas. Medical Journal of Tabriz University of Medical Sciences. 2023;45(4):337-52. doi: 10.34172/mj.2023.037.
10. Yang S, Jiang Q, Li H. The role of telenursing in the management of diabetes: a systematic review and meta-analysis. Public Health Nursing. 2019;36(4):575-86. doi:10.1111/phn.12603.
11. Mahinfar K, Sadooghiasl A, Kazemnejad A. The effect of self-care program on quality of life of women with Breast Cancer having mastectomy. Iranian Journal of Nursing Research. 2021;16(2):11-22.
12. Ghanbari E, Yektatalab S, Mehrabi M. Effects of psychoeducational interventions using mobile apps and mobile-based online group discussions on anxiety and self-esteem in women with breast cancer: randomized controlled trial. JMIR mHealth and uHealth. 2021;9(5):e19262. doi:10.2196/19262.
13. Hosseini h, loripoor m, roeintan F. The effect of palliative-care education on quality of life of women with breast cancer. Iranian Journal of Cancer Care. 2022;1(2):31-8. doi:10.1186/s12904-023-01245-x.
14. Ahmadi babadi s, Sadeghmoghadam L, Delshad Noghabi A. Comparing the effectiveness of telenursing with in-person follow up on the feeling of loneliness among the elderly in community health centers in Ahvaz in 2017. Journal of Gerontology. 2017;2(3):58-65.
15. Barbosa Ide A, Silva KC, Silva VA, Silva MJ. The communication process in Telenursing: integrative review. Rev Bras Enferm. 2016;69(4):765-72. doi: 10.1590/0034-7167.2016690421i.

[DOI: 10.61882/ijbd.18.3.7]

مقاله‌پژوهشی

مجله علمی
بیماری‌های پستان ایران
۱۳۷-۱۱۲ : (۳)۱۸؛۱٤۰٤

رویکرد پیش‌بینی سرطان پستان بر اساس داده‌های میکروآرایه با استفاده از انتخاب ژن ترکیبی و یادگیری عمیق

معصومه متولی الموتی[۱]✉، مجید خلیلیان[۲]، اعظم باستان فرد[۳]

[۱]دانشجوی دکتری کامپیوتر، گروه کامپیوتر دانشگاه آزاد اسلامی، واحد کرج
[۲]استادیار، گروه کامپیوتر دانشکده هوش مصنوعی دانشگاه آزاد اسلامی، واحد کرج
[۳]استادیار، گروه کامپیوتر دانشگاه آزاد اسلامی، واحد کرج

تاریخ ارسال: ۱۴۰۳/۰۸/۲۱
تاریخ پذیرش: ۱۴۰۴/۰۶/۱۲

* نویسنده مسئول:
Khalilian@kiau.ac.ir

چکیده

مقدمه: سرطان پستان، شایع‌ترین سرطان و دومین علت مرگ‌ومیر ناشی از سرطان در زنان است و به‌دلیل تکثیر غیرقابل کنترل سلول‌های بافت پستان رخ می‌دهد. بررسی بیان ژن در این سرطان، اطلاعات ارزشمندی درباره رفتار و پیشرفت بیماری ارائه می‌دهد و با استفاده از داده‌های میکروآرایه DNA، محققان می‌توانند الگوهای خاص ژنی را شناسایی کنند که به تشخیص دقیق‌تر و بهبود درمان کمک می‌کند. با این حال، حجم بالای داده‌های میکروآرایه چالش‌های پیچیده‌ای را برای الگوریتم‌های یادگیری ماشین به همراه دارد. برای رفع این مشکل، از روش انتخاب ویژگی برای کاهش ابعاد داده‌ها استفاده می‌شود که علاوه بر افزایش دقت، زمان پردازش را نیز کاهش می‌دهد و در تشخیص و طبقه‌بندی مؤثرتر سرطان مفید است.

روش بررسی: در این تحقیق، هدف انتخاب ژن‌های مرتبط با سرطان پستان در داده‌های میکروآرایه‌ای است که حاوی اطلاعات بسیار پیچیده‌ای از بیان ژن‌ها هستند. برای بهبود دقت و کارایی طبقه‌بندی این داده‌ها، از یک رویکرد ترکیبی استفاده شده است که ترکیب دو روش انتخاب ویژگی فیلتر و لفاف را با الگوریتم ساخت اهرام جیزه چندهدفه (FWGPC) و یادگیری عمیق به‌کار می‌گیرد.

یافته‌ها: نتایج تجربی نشان می‌دهد که این روش بر روی مجموعه داده‌های مختلف (BC-TCGA، GSE) به ترتیب دقتی برابر با ۹۹/۹۶٪ و ۹۶/۱٪ را به‌دست آورده و از بسیاری از روش‌های طبقه‌بندی دیگر در تشخیص نمونه‌های جدید سرطان پستان عملکرد بهتری دارد.

نتیجه‌گیری: در این رویکرد، جمعیت اولیه الگوریتم GPC به دو گروه تقسیم می‌شود: یکی براساس تابع تناسب مبتنی بر روش فیلتر و دیگری براساس تابع تناسب مبتنی بر روش لفاف. هر بخش ویژگی‌های خود را براساس معیارهای خاص ارزیابی می‌کند و در نهایت، یک مسابقه بین بهترین نتایج دو بخش برگزار می‌شود. یادگیری عمیق با بررسی دقت طبقه‌بندی، برنده نهایی را اعلام می‌کند که بهترین دقت را در طبقه‌بندی داده‌های جدید سرطان پستان دارد.

کلید واژه‌ها: سرطان پستان، داده‌های میکروآرایه، انتخاب ویژگی، الگوریتم سخت اهرام جیزه، یادگیری عمیق

[DOI: 10.61882/ijbd.18.3.7]

مقدمه

سرطان پستان (BC)[1] یکی از بیماری‌های شایع در میان زنان است که سالانه میلیون‌ها نفر را در سراسر جهان مبتلا می‌کند و با وجود پیشرفت‌های پزشکی، هنوز درمان قطعی برای آن یافت نشده است[۲،۱]. تشخیص زودهنگام و دقیق این بیماری می‌تواند شانس بقا را به میزان قابل‌توجهی افزایش دهد. در سال‌های اخیر، روش‌های خودکار مبتنی بر تصویربرداری پزشکی برای تشخیص سرطان پستان توسعه یافته‌اند، اما چالش‌هایی مانند کیفیت تصاویر، کمبود داده و پیچیدگی ذاتی این بیماری، دقت این روش‌ها را محدود کرده است[۴،۳]. یکی از پیشرفت‌های مهم در حوزه ژنتیک، فناوری بیان ژن و استفاده از میکروآرایه‌هاست که امکان بررسی هم‌زمان هزاران ژن را فراهم می‌کند و به شناسایی الگوهای مرتبط با بیماری کمک می‌نماید[۵]. با این حال، داده‌های میکروآرایه دارای ابعاد بسیار بالا و شامل ویژگی‌های نامربوط، نویزی و تکراری هستند که می‌توانند دقت روش‌های یادگیری ماشین و داده‌کاوی را کاهش دهند. این چالش‌ها باعث افزایش پیچیدگی تحلیل داده‌ها شده و الگوریتم‌های هوش مصنوعی را با مشکلات محاسباتی مواجه می‌کنند، که نیاز به بهینه‌سازی و انتخاب ویژگی‌های مرتبط را ضروری می‌سازد [۷،۶].

برای بهبود دقت مدل‌های یادگیری ماشین و کاهش پیچیدگی محاسباتی در تحلیل داده‌های پزشکی، از روش‌های انتخاب ویژگی استفاده می‌شود. این روش‌ها با حذف ویژگی‌های نامربوط و کم‌اهمیت، اطلاعات مفید و مرتبط با بیماری را استخراج کرده و موجب افزایش دقت مدل‌ها می‌شوند[۸]. انتخاب ویژگی نه‌تنها به درک بهتر عوامل ژنتیکی بیماری‌ها کمک می‌کند، بلکه باعث توسعه روش‌های تشخیص و درمان مؤثرتر نیز می‌شود. هدف این فرآیند یافتن کوچک‌ترین مجموعه ویژگی‌هایی است که بدون از دست دادن اطلاعات کلیدی، دقت مدل را حفظ کرده و خطای تعمیم را کاهش دهد[۹]. در این راستا، ویژگی‌های غیرضروری و نویزی حذف می‌شوند تا مدل بتواند سریع‌تر و کارآمدتر عمل کند و عملکرد بهتری روی داده‌های جدید داشته باشد[۱۰]. دو رویکرد اصلی در انتخاب ویژگی شامل روش‌های مبتنی بر فیلتر و مبتنی بر لفاف است. روش‌های فیلتر با استفاده از معیارهای آماری، ویژگی‌های مرتبط را مستقل از الگوریتم یادگیری انتخاب می‌کنند و به دلیل سرعت بالای پردازش، برای مجموعه داده‌های بزرگ مناسب هستند[۱۱]. در مقابل، روش‌های لفاف ویژگی‌ها را بر اساس عملکرد مدل یادگیری ماشین ارزیابی کرده و دقت بالاتری دارند، اما به دلیل پیچیدگی محاسباتی، زمان‌برتر هستند. برای بهره‌گیری از مزایای هر دو روش، رویکرد ترکیبی مورد استفاده قرار می‌گیرد که ابتدا با فیلتر کردن ویژگی‌های نامرتبط، حجم داده‌ها را کاهش داده و سپس با روش لفاف، ویژگی‌های منتخب را بهینه‌سازی می‌کند[۱۲]. در داده‌های پیچیده و پرحجم مانند میکروآرایه‌ها، انتخاب ویژگی ترکیبی با حذف ژن‌های غیرمرتبط و نویزی، دقت مدل‌های یادگیری ماشین را افزایش داده و به ارائه نتایج تشخیصی قابل‌اعتمادتر کمک می‌کند[۱۳].

هدف از این پژوهش، توسعه یک ابزار تشخیصی نوین و دقیق برای کمک به پزشکان و کادر درمان در شناسایی زودهنگام و طبقه‌بندی بیماران مبتلا به سرطان پستان است. با توجه به پیچیدگی داده‌های ژنتیکی، روش پیشنهادی ما با گزینش هوشمندانه ژن‌های کلیدی مرتبط با بیماری از میان حجم وسیع اطلاعات میکروآرایه، یک مدل پیش‌بینی‌کننده قدرتمند مبتنی بر یادگیری عمیق ایجاد می‌کند. این مدل قادر خواهد بود با بررسی پروفایل ژنی بیماران، الگوهای مولکولی مرتبط با سرطان پستان را شناسایی کرده و به تشخیص سریع‌تر و دقیق‌تر موارد مشکوک کمک نماید. در نهایت، این رویکرد می‌تواند به اتخاذ تصمیمات درمانی آگاهانه‌تر و شخصی‌سازی‌شده برای بیماران منجر شده و در بهبود نتایج درمان و کیفیت زندگی آن‌ها نقش مؤثری ایفا کند.

ادبیات و پیشینه پژوهش

با توجه به اهمیت مسئله کشف و پیش‌بینی بیماران مبتلا به سرطان پستان، محققان زیادی در این زمینه تلاش کرده‌اند.

در [۱۴]، داده‌های میکروآرایه مرتبط مسئله از مجموعه داده مختلف به‌دست آمده است. سپس نمونه‌های دورافتاده با استفاده از تکنیک میانگین چندگانه مقاوم[2] حذف می‌شوند. این تکنیک علاوه بر کنترل کیفیت و حذف نقاط

[1] Breast Cancer

[2] Robust Multi-array Average

[DOI: 10.61882/ijbd.18.3.7]

نامطلوب، نرمال‌سازی داده‌ها و تصحیح پس‌زمینه را انجام می‌دهد. پس از نرمال‌سازی، انتخاب ژن‌ها با استفاده از الگوریتم انتخاب ویژگی مبتنی بر مجموعه‌های خشن انجام می‌شود. سپس، دسته‌بند ترکیبی شامل kNN و SVM برای طبقه‌بندی ژن‌های پرخطر (TNBC[1]) و کم‌خطر (غیر TNBC) اعمال می‌شود. این دسته‌بند با رأی‌گیری نرم، که احتمال دسته‌بندی kNN و SVM را تجمیع می‌کند، نتایج بهتری نسبت به دسته‌بندهای تکی ارائه می‌دهد. در نهایت، مدل پیشنهادی براساس معیارهایی همچون MCC، دقت، یادآوری و F-مقیاس ارزیابی می‌شود.

در روش منبع [۱۴]، تمرکز اصلی بر پیش‌پردازش داده‌های میکروآرایه‌ای با حذف نمونه‌های پرت و استفاده از تکنیک‌های نرمال‌سازی و انتخاب ویژگی با مجموعه‌های خشن است. همچنین، ترکیب دسته‌بندهای kNN و SVMبا رأی‌گیری نرم برای بهبود دقت طبقه‌بندی استفاده شده است. در مقابل، روش پیشنهادی با بهره‌گیری از یک الگوریتم تکاملی چندهدفه (FWGPC) و یادگیری عمیق، فرایند انتخاب ویژگی را از دو دیدگاه فیلتر و لفاف به‌صورت همزمان بررسی می‌کند و بهترین ویژگی‌ها را با کمک یک رقابت میان دو مسیر و ارزیابی نهایی توسط شبکه‌های عمیق انتخاب می‌کند. در نتیجه، روش پیشنهادی دقت بالاتری را در طبقه‌بندی داده‌های جدید (تا ۹۹/۹۶٪) ارائه می‌دهد، در حالی‌که این روش عمدتاً بر کاهش نویز و ترکیب ساده دسته‌بندها تکیه دارد.

در [۱۵]، یک مدل یادگیری عمیق ترکیبی مبتنی بر شبکه عصبی کانولوشنال لاپلاسین (LS-CNN[2]) برای طبقه‌بندی داده‌های سرطان مورد استفاده قرار گرفته است. در [۱۵] از یک مدل CNN بهبودیافته (LS-CNN) برای استخراج خودکار ویژگی‌ها از داده‌های سرطان استفاده شده است که قدرت یادگیری ویژگی‌های محلی را دارد. در حالی‌که روش پیشنهادی نیز از یادگیری عمیق بهره می‌برد، اما آن را با انتخاب ویژگی پیشرفته ترکیب کرده و از الگوریتم تکاملی FWGPC برای کاوش همزمان در فضای ویژگی‌ها استفاده می‌کند. بنابراین، روش پیشنهادی علاوه‌بر یادگیری ساختارهای پیچیده، قدرت انتخاب ویژگی‌های بهینه را نیز دارد که باعث بهبود عملکرد نسبت به مدل صرفاً مبتنی بر CNN می‌شود.

در [۱۶]، بهینه‌سازی ازدحام ذرات و یک روش یادگیری گروهی برای انتخاب ویژگی و طبقه‌بندی سرطان با یکدیگر همکاری می‌کنند.

روش [۱۶]، از ترکیب بهینه‌سازی ازدحام ذرات و یادگیری گروهی برای انتخاب ویژگی استفاده می‌کند که نشان‌دهنده تلاش برای همگرایی بهتر به مجموعه ویژگی‌های مؤثر است. با این حال، روش پیشنهادی با تقسیم جمعیت اولیه به دو مسیر مستقل فیلتر و لفاف و رقابت میان آن‌ها، کاوش چندجانبه‌تری را انجام داده و در نهایت با بهره‌گیری از یادگیری عمیق، انتخاب نهایی را انجام می‌دهد. این ساختار ترکیبی چندمرحله‌ای، جامع‌تر از ساختار صرفاً مبتنی بر PSO عمل می‌کند.

در [۱۷]، یک الگوریتم جستجوی فاخته چند هدفه ترکیبی را ارائه می‌کند که با اپراتورهای تکاملی، به ویژه جهش دوگانه و متقاطع منفرد، با هدف افزایش دقت انتخاب ژن ارائه شده است. این تحقیق بر بهبود کارایی انتخاب ژن‌های اطلاعاتی از داده‌های ریزآرایه سرطان با ابعاد بالا تمرکز دارد، در نتیجه مقادیر ابعاد الگوریتم و قابلیت‌های اکتشاف را افزایش می‌دهد.

در [۱۷]، یک الگوریتم جستجوی فاخته چندهدفه با اپراتورهای تکاملی خاص برای افزایش دقت انتخاب ژن‌ها به کار رفته است. این روش از نظر اکتشاف فضای ویژگی‌ها قوی است، اما در مرحله طبقه‌بندی یا تصمیم‌گیری نهایی وابستگی کمتری به یادگیری عمیق دارد. در مقابل، روش پیشنهادی با بهره‌گیری از یادگیری عمیق و رقابت دوطرفه، هم دقت انتخاب ویژگی را تضمین می‌کند و هم قدرت طبقه‌بندی را افزایش می‌دهد.

در [۱۸]، یک رویکرد طبقه‌بندی سرطان ترکیبی معرفی شده است که از تکنیک‌های مختلف یادگیری ماشینی استفاده می‌کند. از ضریب همبستگی پیرسون برای انتخاب و کاهش ویژگی، یک طبقه‌بندی درخت تصمیم برای تفسیرپذیری و عدم نیاز به پارامتر، و Grid Search CV برای بهینه‌سازی فراپارامتر حداکثر عمق استفاده می‌کند.

در [۱۸] از تکنیک‌های یادگیری ماشین کلاسیک مانند درخت تصمیم و Grid Search برای بهینه‌سازی پارامترها استفاده شده است که منجر به مدلی تفسیرپذیر و با تنظیمات خودکار می‌شود. اما این روش بیشتر بر

[1] triple negative breast cancer

[2] Laplacian Score-Convolutional Neural Network

[DOI: 10.61882/ijbd.18.3.7]

ساده‌سازی و تفسیر تمرکز دارد، در حالی‌که روش پیشنهادی بر دقت حداکثری و کاوش ترکیبی در فضای ویژگی‌ها متمرکز است. لذا از نظر دقت، روش پیشنهادی عملکرد بالاتری دارد، اما از نظر سادگی و تفسیرپذیری، این روش ممکن است برتری داشته باشد.

در [۱۹]، دو روش انتخاب ویژگی یعنی Boruta و LASSO و ماشین بردار پشتیبان (SVM[1]) و طبقه‌بندی رگرسیون لجستیک (LR[2]) مورد مطالعه قرار گرفته‌اند. یک مجموعه داده سرطان پستان از وب GEO در این مطالعه اتخاذ شده است.

در [۱۹]، از Boruta و LASSO به‌عنوان تکنیک‌های انتخاب ویژگی و SVM و رگرسیون لجستیک به‌عنوان دسته‌بند استفاده شده است. این ترکیب در چارچوب کلاسیک یادگیری ماشین انجام می‌شود و دقت خوبی دارد، اما انعطاف و قدرت کاوش الگوریتم FWGPC همراه با یادگیری عمیق در روش پیشنهادی می‌تواند ویژگی‌های پیچیده‌تری را استخراج کرده و دقت بالاتری حاصل کند، به‌ویژه برای داده‌هایی با ابعاد بالا و ویژگی‌های غیرخطی.

در [۲۰]، ۷۶۲ بیمار مبتلا به سرطان پستان و ۱۳۸ فرد سالم مورد بررسی قرار گرفتند و از سه گروه الگوریتم‌های یادگیری ماشین برای انتخاب و استخراج ویژگی‌ها و همچنین ۱۳ الگوریتم طبقه‌بندی با تنظیمات خودکار استفاده شد تا عملکرد مدل‌ها با دقت متوازن و مساحت زیر منحنی ارزیابی شود.

در [۲۰]، از مجموعه‌ای گسترده از الگوریتم‌های یادگیری ماشین و ویژگی‌های آماری برای مدل‌سازی استفاده شده است که یک رویکرد جامع و ترکیبی به‌شمار می‌رود. با این‌حال، نبود تمرکز خاص بر بهینه‌سازی انتخاب ویژگی یا استفاده از الگوریتم‌های تکاملی باعث می‌شود روش پیشنهادی، به‌خصوص در انتخاب ویژگی‌های کلیدی با کمک FWGPC و یادگیری عمیق، دقت و بازده بالاتری در تشخیص سرطان داشته باشد.

در [۲۱]، یک چارچوب برای انتخاب ژن و طبقه‌بندی سرطان ارائه شده است که از یک خودرمزگذار و یکی از نه طبقه‌بند استفاده می‌کند که این مدل با کاهش تدریجی تعداد ژن‌ها، بهترین مدل را در طبقه‌بندی سرطان شناسایی کرده و در مقایسه با هفت مدل دیگر در تمامی مجموعه داده‌ها عملکرد بهتری دارد، در حالی که اهمیت بیولوژیکی نتایج نیز مورد بررسی قرار گرفته است.

در [۲۱]، از خودرمزگذار برای کاهش تدریجی تعداد ژن‌ها و یافتن ساختارهای پنهان استفاده شده و سپس با دسته‌بندهای مختلف آزمایش شده است. اگرچه این روش از نظر استخراج ویژگی‌های فشرده و مهم موثر است، اما روش پیشنهادی با رویکرد رقابتی میان دو مسیر انتخاب ویژگی و ارزیابی نهایی با یادگیری عمیق، نتایج دقیق‌تری ارائه می‌دهد و همچنین کنترل بیشتری بر کیفیت انتخاب ویژگی‌ها دارد.

در[۲۲]، روشی جدید بدون اطلاعات قبلی برای شناسایی ژن‌های جهش‌یافته مرتبط با سرطان پستان پیشنهاد می‌شود که شامل پردازش داده‌های جهش somatic به یک ماتریس جهش‌یافته، فیلتر کردن مجموعه‌ای از ژن‌های جهش‌یافته بر اساس فراوانی جهش، و استفاده از الگوریتم انتخاب ویژگی مرحله‌ای با بهینه‌سازی بیزی برای ایجاد یک مدل بهینه و ارزیابی با یک معیار وزنی برای حل مشکل عدم تعادل نمونه‌ها است.

در [۲۲] از داده‌های جهش‌یافته و فیلتر کردن بر اساس فراوانی و انتخاب ویژگی مرحله‌ای با بهینه‌سازی بیزی استفاده شده است که بیشتر بر ژن‌های خاص با اطلاعات قبلی تمرکز دارد. در مقابل، روش پیشنهادی نیازی به اطلاعات قبلی ندارد و با جستجوی هوشمند در فضای ویژگی‌ها و ارزیابی دقیق، ویژگی‌های مؤثر را از داده‌های گسترده استخراج می‌کند.

در [۲۳]، مدلی کارآمد از XGBoost که با الگوریتم جستجوی شبکه بهینه‌سازی شده است، برای شناسایی کمکی تومورهای متاستاتیک پستان بر اساس بیان ژن‌ها ایجاد شده که با اعتبارسنجی ده‌برابری، میانگین بالاتر مساحت زیر منحنی برابر ۰/۸۲ را نسبت به دیگر دسته‌بندها دارد. در ، این مقاله یک مدل CNN الهام گرفته از بیولوژی برای تشخیص سرطان پستان با استفاده از داده‌های بیان ژن دانلود شده از پایگاه داده سرطان ژنوم (TCGA3) ارائه می‌دهد که شامل ۱۲۰۸ نمونه بالینی و با استفاده از الگوریتم جستجوی بهینه‌سازی ابولا (EOSA[4])، برای بهبود دقت تشخیص طراحی شده است.

[1] Support Vector Machine
[2] Logistic Regression

[3] cancer genome atlas
[4] Ebola Optimization Search Algorithm

[DOI: 10.61882/ijbd.18.3.7]

الگوریتم SCACSA در [۲۵] به‌عنوان روشی جدید برای انتخاب ویژگی به‌کار رفته که با SVM ترکیب شده است. اگرچه این الگوریتم تکاملی نوآورانه است، اما برخلاف روش پیشنهادی که از دو مسیر موازی انتخاب ویژگی و یادگیری عمیق برای تأیید نهایی بهره می‌برد، فاقد ارزیابی چندمرحله‌ای و تأیید توسط شبکه‌های عصبی است. به همین دلیل، دقت نهایی و پایداری مدل پیشنهادی بالاتر ارزیابی می‌شود.

در جدول ۱ مزایا و معایب هر یک از روش‌های مطرح شده برای کشف و پیش‌بینی بیماران مبتلا به سرطان پستان آورده شده است.

بر اساس کارهای مرتبط بررسی شده، بدیهی است که مطالعات کمتری از رویکرد انتخاب ویژگی ترکیبی تعاملی فیلتر- لفاف استفاده کرده است. اکثر روش‌ها سعی می‌کنند ویژگی‌های تقریباً بهینه را بدون در نظر گرفتن پیچیدگی الگوریتم‌های فراابتکاری انتخاب کنند. با این حال، کاهش ابعاد ویژگی‌های نامربوط می‌تواند به جلوگیری از تحمیل پیچیدگی اضافی بر روش‌های انتخاب ویژگی، و در نهایت بر روی روش‌های طبقه‌بندی مورد استفاده برای پیش‌بینی بیماران سرطانی کمک کند[۲۶].

ادامه مقاله به شرح زیر سازماندهی شده است. در بخش دوم مواد و روش‌ها بررسی خواهد شد. در بخش سوم یافته‌های روش پیشنهادی بیان خواهد شد. در بخش چهارم بحث در مورد روش پیشنهادی بیان خواهد شد. در بخش پنجم نتیجه گیری این مقاله بیان خواهد شد.

مدل [۲۳] از XGBoost با تنظیم پارامتر از طریق جستجوی شبکه استفاده می‌کند و دقت خوبی با مساحت زیر منحنی بالا فراهم می‌آورد. اما الگوریتم‌های درختی مانند XGBoost با وجود سرعت بالا، در برخی موارد از نظر تفسیر روابط پیچیده و تعامل ویژگی‌ها محدودتر هستند. روش پیشنهادی با ساختار چندمرحله‌ای و بررسی عمیق‌تری از ویژگی‌ها، قادر به استخراج ویژگی‌های معنادارتر و ارائه دقت بسیار بالاتری است.

در [۲۴]، یک چارچوب جدید انتخاب ویژگی ترکیبی شامل حداکثر ارتباط-حداقل تکرار (mRMR[1]) و آزمون t دو طرفه را برای شناسایی بهینه‌ترین نشانگرهای ژنی برای پیش‌بینی سرطان پستان پیشنهاد شده است که شامل سه نشانگر ژنی MAPK1، APOBEC3B و ENAH است و برای ارزیابی قابلیت پیش‌بینی این نشانگرها از الگوریتم‌های یادگیری ماشین پیشرفته استفاده شده است.

در [۲۴] رویکرد mRMR و آزمون t دو طرفه برای انتخاب نشانگرهای ژنی خاص به کار گرفته شده که به یافتن نشانگرهای بیولوژیکی مهم کمک می‌کند. اگرچه این روش از نظر تفسیر زیستی قوی است، اما از نظر مدل‌سازی دقیق و تطبیق با یادگیری عمیق و جستجوی چندهدفه، نسبت به روش پیشنهادی در سطح پایین‌تری قرار دارد.

در [۲۵]، یک روش جدید به نام الگوریتم جستجوی سینوس-کسینوس و ککوی (SCACSA) برای انتخاب ژن ارائه شده که با دسته‌بندهای یادگیری ماشین معروف مانند SVM کار می‌کند و عملکرد آن با استفاده از یک مجموعه داده مربوط به سرطان پستان ارزیابی و مقایسه شده است.

جدول ۱: مزایا و معایب روش‌های پیشین

Table 1: Advantages and disadvantages of previous methods

Ref.	Method	Advantages	Disadvantages
[14]	Multiple Mean Technique	Control quality, remove noise, normalize data	May require large and complex processing for large datasets
[15]	LS-CNN	Deep learning capability and feature extraction	Requires large datasets and long processing times

[1] minimum Redundancy-Maximum Relevance

[DOI: 10.61882/ijbd.18.3.7]

Ref.	Method	Advantages	Disadvantages
[16]	Particle Swarm Optimization	Improved feature selection and categorization	May not converge to the best solution
[17]	Multi-objective Optimization	Increased complexity with the use of operators	High complexity and requires precise parameter tuning
[18]	Decision Tree Ensemble Methods	Interpretability using decision trees	May require precise parameter tuning
[19]	Boruta and LASSO Methods	Optimal feature selection and weight reduction	May be unreliable in some cases
[20]	Adaptive Algorithms	High diversity in feature selection and algorithms	Automatic configurations may reduce usability
[21]	Framework	Identifying the best model in reducing overfitting	Complex interactions between models may lead to overfitting
[22]	Gender Recognition	Processing all aspects using algorithms	Requires accurate and extensive datasets
[23]	XGBoost Model	High AREA UNDER THE CURVE compared to other algorithms	May require longer prediction times
[24]	mRMR and t-test Ensemble	Identifying optimal features using statistical techniques	May affect selection accuracy under noisy data
[25]	SCASCA	Improved feature selection and SVM performance	Complex method may require prior knowledge

مواد و روش‌ها

در این مقاله، رویکردی ترکیبی از روش‌های فیلتر و لفاف برای انتخاب ویژگی در تشخیص سرطان پستان معرفی شده است. این روش از الگوریتم بهینه‌سازی چندهدفه اهرام جیزه به‌عنوان چارچوب اصلی بهره می‌برد و شبکه‌های عصبی کانولوشنی را برای ارزیابی عملکرد استفاده می‌کند. ابتدا جمعیت اولیه‌ای از راه‌حل‌های بالقوه به دو بخش تقسیم می‌شود: بخشی که با استفاده از الگوریتم اهرام جیزه مبتنی بر فیلتر ویژگی‌ها را انتخاب می‌کند، و بخش دیگری که الگوریتم اهرام جیزه را برای روش لفاف به کار می‌برد. هر کدام از این بخش‌ها تابع تناسب مخصوص به خود را دارند که راه‌حل‌های اولیه را ارزیابی می‌کند و سپس بهترین راه‌حل‌های موجود انتخاب می‌شوند.

در مرحله بعد، دو راه‌حل منتخب به‌عنوان ورودی برای شبکه عصبی کانولوشنی (CNN) در نظر گرفته می‌شوند ۲۵/۹/۲۰۲۵. شبکه عصبی عمیق بر روی این دو مجموعه داده اعمال می‌شود و دقت طبقه‌بندی برای هر یک از راه‌حل‌ها محاسبه و مقایسه می‌شود. اگر روش لفاف دقت بالاتری در طبقه‌بندی بیماران داشته باشد، نسل جدید از روی این روش ایجاد می‌شود؛ در غیر این صورت، نسل جدید بر اساس راه‌حل روش فیلتر تولید می‌شود. این فرآیند تکرار می‌شود و در هر مرحله نسل جدیدی ایجاد می‌گردد، که ترکیبی از راه‌حل‌های دو روش است و به تدریج به یک مجموعه بهینه و متعادل از ژن‌های انتخاب‌شده منجر می‌شود.

در نهایت، این فرآیند تکرار می‌شود تا شرط توقف، که در این روش به ۱۰۰ بار تکرار الگوریتم تنظیم شده، محقق شود[۲۸].

[DOI: 10.61882/ijbd.18.3.7]

در آخرین مرحله، از میان دو راه‌حل موجود، بهترین راه‌حل به عنوان مجموعه ژن‌های نهایی انتخاب می‌شود. این روش ترکیبی نه تنها امکان بهره‌مندی از مزایای هر دو روش فیلتر و لفاف را فراهم می‌کند، بلکه الگوریتم ساخت اهرام جیزه به بهبود دقت و پایداری مدل کمک می‌کند و فرآیند انتخاب ژن‌های کلیدی برای تشخیص سرطان پستان را به‌طور مؤثری تسهیل می‌نماید. مشارکت‌های کلیدی این پژوهش به شرح زیرمی‌باشند:

۱. **ارائه یک چارچوب نوین انتخاب ویژگی** بر پایه الگوریتم بهینه‌سازی الهام‌گرفته از ساخت اهرام جیزه (Giza Pyramids Construction Optimization – FWGPC) در قالب یک روش فیلتر محور (filter-based).

۲. **طراحی یک رویکرد تعاملی دومرحله‌ای** که در آن جمعیت اولیه‌ی راه‌حل‌ها بین دو رویکرد فیلتر و لفاف (wrapper-based) تقسیم شده و از طریق یک مکانیسم رقابتی، به بهبود عملکرد کلی انتخاب ویژگی کمک می‌شود.

۳. **به‌کارگیری توابع تناسب (fitness functions)** متفاوت برای هر یک از دو رویکرد فیلتر و لفاف به‌منظور بهینه‌سازی بهتر ویژگی‌ها متناسب با اهداف خاص هر رویکرد.

۴. **استفاده از شبکه‌های عصبی کانولوشن (CNNs)** جهت ارزیابی و طبقه‌بندی راه‌حل‌های تولیدشده توسط هر دو رویکرد، و تحلیل دقت طبقه‌بندی آن‌ها.

۵. **انتخاب راه‌حل بهینه نهایی** با بالاترین دقت طبقه‌بندی به‌عنوان مرجع (reference solution) برای تولید نسل جدیدی از راه‌حل‌ها در هر دو رویکرد، با هدف تسریع همگرایی و افزایش بهره‌وری الگوریتم.

۶. **ارزیابی عملکرد روش پیشنهادی** از طریق پیاده‌سازی آن بر روی دو مجموعه داده مرتبط با سرطان پستان و مقایسه آن با روش‌های طبقه‌بندی مختلف به‌منظور اعتبارسنجی و تحلیل دقت، کارایی و پایداری مدل.

در روش پیشنهادی، انتخاب ویژگی‌ها با استفاده از یک رویکرد ترکیبی و تعاملی انجام می‌شود که ترکیبی از دو روش انتخاب ویژگی مبتنی بر فیلتر و انتخاب ویژگی مبتنی بر Wrapper است. این ترکیب با بهره‌گیری از الگوریتم چندهدفه ساخت اهرام جیزه (MOGPC[1]) انجام می‌گیرد و در آن، یادگیری عمیق نقش هسته اصلی انتخاب ویژگی‌ها را ایفا می‌کند. با توجه به معماری روش پیشنهادی که در شکل ۱ نشان داده شده است. ابتدا، داده‌ها از مجموعه داده بیماران سرطان پستان ستخراج شده به جمعیت اولیه الگوریتم GPC تبدیل شده است. سپس این جمعیت اولیه به دو بخش تقسیم می‌شود؛ یک بخش جهت انتخاب ویژگی‌های مبتنی بر فیلتر و بخش دیگر برای روش مبتنی بر Wrapper اختصاص داده می‌شود. هر یک از این روش‌ها دارای تابع تناسب خاص خود هستند که برای ارزیابی ویژگی‌ها استفاده می‌شوند. به این ترتیب، جمعیت اولیه به‌طور جداگانه در هر دو روش مورد ارزیابی قرار می‌گیرد و سپس، بهترین ویژگی‌ها بر اساس نتایج هر دو روش انتخاب می‌شوند. در واقع، نیمی از جمعیت اولیه از طریق فیلتر و نیمه دیگر با استفاده از روش Wrapper تحلیل می‌شود.

روش‌های انتخاب زیرمجموعه ویژگی‌ها از الگوریتم چندهدفه ساخت اهرام جیزه (MOGPC) و یادگیری عمیق استفاده می‌کند[۲۸]. این الگوریتم بر روی داده‌های آموزشی و تست مجموعه داده‌های بیماران مبتلا به سرطان پستان[2] اعمال می‌شود تا بتوان الگوهای تشخیص و شناسایی سرطان پستان را به‌دقت تعیین کرد. الگوریتم ساخت اهرام جیزه با ایجاد یک جمعیت اولیه از راه‌حل‌های مختلف، فرآیند بهینه‌سازی را آغاز می‌کند. هر راه‌حل اولیه بر اساس تابع تناسب ارزیابی شده و میزان بهینگی آن تعیین می‌شود. سپس، این راه‌حل‌ها به ترتیب بهینگی مرتب می‌شوند و در مرحله اکتشاف، راه‌حل‌های بهینه انتخاب می‌گردند. راه‌حلی که بهترین تابع تناسب را دارد، به‌عنوان راه‌حل غالب انتخاب می‌شود و به‌عنوان مرجعی برای جهت‌دهی به نسل‌های بعدی مورد استفاده قرار می‌گیرد. با حذف راه‌حل‌های نامناسب و ایجاد جمعیتی جدید که به سمت راه‌حل‌های بهینه گرایش دارد، فرآیند بهبود تدریجی انجام می‌شود. این مراحل تا زمانی که شرط توقف تحقیق حاصل شود، تکرار می‌گردند.

[1] Multi-Objective Giza Pyramids Construction

[2] https://data.mendeley.com/datasets/v3cc2p38hb/1

[DOI: 10.61882/ijbd.18.3.7]

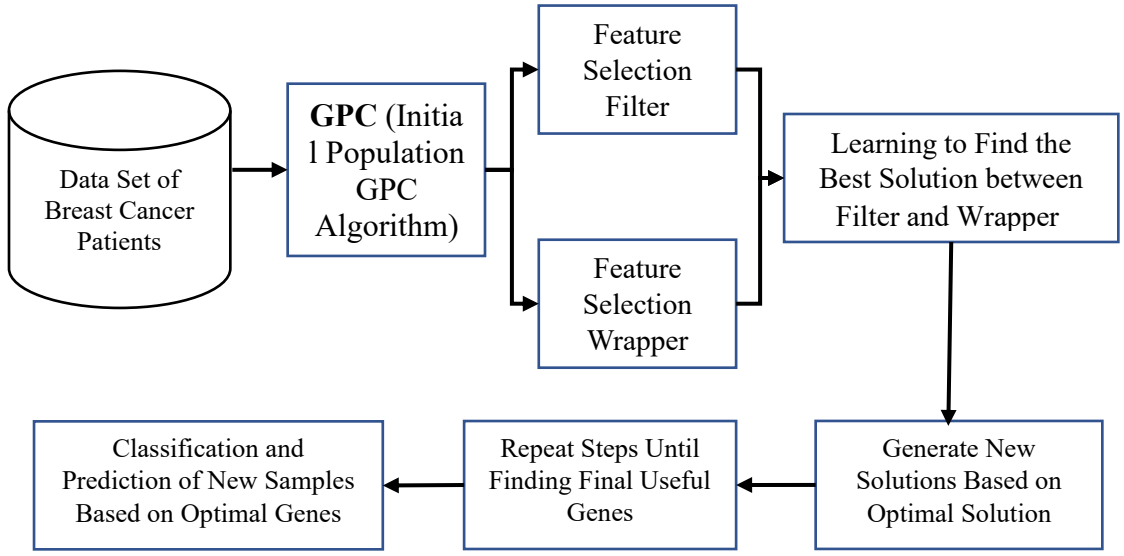


Figure 1: Architecture of the proposed method

شکل ۱: معماری روش پیشنهادی

در روش پیشنهادی، اهداف ارزیابی به دو دسته مبتنی بر فیلتر و مبتنی بر Wrapper تقسیم شده‌اند. اهداف مبتنی بر فیلتر شامل کاهش آنتروپی و افزایش اطلاعات مشترک بین ویژگی‌ها هستند، در حالی که اهداف مبتنی بر Wrapper به کاهش تعداد ویژگی‌ها و بهبود عملکرد طبقه‌بندی تمرکز دارند. در نتیجه، راه‌حل نهایی مجموعه‌ای از راه‌حل‌های غالب است که هر یک دارای برداری از مؤلفه‌های آنتروپی، میزان اطلاعات مشترک، تعداد ویژگی‌ها و نرخ خطای طبقه‌بندی هستند. بهینه‌سازی در این مسئله به این صورت فرموله می‌شود که تعداد ویژگی‌های غیرمرتبط و نرخ خطای طبقه‌بندی به حداقل برسد تا بهترین ویژگی‌های ممکن برای پیش‌بینی سرطان استخراج شوند.

فرموله سازی مسئله

در این مقاله، انتخاب ویژگی به‌عنوان یک مسئله بهینه‌سازی چندمعیاری در نظر گرفته شده است که با استفاده از الگوریتم چندهدفه ساخت اهرام جیزه حل می‌شود. اهداف این بهینه‌سازی به دو دسته تقسیم می‌شوند: یکی کاهش آنتروپی و افزایش اطلاعات مشترک بین ویژگی‌ها و دیگری کاهش تعداد ویژگی‌ها و کاهش نرخ خطای طبقه‌بندی. هر ویژگی در مجموعه داده به‌صورت مستقل در فضای جستجوی باینری در بازه [۰، ۱] مدل می‌شود که نشان‌دهنده حضور یا عدم حضور آن ویژگی در راه‌حل نهایی است. از آنجا که الگوریتم ساخت اهرام جیزه یک روش فرااکتشافی مبتنی بر جمعیت اولیه است، جمعیت اولیه با انتخاب زیرمجموعه‌ای از ویژگی‌ها که به برچسب‌های کلاس مرتبط هستند، آغاز می‌شود[۲۸]. در این روش پیشنهادی، هر راه‌حل به‌صورت یک بردار باینری تعریف شده است که طول آن برابر با تعداد ویژگی‌ها در مجموعه داده است. اگر یک ویژگی خاص در راه‌حل حضور داشته باشد، مقدار درایه مربوط به آن در بردار برابر یک است و در غیر این صورت صفر خواهد بود[۲۸].

در این روش، ابتدا ماتریس باینری جمعیت اولیه به‌عنوان ورودی الگوریتم ساخت اهرام جیزه به کار می‌رود و به دو بخش تقسیم می‌شود. نیمی از این جمعیت با استفاده از رویکرد فیلتر و نیمی دیگر با رویکرد Wrapper ارزیابی می‌شوند. در هر بخش، تابع تناسب اختصاصی برای ارزیابی راه‌حل‌ها به کار می‌رود. الگوریتم ابتدا راه‌حل‌ها را با استفاده از این توابع تناسب بررسی می‌کند. در بخش Wrapper، هر راه‌حل که شامل مجموعه‌ای از ویژگی‌های انتخابی است، به یک طبقه‌بند KNN وارد می‌شود تا میزان خطای طبقه‌بندی آن مشخص گردد، زیرا یکی از اهداف تابع تناسب، کاهش خطای طبقه‌بندی است. در بخش فیلتر نیز تابع تناسب بر اساس میزان آنتروپی و اطلاعات مشترک بین ویژگی‌ها عمل می‌کند[۲۹]. پس از ارزیابی‌ها، از هر رویکرد یک راه‌حل بهینه

[DOI: 10.61882/ijbd.18.3.7]

با بهترین مقدار تابع تناسب انتخاب می‌شود و به‌عنوان ورودی روش یادگیری عمیق در نظر گرفته می‌شود.

در ادامه، روش یادگیری عمیق با استفاده از اعتبارسنجی متقابل و با حجم نمونه ۱۰ درصد، داده‌ها را با توجه به ویژگی‌های انتخاب شده از این دو راه‌حل طبقه‌بندی کرده و میزان خطا را به‌عنوان خروجی تولید می‌کند. راه‌حلی که کمترین میزان خطا و بیشترین دقت را دارد، به‌عنوان راه‌حل غالب آن مرحله انتخاب می‌شود۲۵/۹/۲۰۲۵. در مراحل بعدی، راه‌حل‌های جدید با توجه به این راه‌حل غالب تولید و به مرحله اکتشاف راه می‌یابند. در این فرآیند اکتشاف، راه‌حل برنده به‌عنوان مرجعی برای تولید نسل جدیدی از راه‌حل‌ها به کار می‌رود. این نسل نیز مانند قبل به دو بخش تقسیم شده و با توابع تناسب مرتبط با فیلتر و Wrapper ارزیابی می‌شود. این روند تا زمانی ادامه می‌یابد که شرط توقف برقرار شود[۳۰].

تابع تناسب پیشنهادی

تابع تناسب پیشنهادی در این مقاله با توجه به روش انتخاب ویژگی فیلتر یا Wrapper متفاوت است. روش انتخاب ویژگی مبتنی بر فیلتر که در این پژوهش به کار رفته است، با استفاده از ترکیب دو معیار آنتروپی و اطلاعات مشترک (MI[1])، ویژگی‌های موجود در مجموعه داده میکروآرایه را برای تشخیص سرطان ارزیابی می‌کند[۳۱]. هدف اصلی این روش بررسی ارتباط بین ژن‌ها و بیماری و انتخاب ویژگی‌هایی است که بیشترین ارتباط با برچسب کلاس (مثلاً وجود یا عدم وجود سرطان پستان) را دارند. در این فرآیند، ابتدا میزان آنتروپی هر ویژگی و اطلاعات مشترک بین هر ویژگی و برچسب کلاس محاسبه می‌شود. سپس، زیرمجموعه‌ای از ویژگی‌ها که کمترین آنتروپی و بیشترین اطلاعات مشترک را دارند، به‌عنوان بهترین زیرمجموعه انتخاب می‌شوند[۳۱]. تابع تناسبی نیز به‌طور خاص برای این روش پیشنهاد شده است که با کاهش آنتروپی و افزایش اطلاعات مشترک میان ویژگی‌ها و برچسب کلاس، جمعیت اولیه از ویژگی‌ها را ارزیابی می‌کند. این تابع در طول مراحل تکرار الگوریتم به انتخاب زیرمجموعه‌ای از ویژگی‌ها منجر می‌شود که پیش‌بینی دقیق‌تری از بیماری سرطان ارائه دهد. در انتخاب ویژگی مبتنی بر فیلتر در روش پیشنهادی از توابع آنتروپی و اطلاعات مشترک استفاده می‌شود که در روابط زیر تعریف شده است[۳۲].

$$E(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y) log p(x,y) \quad (۱)$$

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x).p(y)} \quad (۲)$$

که در آن X مقادیر کل برای یک ویژگی و x یک مقدار از ویژگی، Y کل کلاس‌ها و y نماینده هر کلاس در مجموعه داده، $p(x,y)$ احتمال حضور مقدار خاص x در کلاس y، $p(x)$ احتمال کل مقدار خاص x در ویژگی و $p(y)$ احتمال کلاس در کل مجموعه داده است. بر این اساس تابع تناسب پیشنهادی برای روش انتخاب ویژگی مبتنی بر فیلتر به صورت زیر خواهد بود

$$\begin{aligned} &min\, f_f = E(X,Y) - MI(X,Y) \\ &s.t. \\ &\sum_{x\in X}\sum_{y\in Y} p(x,y) \le 1 \\ &\sum_{x\in X} p(x) \le 1 \\ &\sum_{y\in Y} p(y) \le 1 \\ &x, y > 0 \end{aligned} \quad (۳)$$

در روش انتخاب ویژگی مبتنی بر فیلتر، تابع تناسب پیشنهادی برای هر زیرمجموعه ویژگی محاسبه می‌شود تا بهترین زیرمجموعه ممکن شناسایی شود[۲۹]. هدف اصلی این روش، یافتن زیرمجموعه‌ای از ویژگی‌هاست که بیشترین میزان اطلاعات مشترک با برچسب کلاس‌ها و کمترین آنتروپی را داشته باشد. در هر مرحله، زیرمجموعه‌ای که بالاترین مقدار تابع تناسب را داشته باشد به‌عنوان راه‌حل غالب (Sf) انتخاب می‌شود. این راه‌حل غالب، مجموعه‌ای از ویژگی‌ها را شامل می‌شود که با داشتن بالاترین اطلاعات مشترک و کمترین آنتروپی، می‌تواند به بهترین شکل بیماران سرطانی را پیش‌بینی کند. بنابراین، این فرآیند تکراری به دنبال یافتن راه‌حلی از میان جمعیت اولیه است که با معیارهای تناسب بهترین همخوانی را داشته باشد و در نتیجه، دقت طبقه‌بندی را افزایش دهد.

[1] Mutual Information

[DOI: 10.61882/ijbd.18.3.7]

در روش انتخاب ویژگی مبتنی بر Wrapper، جمعیت اولیه از زیرمجموعه‌های ویژگی بر اساس ترکیبی از تعداد ویژگی‌ها و میزان خطای طبقه‌بندی ارزیابی می‌شود. برای کاهش پیچیدگی محاسباتی، از یک طبقه‌بند سبک مانند نزدیک‌ترین همسایه (KNN) استفاده می‌شود که امکان ارزیابی سریع ویژگی‌ها را فراهم می‌کند. تابع تناسب پیشنهادی در این روش، ترکیبی از نرخ انتخاب ویژگی‌ها و خطای طبقه‌بندی است، به طوری که زیرمجموعه‌هایی از ویژگی‌ها که تعداد ویژگی‌های کمتر و دقت طبقه‌بندی بالاتری دارند، امتیاز بیشتری می‌گیرند۲۵/۹/۲۰۲۵. این رویکرد به انتخاب بهترین مجموعه ویژگی‌هایی منجر می‌شود که نه‌تنها به‌خوبی با داده‌ها مرتبط هستند، بلکه دقت بالایی در طبقه‌بندی دارند. تابع تناسب مربوط به انتخاب ویژگی Wrapper را می‌توان به صورت رابطه (۴) می‌توان تعریف کرد:

$$\min f_w = \sum_{j=1}^{M} \frac{n_j}{N} + \sum_{i=1}^{n} \sum_{j=1}^{M} s_{ij} x_{ij}$$
s.t.
$$\sum_{j=1}^{k} \frac{n_j}{N} \leq 1$$
$$\sum_{i=1}^{n} S_i \leq \alpha$$
$$\sum_{i=1}^{k} x_j \leq \beta \quad (۴)$$

که در آن N تعداد کل ویژگی‌ها و n تعداد ویژگی‌های انتخاب شده در هر راه حل، M تعداد راه حل‌ها، S_{ij} جریمه مربوط به راه حل، x_{ij} مقدار خطای طبقه‌بندی توسط راه حل، α حداکثر مقدار جریمه و β حداکثر مقدار خطای مجاز برای هر راه حل می‌باشد. پس از محاسبه مقدار تابع تناسب برای همه راه حل‌ها، بهترین راه حل یافت شده به عنوان راه حل غالب (S_w) در روش انتخاب ویژگی مبتنی بر Wrapper انتخاب می‌شود.

در روش پیشنهادی، دو راه‌حل بهینه‌ی حاصل از روش‌های انتخاب ویژگی مبتنی بر فیلتر (Sf) و Wrapper (Sw) به‌عنوان ورودی به شبکه عصبی عمیق اعمال می‌شوند. این کار باعث کاهش پیچیدگی محاسباتی می‌شود، زیرا به‌جای اجرای شبکه عصبی برای تعداد زیادی از زیرمجموعه‌ها، تنها برای این دو راه‌حل اجرا می‌شود. در نهایت، دقت هر یک از این راه‌حل‌ها در طبقه‌بندی بیماران سرطانی محاسبه و مقایسه می‌شود. اگر عملکرد روش مبتنی بر Wrapper بهتر باشد، نسل جدید الگوریتم ساخت اهرام جیز بر پایه Sw شکل می‌گیرد و در غیر این صورت بر اساس Sf ایجاد می‌شود. سپس جمعیت جدیدی با ترکیب این دو رویکرد ایجاد شده و این مراحل تکرار می‌شوند تا شرط توقف، یعنی ۱۰۰ بار اجرای الگوریتم، محقق شود. در پایان، از میان دو راه‌حل نهایی Sf و Sw، بهترین انتخاب می‌شود.

در این روش از الگوریتم‌های بهینه‌سازی چندهدفه برای انتخاب ژن‌های مؤثر در داده‌های میکروآرایه استفاده می‌شود و نتایج به‌دست‌آمده از این الگوریتم‌ها به‌عنوان ورودی برای شبکه عصبی عمیق به کار می‌روند. داده‌های ورودی، ژن‌های گزینش‌شده‌ای هستند که نمایانگر ویژگی‌های مؤثر ژنتیکی برای تشخیص بیماری سرطان پستان در داده‌های میکروآرایه هستند. در نهایت، عملکرد الگوریتم‌های فراابتکاری برای انتخاب ژن‌ها و دقت این انتخاب‌ها ارزیابی می‌شود. این روش ترکیبی از فیلتر و Wrapper با استفاده از الگوریتم‌های بهینه‌سازی چندهدفه و روش‌های فراابتکاری است که به همراه شبکه عصبی عمیق، دقت و کارایی بالایی در پیش‌بینی و طبقه‌بندی داده‌های میکروآرایه ارائه می‌دهد.

یافته‌ها

در این بخش، یافته‌های کلیدی حاصل از پیاده‌سازی و ارزیابی روش ترکیبی پیشنهادی برای تشخیص سرطان پستان با استفاده از داده‌های میکروآرایه DNA ارائه می‌شود برای ارزیابی کارایی روش پیشنهادی، از دو مجموعه‌داده میکروآرایه DNA معتبر و با دسترسی عمومی استفاده شد: BC-TCGA و GSE [33]. این مجموعه‌داده‌ها حاوی اطلاعات بیان ژن‌های مرتبط با سرطان پستان در دو گروه اصلی بیماران مبتلا به سرطان پستان و افراد سالم هستند.

جدول۲: اطلاعات آماری مربوط به مجموعه داده‌ها

Table 2: Statistical information related to the datasets

Dataset	Number of instance	Number of genes	Number of normal instance	Rate of normal instance	Number of cancer instance	Rate of cancer instance
BC-TCGA	590	17814	61	0.1034	529	0.8966
GSE	200	10000	100	0.5	100	0.5

[DOI: 10.61882/ijbd.18.3.7]

جزئیات آماری مربوط به این مجموعه‌داده‌ها، از جمله تعداد نمونه‌ها و ژن‌های اولیه، در جدول ۲ ارائه شده است. انتخاب این مجموعه داده‌ها، امکان مقایسه و اعتبارسنجی دقیق عملکرد الگوریتم را فراهم آورد.

یکی از مهم‌ترین یافته‌ها، افزایش تدریجی و پایدار دقت روش پیشنهادی در طول مراحل تکرار الگوریتم است. همانطور که در شکل ۲ به وضوح قابل مشاهده است، با پیشرفت در چرخه بهینه‌سازی، راه‌حل‌های نهایی به طور مداوم بهبود یافته و به سمت انتخاب بهینه‌ترین ژن‌ها برای طبقه‌بندی بیماران سرطانی و پیش‌بینی وضعیت نمونه‌های جدید حرکت می‌کنند. این روند بهبود تا آخرین مرحله تکرار ادامه می‌یابد و در نهایت، بالاترین دقت طبقه‌بندی در آخرین دور از تکرار الگوریتم پیشنهادی حاصل می‌شود. این همگرایی نشان‌دهنده قابلیت الگوریتم در هدایت جستجو به سمت ترکیبات ویژگی‌های موثر است.

جدول ۲ به تفصیل نتایج مربوط به راه‌حل‌هایی که توسط الگوریتم ساخت اهرام جیزه برای هر یک از مجموعه‌داده‌های BC-TCGAو GSE ارائه شده‌اند را نشان می‌دهد. این نتایج حاکی از آن است که فرآیند انتخاب ویژگی در هر مرحله، به طور مؤثر کیفیت طبقه‌بندی را بهبود می‌بخشد و در نهایت به انتخاب ژن‌هایی منجر می‌شود که بیشترین ارتباط را با تشخیص سرطان پستان دارند. این جدول، تأثیر مستقیم بهینه‌سازی انتخاب ویژگی را بر دقت نهایی طبقه‌بندی و پیش‌بینی برای هر مجموعه داده مشخص می‌کند. الگوریتم پیشنهادی، با تنظیم معیار ارزیابی بر اساس مقدار تابع تناسب، تقریباً نیمی از ژن‌های هر مجموعه داده را به عنوان ورودی برای مرحله یادگیری عمیق انتخاب می‌کند. این رویکرد تضمین می‌کند که تنها ژن‌های با بیشترین احتمال ارتباط با بیماری در تحلیل‌های بعدی مورد استفاده قرار می‌گیرند.

نمودار شکل ۳ فرآیند همگرایی الگوریتم پیشنهادی به سمت یک راه‌حل بهینه را به تصویر می‌کشد. این نمودار نشان می‌دهد که چگونه راه‌حل‌ها در هر مرحله از تکرار، به طور تدریجی به هدف نهایی (یعنی انتخاب بهینه‌ترین مجموعه ژن‌ها) نزدیک‌تر می‌شوند. این همگرایی به معنای توانایی الگوریتم در یافتن یک نقطه مطلوب در فضای جستجو و تثبیت آن برای دستیابی به بالاترین عملکرد طبقه‌بندی در هر مجموعه داده است.

همان‌طور که در شکل ۳ نشان داده شده است، مقدار تابع تناسب راه‌حل‌های به دست آمده از الگوریتم پیشنهادی، با افزایش تکرارها به سمت بهبود و کاهش پیش می‌رود. این همگرایی در حدود ۱۰۰ مرحله به یک مقدار ثابت و بهینه می‌رسد. این یافته نشان می‌دهد که الگوریتم پیشنهادی، قبل از رسیدن به ۱۰۰ تکرار به نقطه بهینه خود می‌رسد و تکرارهای بیشتر عملاً تغییر معناداری در بهینه‌سازی تابع تناسب ایجاد نمی‌کنند. به عبارت دیگر، پس از اینکه تابع تناسب به یک نقطه بهینه با خطای نزدیک به صفر درصد همگرا می‌شود، ادامه تکرارها تنها منجر به افزایش زمان اجرا می‌شود بدون اینکه بهبود بیشتری در دقت یا کاهش خطا حاصل شود. این یافته بر کارایی و پایداری روش پیشنهادی در یافتن ژن‌های مرتبط و رسیدن به یک راه‌حل بهینه تأکید دارد، و در نهایت، راه‌حل بهینه نهایی برای هر نمونه در مرحله آموزش الگوریتم تعیین می‌شود.

برای پیش‌بینی برچسب کلاس نمونه‌های تست، روش پیشنهادی از شبکه‌های عصبی کانولوشنی (CNN) بهره می‌برد. این شبکه‌ها با استفاده از تابع تناسب به عنوان هسته اصلی خود، به طور دقیق آموزش داده می‌شوند. علاوه بر CNN، چهار روش طبقه‌بندی دیگر شامل kنزدیک‌ترین همسایه(kNN)، بیزین ساده(NB)، درخت تصمیم، و رگرسیون لجستیک نیز مورد استفاده قرار می‌گیرند. هر یک از این روش‌ها با بهره‌گیری از زیرمجموعه ویژگی‌هایی که توسط الگوریتم انتخاب ویژگی GPC تعیین شده‌اند، قادر به پیش‌بینی دقیق برچسب کلاس نمونه‌های تست هستند. این ترکیب چندگانه از یک شبکه عصبی پیشرفته و الگوریتم‌های طبقه‌بندی متنوع، امکان دست‌یابی به دقت بالاتر و پوشش بهتر برای شناسایی ژن‌های مرتبط را فراهم می‌کند و از تنوع روش‌ها برای افزایش اطمینان در نتایج بهره می‌برد.

با استفاده از روش پیشنهادی برای انتخاب ژن به عنوان مدل پیش‌بینی، امکان ارزیابی وضعیت سلامت و پیش‌بینی نمونه‌های جدید DNA در مجموعه داده‌های میکروآرایه فراهم می‌شود. در این فرآیند، ابتدا ژن‌های مرتبط با بیماری با استفاده از الگوریتم انتخاب ژن شناسایی می‌شوند. سپس

[DOI: 10.61882/ijbd.18.3.7]

این ژن‌های منتخب به عنوان ویژگی‌های مهم و کلیدی در مدل طبقه‌بندی مورد استفاده قرار می‌گیرند. الگوریتم GPC به طور خاص، ژن‌هایی را که با ویژگی‌های بیماری مرتبط هستند، از داده‌های میکروآرایه استخراج کرده و آن‌ها را به عنوان ورودی به سیستم پیش‌بینی می‌دهد. در مرحله بعد، نمونه‌های جدید در فضای داده‌های میکروآرایه براساس ژن‌های انتخاب‌شده قرار می‌گیرند و از این طریق احتمال ابتلا به بیماری در فرد مورد نظر شناسایی می‌شود .روند آموزش شبکه‌های عصبی کانولوشن بر روی ژن‌های باقیمانده از مرحله انتخاب ویژگی در شکل ۴ قابل مشاهده است .این مراحل، چارچوب جامعی را برای تشخیص زودهنگام و دقیق سرطان پستان ارائه می‌دهند.

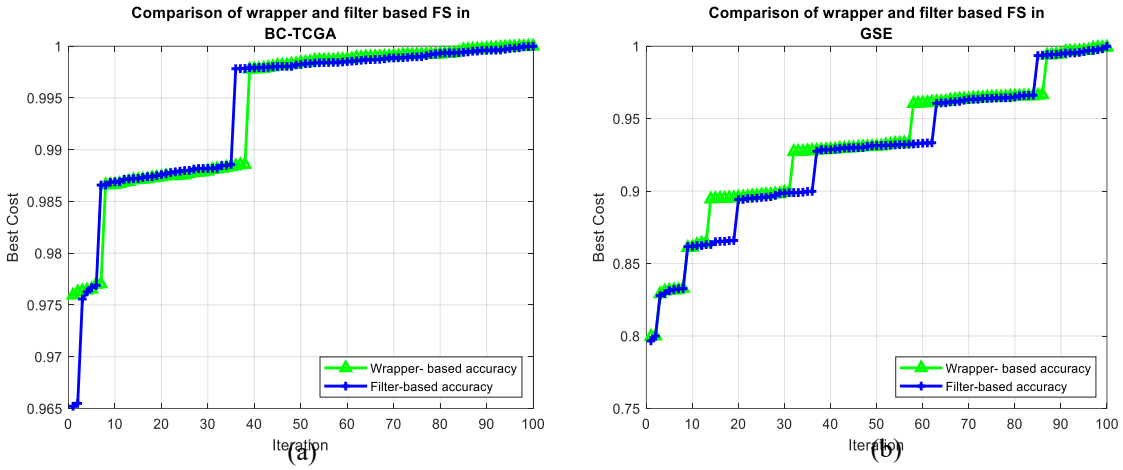


Figure 2: Comparison of accuracy of filter-based and wrapper-based feature selection approaches based on deep learning (a) BC-TCA (b) GSE

شکل ۲: مقایسه دقت رویکردهای انتخاب ویژگی مبتنی بر فیلتر و Wrapper بر اساس یادگیری عمیق (a) BC-TCA (b) GSE

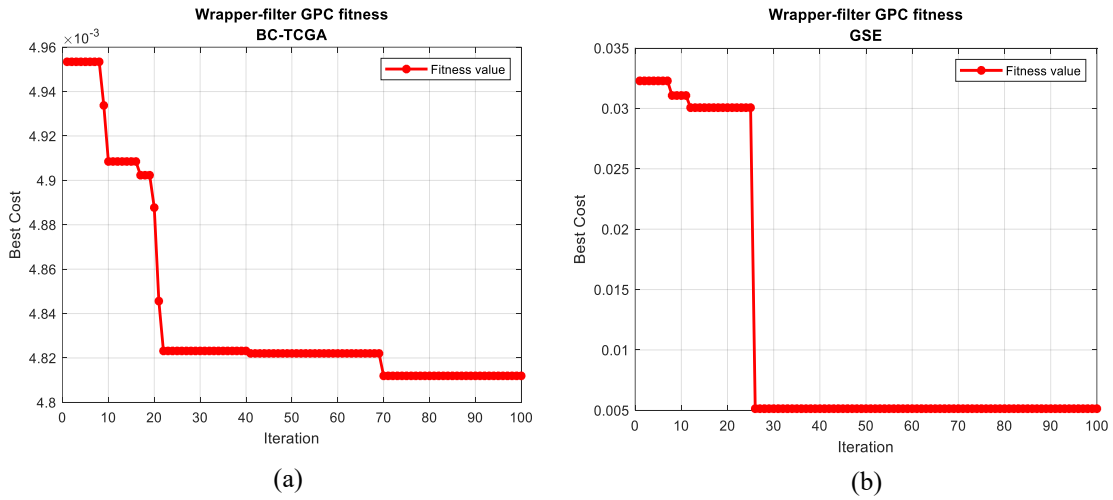
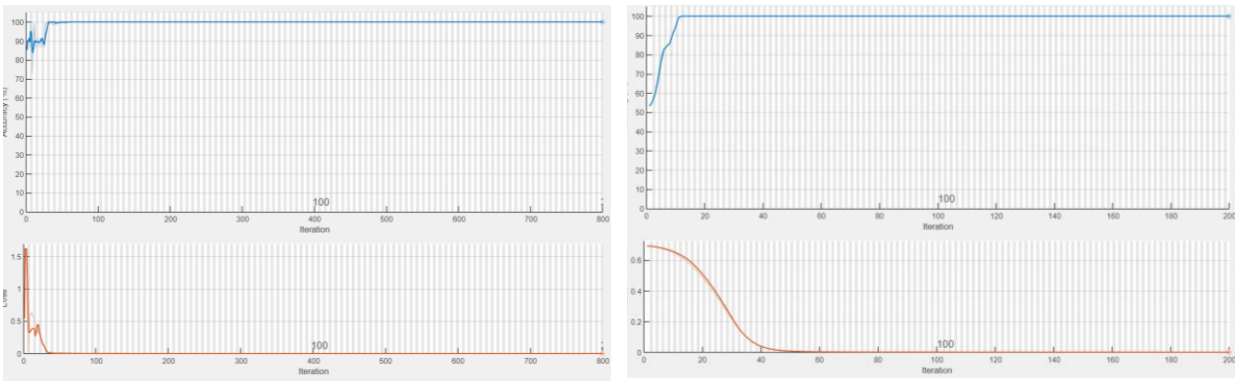


Figure 3: Comparison of accuracy of filter-based and wrapper-based feature selection approaches based on deep learning (a) BC-TCA (b) GSE

شکل ۳: مقایسه دقت رویکردهای انتخاب ویژگی مبتنی بر فیلتر و Wrapper بر اساس یادگیری عمیق (a) BC-TCA (b) GSE

[DOI: 10.61882/ijbd.18.3.7]



(a) (b)

Figure 4: Convolutional neural network training process on the dataset (a) BC-TCA (b) GSE

شکل ۴: روند آموزش شبکه‌های عصبی کانولوشن بر روی مجموعه داده‌ها (a) BC-TCA (b) GSE

برای ارزیابی دقت روش پیشنهادی در پیش‌بینی نمونه‌های جدید DNA، از ماتریس آشفتگی استفاده می‌شود. این ماتریس ابزاری برای اندازه‌گیری عملکرد سیستم طبقه‌بندی است که به‌طور دقیق تعداد نمونه‌هایی را که به‌درستی یا نادرستی طبقه‌بندی شده‌اند، نشان می‌دهد. ماتریس آشفتگی شامل چهار بخش اصلی است: مثبت صحیح(TP[1])، منفی صحیح (TN[2])، مثبت کاذب (FP[3]) و منفی کاذب (FN[4]). این چهار بخش به‌طور دقیق نحوه عملکرد مدل در شناسایی نمونه‌های درست و نادرست را نشان می‌دهند و می‌توانند به‌عنوان معیاری برای ارزیابی کارایی مدل در تشخیص بیماری‌ها و پیش‌بینی وضعیت سلامت افراد جدید استفاده شوند. استفاده از ماتریس آشفتگی این امکان را فراهم می‌آورد که نقاط ضعف و قوت مدل شناسایی شده و در صورت لزوم بهبودهایی در الگوریتم‌های انتخاب ژن یا طبقه‌بندی اعمال شود. پارامترهای مربوط به ماتریس آشفتگی در شبکه‌های عصبی کانولوشن و سایر روش‌های طبقه بندی در هر یک از مجموعه داده‌های مورد استفاده در روش پیشنهادی در جدول ۳ نشان داده شده است.

در جدول ۳، نتایج مربوط به پارامترهای ماتریس آشفتگی برای هر روش طبقه‌بندی، به منظور بررسی دقیق‌تر عملکرد این روش‌ها محاسبه شده‌اند. چهار معیار اصلی برای ارزیابی کارایی مدل‌ها به‌کار گرفته شده است: دقت، حساسیت، صحت، و معیار F. دقت[5] به عنوان یک شاخص کلی نشان می‌دهد چه نسبتی از کل نمونه‌ها به‌درستی طبقه‌بندی شده‌اند و برای بررسی عملکرد کلی مدل مفید است. حساسیت[6] نشان‌دهنده توانایی مدل در شناسایی درست نمونه‌های مثبت است و کاربرد ویژه‌ای در مسائلی مانند تشخیص بیماری‌ها دارد، چرا که تشخیص دقیق موارد مثبت از اهمیت بالایی برخوردار است. صحت[7] نیز بر تشخیص درست نمونه‌های منفی تمرکز دارد و به‌ویژه در مسائلی اهمیت دارد که نادرست شناسایی‌شدن موارد منفی می‌تواند پیامدهای منفی به دنبال داشته باشد. معیار F[8] نیز به‌عنوان میانگین‌هارمونیک دقت و حساسیت، تعادلی بین این دو معیار ایجاد می‌کند و برای مسائلی مفید است که به طور همزمان به دقت و حساسیت نیاز دارند.

هر یک از این معیارها، مزایا و محدودیت‌های خاص خود را دارند. به‌عنوان مثال، دقت به‌تنهایی ممکن است در مسائلی که عدم تعادل در تعداد نمونه‌های مثبت و منفی وجود دارد، به اندازه کافی گویا نباشد و در این موارد از حساسیت و صحت نیز استفاده می‌شود. به همین ترتیب، معیار F برای مسائلی که هم خطای مثبت کاذب و هم منفی کاذب مهم هستند، انتخاب مناسبی است. انتخاب معیار یا معیارهای مناسب به ماهیت مسئله و نیازهای خاص پژوهش بستگی دارد و معمولاً

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

[5] Accuracy
[6] Recall
[7] Precision
[8] F-score

[DOI: 10.61882/ijbd.18.3.7]

برای ارزیابی جامع، از چندین معیار به صورت ترکیبی استفاده می‌شود.

به منظور بررسی جامع‌تر و دقیق‌تر عملکرد روش پیشنهادی، ترکیب روش انتخاب ژن با استفاده از الگوریتم GPC با روش‌های مختلف طبقه‌بندی همچون نزدیکترین همسایه (KNN)، درخت تصمیم، بیزین سادهو رگرسیون لجستیک مقایسه شده است. این مقایسه‌ها به منظور شناسایی بهترین روش طبقه‌بندی که بیشترین دقت و کارایی را در پیش‌بینی بیماران مبتلا به سرطان پستان دارد، انجام شده است. در شکل ۵ تا ۸، نتایج مقایسه نتایج معیارهای ارزیابی برای پیش‌بینی بیماران مبتلا به سرطان پستان در مجموعه داده‌های مختلف نشان داده شده است.

جدول ۳: پارامترهای ماتریس آشفتگی برای روش‌های طبقه بندی مختلف در مجموعه داده‌ها

Table 3: Confusion matrix parameters for different classification methods on datasets

Dataset	Classifier	TP	FP	TN	FN
BC-TCGA	CNN	25	0	140	12
	KNN	15	10	105	47
	DT	14	11	138	14
	NB	16	9	138	14
	LR	5	20	140	12
GSE	CNN	31	0	20	9
	KNN	26	5	4	25
	DT	28	3	18	11
	NB	26	5	8	21
	LR	19	12	18	11

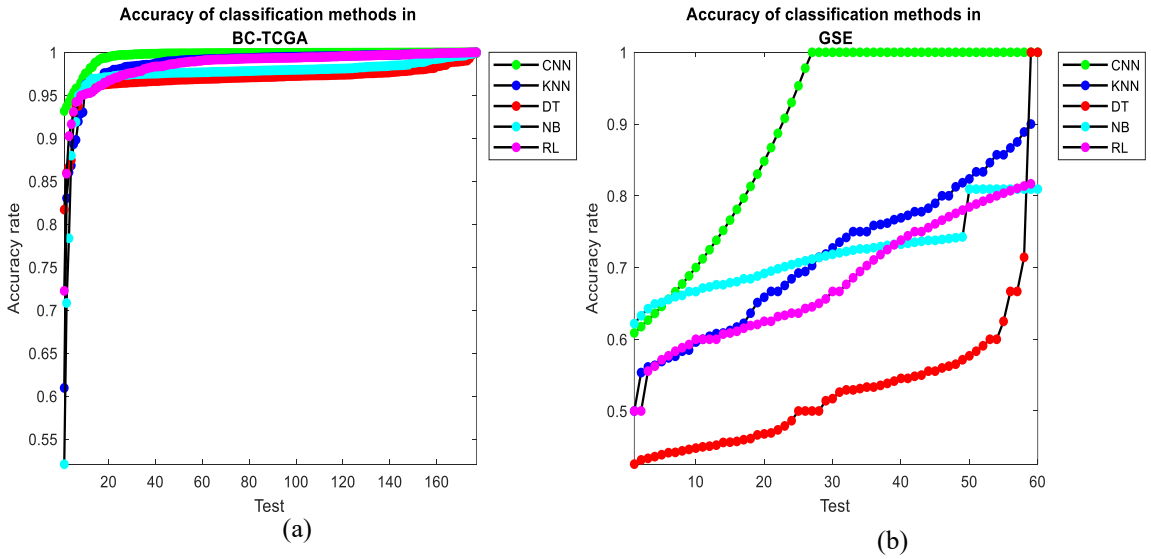
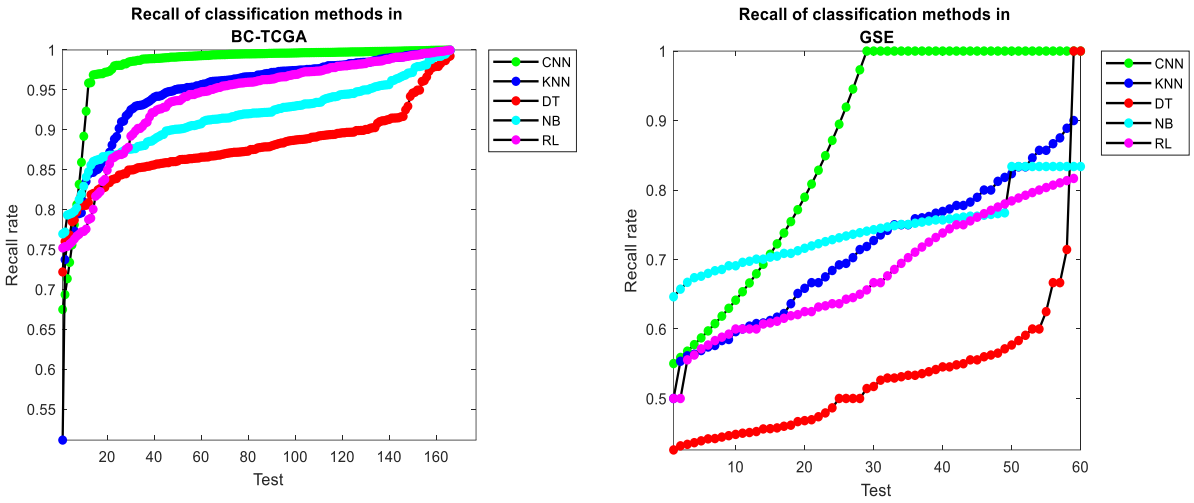


Figure 5: Comparison of the accuracy of the proposed method with classification algorithms on the datasets (a) BC-TCA (b) GSE

شکل ۵: مقایسه دقت روش‌پیشنهادی با الگوریتم‌های طبقه‌بندی بر روی مجموعه داده‌ها (a) BC-TCA (b) GSE

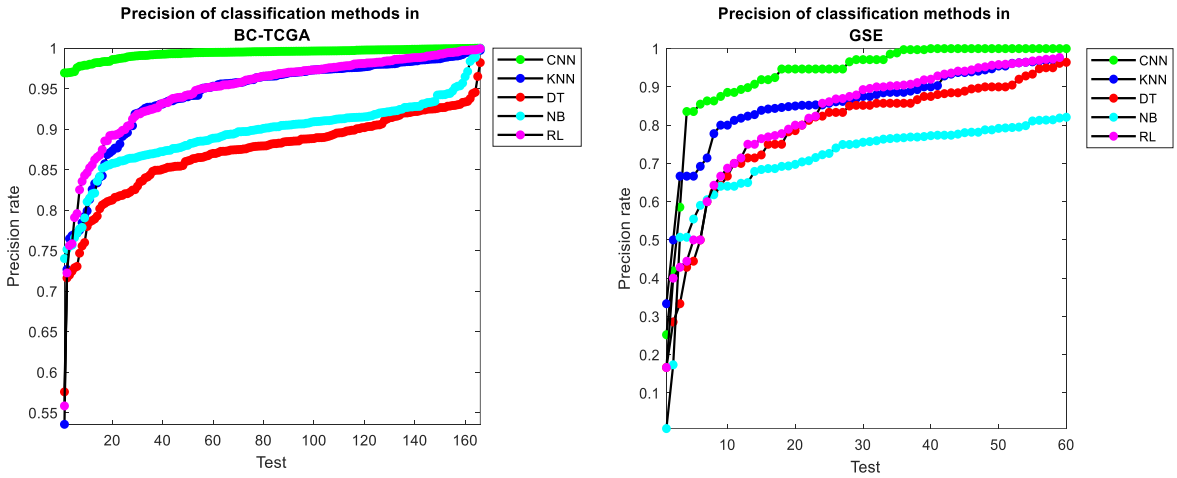
[DOI: 10.61882/ijbd.18.3.7]



(a) (b)

Figure 6: Comparison of the recall of the proposed method with classification algorithms on the datasets (a) BC-TCA (b) GSE

شکل ۶: مقایسه حساسیت روش‌پیشنهادی با الگوریتم‌های طبقه‌بندی بر روی مجموعه داده‌ها (a) BC-TCA (b) GSE



(a) (b)

Figure 7: Comparison of the precision of the proposed method with classification algorithms on the datasets (a) BC-TCA (b) GSE

شکل ۷: مقایسه صحت روش‌پیشنهادی با الگوریتم‌های طبقه‌بندی بر روی مجموعه داده‌ها (a) BC-TCA (b) GSE

[DOI: 10.61882/ijbd.18.3.7]

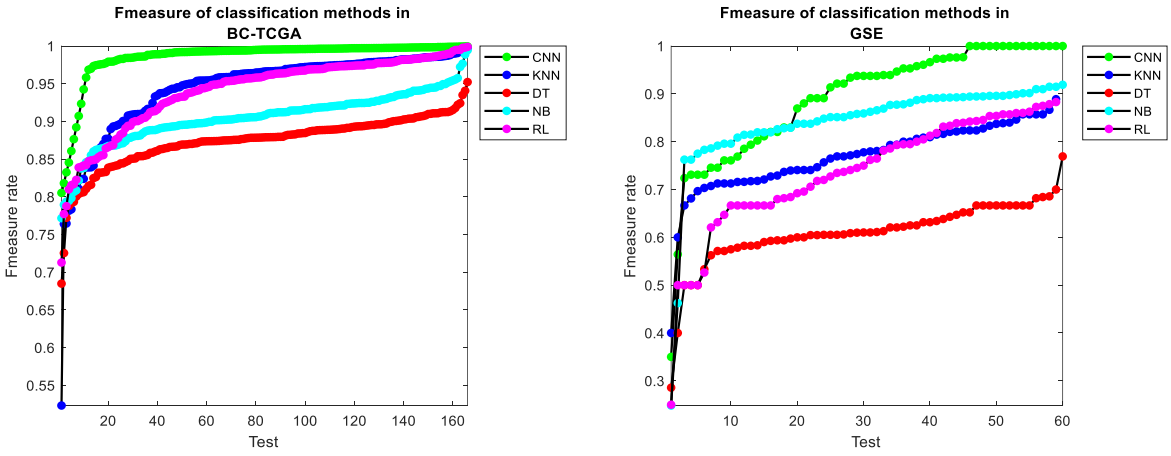


Figure 8: Comparison of the F-measure of the proposed method with classification algorithms on the datasets (a) BC-TCA (b) GSE

شکل ۸: مقایسه معیار F روش پیشنهادی با الگوریتم‌های طبقه‌بندی بر روی مجموعه داده‌ها (a) BC-TCA (b) GSE

با توجه به شکل‌های ۵ تا ۸ می‌توان دید استفاده از ژن‌های انتخاب شده با رویکرد ترکیبی تعاملی پیشنهادی بر اساس الگوریتم بهینه‌سازی GPC، توانسته است معیارهای ارزیابی برای پیش بینی زودهنگام بیماران مبتلا به سرطان پستان، به میزان قابل‌توجهی افزایش یافته است و در مقایسه با روش‌های دیگر، عملکرد بهتری را در تشخیص صحیح بیماران ارائه دهد.

این یافته‌ها نشان‌دهنده اثربخشی روش پیشنهادی در انتخاب ژن‌های مرتبط و بهبود دقت مدل‌های طبقه‌بندی برای کاربردهای بالینی می‌باشد. در شکل ۹ و جدول ۴ مقادیر میانگین معیارهای ارزیابی برای روش‌های طبقه بندی مختلف بر اساس ترکیب با انتخاب ژن مبتنی بر الگوریتم GPC در مجموعه داده‌های مختلف نشان داده شده است.

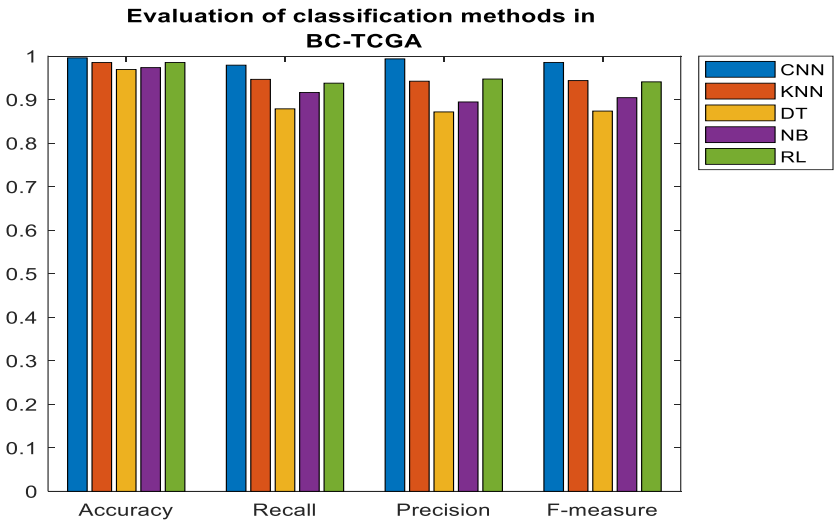


Figure 9: Bar chart comparing the average evaluation criteria for classification methods on the dataset (a) BC-TCA (b) GSE

شکل ۹: نمودار میله‌ای مقایسه میانگین معیارهای ارزیابی برای روش‌های طبقه‌بندی روی مجموعه داده‌ها (a) BC-TCA (b) GSE

[DOI: 10.61882/ijbd.18.3.7]

جدول ۴: مقایسه مقادیر میانگین معیارهای ارزیابی برای روش‌های طبقه‌بندی روی مجموعه داده‌ها

Table 4: Comparison of average values of evaluation criteria for classification methods on datasets

Dataset	Classifier	Accuracy	Recall	Precision	F-measure
BC-TCGA	CNN	0.9996	0.9792	0.9936	0.9854
	KNN	0.9851	0.9466	0.9423	0.9437
	DT	0.9692	0.8788	0.8717	0.8737
	NB	0.9735	0.9164	0.8946	0.9045
	LR	0.9853	0.9378	0.9473	0.9408
GSE	CNN	0.9610	0.9110	0.9555	0.9262
	KNN	0.7073	0.7073	0.8686	0.7806
	DT	0.8481	0.8481	0.8966	0.8564
	NB	0.7134	0.7134	0.8698	0.7813
	LR	0.8293	0.8293	0.8885	0.8470

در شکل ۹ و جدول ۴ مشاهده می‌شود که روش پیشنهادی، که ترکیبی از انتخاب ویژگی با الگوریتم GPC و شبکه‌های عصبی کانولوشنی است، به‌طور قابل توجهی توانسته به مقادیر نزدیک به بهینه در معیارهای ارزیابی دست یابد. این موفقیت به دلیل انتخاب دقیق ژن‌های مرتبط و مؤثر توسط الگوریتم GPC و استفاده از شبکه‌های عصبی کانولوشن برای آموزش عمیق روی این ژن‌هاست که موجب افزایش دقت در پیش‌بینی می‌شود. هرچند سایر روش‌های طبقه‌بندی نیز نتایج مناسبی ارائه کرده‌اند، اما این نتایج اهمیت دقت انتخاب ژن‌های مؤثر را در بهبود طبقه‌بندی بیماران مبتلا به سرطان پستان نشان می‌دهد و تأکیدی بر کارایی روش GPC در شناسایی ژن‌های حیاتی برای طبقه‌بندی دقیق‌تر است.

بحث

پژوهش حاضر با ارائه یک رویکرد نوین در تشخیص سرطان پستان بر پایه تلفیق هوشمندانه انتخاب ویژگی و یادگیری عمیق، گامی مهم در جهت بهبود دقت و کارایی سیستم‌های پشتیبان تصمیم‌گیری بالینی برداشته است. نوآوری اصلی در این مطالعه، توسعه یک روش ترکیبی برای انتخاب ویژگی است که نقاط قوت رویکردهای فیلتر و Wrapper را با یکدیگر ادغام می‌کند. مزیت این ترکیب، بهره‌گیری از سرعت و استقلال مدل در مرحله فیلتر و دقت بالای انتخاب ویژگی در مرحله Wrapper است، که به‌صورت تعاملی و بهینه با یکدیگر عمل می‌کنند. فراتر از آن، الهام‌گیری از الگوریتم ساخت اهرام جیزه برای بهینه‌سازی فرآیند انتخاب ویژگی، یک نوآوری قابل توجه محسوب می‌شود. این الگوریتم فراابتکاری، برخلاف روش‌های جستجوی سنتی، قادر به جستجوی جامع‌تر در فضای ویژگی‌ها و اجتناب از گیر افتادن در بهینه‌های محلی است، که به شناسایی بهینه‌ترین زیرمجموعه از ویژگی‌ها کمک شایانی می‌کند.

در مرحله بعدی، استفاده از شبکه‌های عصبی کانولوشنی (CNN) به عنوان مدل یادگیری عمیق، نقش محوری در استخراج الگوهای پیچیده و انتزاعی از داده‌های میکروآرایه‌ای ژنتیکی ایفا کرده است CNNها با قابلیت یادگیری سلسله‌مراتبی ویژگی‌ها، به‌طور موثری نویز و ویژگی‌های غیرمرتبط را در داده‌های زیستی پرحجم فیلتر می‌کنند. این ترکیب بهینه از انتخاب ویژگی و قدرت تحلیلیCNN، منجر به افزایش قابل توجه دقت پیش‌بینی در مقایسه با مطالعات پیشین شده است. در حالی که بسیاری از پژوهش‌ها بر پیش‌پردازش ساده یا استفاده از دسته‌بندهای کلاسیک مانند kNN و SVM تمرکز داشتند [۱۴]، رویکرد ما با فراتر رفتن از این محدودیت‌ها و بهره‌گیری از پیچیدگی یادگیری عمیق، توانسته است عملکردی فراتر از حد انتظار ارائه دهد.

[DOI: 10.61882/ijbd.18.3.7]

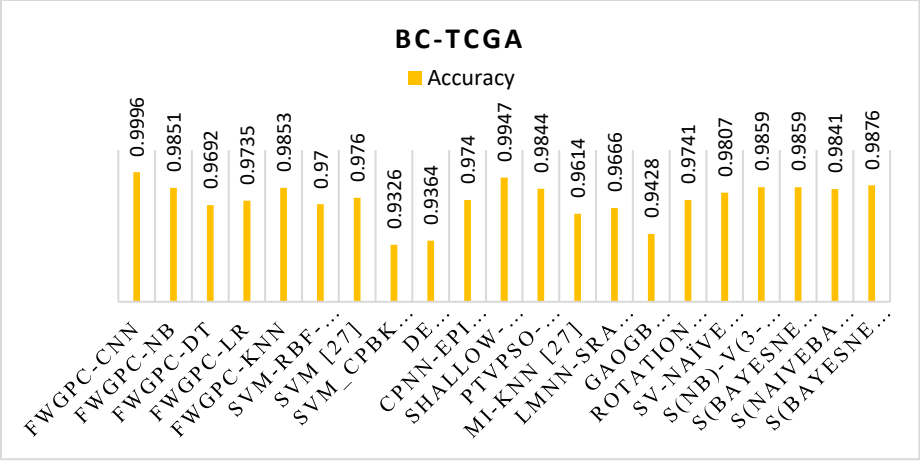


Figure 10: Comparison of the proposed method with previous methods in terms of accuracy criteria on the BC-TCGA dataset.

شکل ۱۰: مقایسه روش پیشنهادی با روش‌های پیشین از نظر معیار دقت در مجموعه داده BC-TCGA

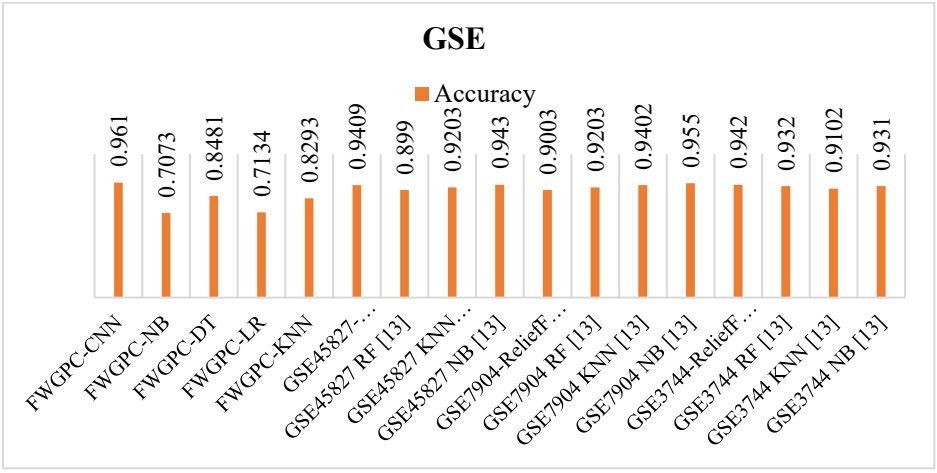


Figure 11: Comparison of the proposed method with previous methods in terms of accuracy criteria on the GSE dataset.

شکل ۱۱: مقایسه روش پیشنهادی با روش‌های پیشین از نظر معیار دقت در مجموعه داده GSE

یادگیری عمیق در پردازش داده‌ها بوده است که امکان تجزیه و تحلیل دقیق‌تری را فراهم کرده است.

روش پیشنهادی ما از نظر انعطاف‌پذیری و قابلیت تعمیم در کار با مجموعه‌داده‌های مختلف، برتری چشمگیری دارد. برخلاف مدل‌های سنتی که اغلب در شناسایی الگوهای غیرخطی پیچیده با چالش مواجه هستند[18] ، CNNتوانایی بالایی در استخراج این الگوها دارد. این موضوع، به همراه بهینه‌سازی ویژگی‌ها، منجر به عملکرد بهتر مدل در معیارهای کلیدی مانند دقت، یادآوری و مساحت زیر منحنی در مقایسه با سایر روش‌های بررسی‌شده در ادبیات شده است. این دستاورد به دلیل بهره‌گیری از قدرت اکتشافی الگوریتم‌های فراابتکاری در کنار قدرت یادگیری عمیق است که امکان شناسایی ترکیب‌های بهینه‌تر ویژگی‌ها را فراهم می‌کند.

همان‌گونه که در شکل‌های ۱۰ و ۱۱ نشان داده شده است، روش پیشنهادی در این پژوهش برای بهبود دقت در پیش‌بینی بیماران سرطان پستان از یک تکنیک انتخاب ویژگی ترکیبی استفاده کرده است. این روش، که ترکیبی از رویکردهای فیلتر و Wrapper است، با الهام از الگوریتم ساخت اهرام جیزه بهینه شده است. به کمک این ترکیب، ویژگی‌های مهم و تأثیرگذار از میان داده‌ها انتخاب می‌شوند و در نهایت، مدل پیش‌بینی دقیق‌تری ساخته می‌شود. پس از انتخاب ویژگی‌ها، یک مدل یادگیری عمیق مبتنی بر شبکه‌های عصبی کانولوشنی برای تحلیل و پیش‌بینی نهایی به کار گرفته شده است. با استفاده از این روش، دقت مدل در پیش‌بینی بیماران جدید مبتلا به سرطان پستان در مجموعه داده‌های مختلف بهبود یافته و نسبت به روش‌های پیشین منتشر شده در مقالات، نتایج بهتری از نظر دقت به دست آورده است. این بهبود دقت به‌دلیل استفاده از تکنیک‌های مؤثر در انتخاب ویژگی و بهره‌گیری از قدرت

[DOI: 10.61882/ijbd.18.3.7]

با این حال، مانند هر روش محاسباتی پیشرفته‌ای، رویکرد پیشنهادی نیز با محدودیت‌هایی همراه است. یکی از مهم‌ترین چالش‌ها، پیچیدگی زمانی و محاسباتی بالا است. اجرای الگوریتم‌های فراابتکاری، به ویژه در ترکیب با روش‌هایWrapper ، نیازمند منابع محاسباتی قابل توجهی است. علاوه بر این، آموزش مدل‌های یادگیری عمیق نظیر CNNبه مجموعه‌داده‌های بزرگ و متوازن نیاز دارد. در محیط‌های بالینی، داده‌های زیستی اغلب با چالش عدم تعادل کلاس‌ها یا حجم محدود مواجه هستند که می‌تواند بر دقت و تعمیم‌پذیری مدل تأثیر بگذارد.

نکته دیگر، نوظهور بودن الگوریتم الهام‌گرفته از ساخت اهرام جیزه است. اگرچه این الگوریتم نوآورانه و مؤثر ظاهر شده است، اما عدم ارزیابی گسترده آن در ادبیات علمی می‌تواند سوالاتی در مورد قابلیت اطمینان، پایداری و عملکرد آن در شرایط مختلف ایجاد کند. برای افزایش اعتبار، مقایسه این الگوریتم با سایر الگوریتم‌های فراابتکاری شناخته‌شده مانند PSO، GA، یا DE ضروری به نظر می‌رسد.

در مقایسه با مطالعات پیشین که تمرکز بر پیش‌پردازش و انتخاب ویژگی با روش‌های سنتی مانند مجموعه‌های سخت و دسته‌بندهای پایه مانند kNN و SVM داشتند، روش حاضر به‌روزتر و پیشرفته‌تر است. همچنین برخلاف روش‌هایی که صرفاً بر انتخاب ویژگی آماری متکی هستند (مثل ضریب همبستگی پیرسون)، این پژوهش از قدرت اکتشافی الگوریتم‌های بهینه‌سازی فراابتکاری استفاده کرده که امکان یافتن ترکیب‌های بهتر از ویژگی‌ها را دارد.

در مقایسه با مطالعات مرتبط، می‌توان مشاهده کرد که اگرچه روش‌های دیگری نیز از ترکیب انتخاب ویژگی و دسته‌بندها استفاده کرده‌اند [۱۴]، اما مدل پیشنهادی ما با بهره‌گیری از یادگیری عمیق و الگوریتم ترکیبی فیلتر-لفاف (FWGPC)، دقت بالاتری ارائه داده و فراتر از ترکیب دسته‌بندهای کلاسیک عمل کرده است. در حالی که برخی پژوهش‌ها صرفاً بر قدرت شبکه‌های عصبی تمرکز دارند [۱۵]، مدل ما با داشتن یک ساختار انتخاب ویژگی پیچیده و چندلایه، توانسته است دقت نهایی را از بسیاری از مدل‌های صرفاً عمیق فراتر ببرد. همچنین، در مقایسه با روش‌هایی که از بهینه‌سازی ازدحام ذرات استفاده می‌کنند [۱۶]، الگوریتم FWGPC با طراحی رقابتی و چندجمعیتی، ظرفیت اکتشافی و استثماری بالاتری نشان داده است. این پژوهش، برخلاف روش‌هایی که صرفاً بر جستجوی فراگیر با الگوریتم‌های تکاملی تکیه دارند [۱۷]، با ترکیب دو تابع تناسب (فیلتر و لفاف) و ارزیابی عملکرد مبتنی بر شبکه عصبی، رویکردی هدفمندتر و دقیق‌تر را ارائه می‌دهد. حتی در مقایسه با رویکردهایی که از خودرمزگذارها برای کاهش ویژگی‌ها استفاده می‌کنند [۲۱]، مدل پیشنهادی ما با مرحله انتخاب ویژگی ترکیبی و طراحی الگوریتمی خاص، قابلیت بهینه‌سازی دقیق‌تری را فراهم کرده است.

نتیجه‌گیری

پژوهش حاضر با ارائه یک رویکرد نوین مبتنی بر تلفیق الگوریتم بهینه‌سازی الهام‌گرفته از ساخت اهرام جیزه و شبکه‌های عصبی کانولوشنی(CNN)، توانسته است دقت پیش‌بینی در تشخیص سرطان پستان را به‌طور قابل‌توجهی افزایش دهد. نتایج به‌دست‌آمده نشان‌دهنده پتانسیل بالای این روش در بهبود شناسایی ژن‌های مرتبط با انواع خاص سرطان پستان و کمک به انتخاب درمان‌های هدفمند و شخصی‌سازی شده برای بیماران است. این پیشرفت می‌تواند به‌عنوان یک ابزار مکمل در کنار تجربه بالینی پزشکان، نقش مؤثری در تشخیص زودهنگام و تصمیم‌گیری‌های درمانی ایفا کند. با این‌حال، برای بهره‌برداری عملی و بالینی از این مدل، لازم است مطالعات گسترده‌تری بر روی داده‌های واقعی، نویزی و متنوع از مراکز درمانی انجام شود. همچنین، کاهش پیچیدگی محاسباتی و مدیریت چالش‌هایی مانند عدم تعادل داده‌ها از اولویت‌های اصلی در ادامه مسیر توسعه این روش است. در نهایت، هدف این پژوهش، فراهم‌سازی بستری برای توسعه سامانه‌های هوشمند پشتیبان تصمیم‌گیری بالینی است که بتوانند کیفیت مراقبت از بیماران مبتلا به سرطان پستان را ارتقاء دهند

تعارض منافع

نویسنگان این مقاله هیچ‌گونه تعارض منافعی ندارند.

References

[1] Bissanum R, Chaichulee S, Kamolphiwong R, Navakanitworakul R, Kanokwiroon K. Molecular Classification Models for Triple Negative Breast Cancer Subtype Using Machine Learning. J Pers Med. 2021 Sep 1;11(9):881. doi: 10.3390/jpm11090881.

[2] M. Sugimoto, S. Hikichi, M. Takada, M. Toi. Machine learning techniques for breast cancer diagnosis and treatment: a narrative review, Annals of Breast Surgery. 2023; 7:1-13. doi:10.21037/abs-21-63.

[3] Abhisheka, B, Biswas, S.K. & Purkayastha, B. A Comprehensive Review on Breast Cancer Detection, Classification and Segmentation Using Deep Learning. Arch Computat Methods Eng. 2023;30: 5023–5052. doi:10.1007/s11831-023-09968-z

[4] Gupta S, Gupta MK, Shabaz M, Sharma A. Deep learning techniques for cancer classification using microarray gene expression data. Front Physiol. 2022;13:952709. doi: 10.3389/fphys.2022.952709.

[5] Moshood A. Hambali, Tinuke O. Oladele, Kayode S. Adewole. Microarray cancer feature selection: Review, challenges and research directions. International Journal of Cognitive Computing in Engineering. 2020;1:78-97.

[6] Turgut S, Dağtekin M, Ensari T.Microarray breast cancer data classification using

[DOI: 10.61882/ijbd.18.3.7]

machine learning methods. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018; 1-3, doi: 10.1109/EBBT.2018.8391468.
[7] Thalor A, Kumar Joon H, Singh G, Roy S, Gupta D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. Comput Struct Biotechnol J. 2022;20: 18-1631. doi: 10.1016/j.csbj.2022.03.019.
[8] Alhenawi E. a., Al-SayyedR, Hudaib A, Mirjalili S. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. Computers in biology and medicine. 2022; 140: 105051. doi:10.1016/j.compbiomed.2021.105051
[9] Bellarmino N, Cantoro R, Huch M, Kilian T, Schlichtmann U, Squillero G. Feature Selection for Cost Reduction In MCU Performance Screening. in 2023 IEEE 24th Latin American Test Symposium (LATS), 2023: 1-6. doi: 10.1109/LATS58125.2023.10154495
[10] Theng D, Bhoyar K. K,. Feature selection techniques for machine learning: a survey of more than two decades of research. Knowledge and Information Systems. 2024;66(3):1575-1637. doi: 10.1007/s10115-023-02010-5
[11]M. Z. Ali, A. Abdullah, A. M. Zaki, F. H. Rizk, M. M. Eid, E. M. El-Kenway. Advances and challenges in feature selection methods: a comprehensive review. J. Artif. Intell. Metaheuristics. 2024; 1: 67-77. doi: 10.54216/JAIM.070105
[12]K. Navin, H. K. Nehemiah, Y. Nancy Jane, and H. Veena Saroji. A classification framework using filter–wrapper based feature selection approach for the diagnosis of congenital heart failure. Journal of Intelligent & Fuzzy Systems. 2023; 44(4):6183-218. doi:10.3233/JIFS-230004
[13]Kalaiyarasi M, Rajaguru H. Performance analysis of ovarian cancer detection and classification for microarray gene data," *BioMed Research International,* 2022;2022(1):6750457. doi:10.1155/2022/6750457
[14]Patil S., Balmuri K. R., Frnda J., Parameshachari B., Konda S., Nedoma J,. Identification of Triple Negative Breast Cancer Genes Using Rough Set Based Feature Selection Algorithm & Ensemble Classifier. Human-centric computing and information sciences. 2022; 12. doi:10.22967/HCIS.2022.12.054
[15]S. H. Shah, M. J. Iqbal, I. Ahmad, S. Khan, and J. J. P. C. Rodrigues. Optimized Shah, S.H., Iqbal, M.J., Ahmad, I. *et al.* Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Comput & Applic* (2020). doi:10.1007/s00521-020-05367-8
[16] Alrefai, N., Ibrahim, O. Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Comput & Applic* . 2022;(34):13513–28. doi:10.1007/s00521-022-07147-y
[17]Othman M. S., Kumaran S. R., Yusuf L. M., Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data. IEEE Access. 2020;(8): 186348-61. doi:10.1109/ACCESS.2020.3029890
[18]H. Fathi, H. AlSalman, A. Gumaei, I. I. Manhrawy, A. G. Hussien, and P. El-Kafrawy, "An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data," Computational Intelligence and Neuroscience. 2021; 2021(1). 7231126. doi:10.1155/2021/7231126
[19]N. M. Ali, N. Aziz, and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection," Indones J Electr Eng Comput Sci.2020;20(2):712-9. doi:10.11591/ijeecs.v20.i2.pp712-719.
[20]Taghizadeh E., Heydarheydari S., Saberi A., JafarpoorNesheli S., Rezaeijo S. M., Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. BMC Bioinformatics. 2022;23(1):410. doi:10.1186/s12859-022-04965-8
[21]Gokhale M., Mohanty S. K., Ojha A., A stacked autoencoder based gene selection and cancer classification framework. Biomedical Signal Processing and Control. 2022;78: 103999. doi:10.1016/j.bspc.2022.103999
[22]Jiang Q. Jin M., Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression. (in English). Frontiers in Genetics, Original Research. 2021;12. doi:10.3389/fgene.2021.629946
[23]Li Q, Yang H, Wang P, Liu X, Lv K, Ye M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. J Transl Med. 2022;20(1):177. doi: 10.1186/s12967-022-03369-9.

[DOI: 10.61882/ijbd.18.3.7]

[24]Alromema N, Syed AH, Khan T. A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data. Diagnostics (Basel). 2023;13(4):708. doi: 10.3390/diagnostics13040708.
[25]Yaqoob, A., Verma, N.K. & Aziz, R.M. Optimizing Gene Selection and Cancer Classification with Hybrid Sine Cosine and Cuckoo Search Algorithm. Journal of Medical Systems, 2024;48(1):10 doi:10.1007/s10916-023-02188-3
[26]Kramer B., Peherstorfer B., Willcox K. E., Learning nonlinear reduced models from data with operator inference. Annual Review of Fluid Mechanics. 2024;56(1): 521-48. doi:10.1146/annurev-fluid-121021-025220
[27]Cong, S., Zhou, Y. A review of convolutional neural network architectures and their optimizations. Artificial Intelligence Review. 2023;56(3):1905-69. doi:10.1007/s10462-022-10213-5
[28] Corominas A. On deciding when to stop metaheuristics: Properties, rules and termination conditions. Operations Research Perspectives. 2023;10:100283. doi:10.1016/j.orp.2023.100283
[29]Karlupia N. Abrol P., Wrapper-based optimized feature selection using nature-inspired algorithms. Neural Computing and Applications. 2023;35(17):12675-89. doi:10.1007/s00542-023-08383-6.
[30]Nssibi M., Manita G., Korbaa O., Binary Giza pyramids construction for feature selection. Procedia Computer Science. 2021;192: 676-87. doi:10.1016/j.procs.2021.08.070
[31]Zhou H., X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient. Applied intelligence. 2022; 52(5):5457-74. doi:10.1007/s10489-021-02524-x
[32]Wang Z., Li M., Li J., A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure. Information Sciences. 2015;(307):73-88. doi:10.1016/j.ins.2015.02.031
[33]H. L. Xie, Jie; Jatkoe, Tim; Hatzis, Christos, Gene Expression Profiles of Breast Cancer. Mendeley Data. 2017;1: doi:10.17632/v3cc2p38hb.1.
[34]Singh K., Shastri S., Kumar S., Mansotra V., BC-Net: Early Diagnostics of Breast Cancer Using Nested Ensemble Technique of Machine Learning. Automatic Control and Computer Sciences. 2023;57(6):646-59. doi:10.3103/S0146411623060093

[DOI: 10.61882/ijbd.18.3.7]