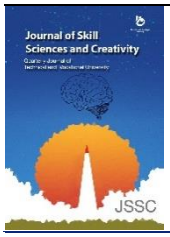


Quarterly Scientific Journal of Technical and Vocational University

Spring 2024, Vol. 1, No. 1, p. 13-34

Journal Homepage: https://jssc.tvu.ac.ir/?lang=en

TVU

doi:10.48301/JSSC.2024.190496

Early Detection of Breast Cancer by LSD Analysis and KNN Classification on MIAS Mammography Database

Farnaz Hoseini[1*], Hamed Sepehrzadeh[2], Masume Kheyri[3]

[1,2]Assistant Professor, Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran.
[3]The Coach, Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 10.17.2022
Revised: 12.11.2023
Accepted: 02.12.2024

Keyword:
Breast Cancer
Mammography Images
Gabor Wavelet
LSD Analysis
Feature Extraction
MIAS Database

***Corresponding Author:**
Farnaz Hoseini
Email: *f-hoseini@tvu.ac.ir*

ABSTRACT

Breast cancer is the most common cancer among women, particularly among women over 50 years old. Recent studies have proven that if breast cancer is diagnosed in the early stages of the formation of cancerous tissues, the chance of survival increases significantly and the costs of controlling the disease are greatly reduced. Therefore, the main solution is early detection of breast cancer. Until now, various research has been presented to diagnose breast cancer, but due to the selection of ineffective features and also the lack of using a suitable analytical method on the features, they could not achieve sufficient accuracy. In this study, LSD analysis and extraction of effective features by KNN classifier are used for automatic detection of breast cancer. The purpose of presenting the proposed method is to increase the accuracy of diagnosis for normal and non-normal classes. The proposed method was implemented in MATLAB environment and on the MIAS mammography image bank. The results obtained from the implementation output demonstrated the detection of breast cancer with 92% accuracy. The obtained results were compared with other methods, which shows the better performance of the proposed method in terms of accuracy criteria.

EXTENDED ABSTRACT

Introduction

Breast cancer is a type of disease that causes the cells of a part of the body to produce abnormally and excessively. Generally, the produced cells gather together and form a mass or tumour. Cancers can be of two types, benign or malignant, in which the cancer cells are fixed in the benign type, but in the malignant type, these cells are transferred to other parts of the body and enable the growth of cancer cells again. Breast cancer is one of the most common types of cancer that is seen mostly in women. To determine the level of risk of a cancerous mass using criteria such as the size of the mass and the extent of its spread and penetration in nearby organs, they determine the stage of progress of a mass. The direct relationship between the risk and the progress of a mass increases the importance of early diagnosis of the presence of a mass or the possibility of a mass occurring. In recent years, much research has been conducted on breast cancer detection in mammography images. Traditional artificial intelligence methods and statistical pattern recognition methods are not valid today, and the use of wavelet transformations can be more effective for recognition. These transformations are incompatible techniques for multiscale object representation. In addition, these techniques are popular in similar fields such as image processing, and their application in the field of breast cancer diagnosis is exemplary. According to the stated content, breast cancer is one of the most common cancers in women, and its timely diagnosis plays an important role in the continuation of life and its treatment; but even the most common diagnostic techniques such as mammography cannot detect 100%, so it is necessary to investigate more optimal diagnostic techniques. For this purpose, in this study, a method for early detection of breast cancer based on Gabor wavelet and LSD (least significant difference) analysis was presented, with the main objective of detecting breast cancer tumours in mammogram images with high accuracy.

Methodology

Breast cancer is one of the most common causes of death among women. As early detection of breast cancer increases the probability of survival, it is very important to develop a system with high accuracy output to detect suspicious masses in mammography images. Different methods have been suggested for breast cancer diagnosis for mammography images, but none of these methods have been able to accurately distinguish between the two classes of normal/abnormal and benign/malignant. In this study, an automatic method for detecting benign and malignant masses was developed, which can perform classification and estimation for normal/abnormal and benign/malignant classes with high accuracy compared to other methods. In this section, a method for early detection of breast cancer based on Gabor wavelet and LSD analysis on mammography images is studied. In the proposed method, first, the information (mammography images) is loaded from the database, and then the area of interest is extracted by selecting an image from the database. Applying the Gabor wavelet transformation on the interested area of the image, the features are extracted by considering the coefficients resulting from the transformation. The obtained features are

entered in the features table and then by using LSD analysis on the features table, the effective features are selected, and finally, by applying the KNN classifier, as shown in part (c) of Figure 1, the accuracy of the normal class classification/ abnormal is determined. 330 mammography images were considered in the proposed method.

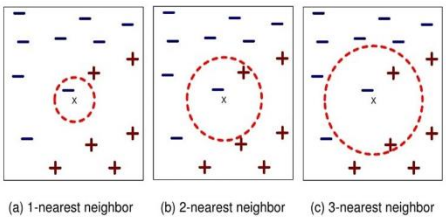


Figure 1. KNN Classifier.

In the database of images, in addition to mammography images, there was information for each image that indicated whether there was a complication or mass in the image in question. Therefore, based on the available information, if the input image was normal, a piece of 100 x 100 pixels was extracted from the middle of the image, and if the image was abnormal, a segment of 100 x 100 pixels was extracted (extraction of the area of interest). Therefore, for each image (normal and non-normal), a piece of 100 x 100 pixels was extracted from the database. Thus, all 330 mammography images became images with dimensions of 100 x 100 pixels.

Results and discussion

The proposed method was implemented with the help of MATLAB software version 2019. In the obtained results, the detection accuracy in the breast cancer training dataset was 92% for the normal/abnormal class and 81% for the benign/malignant class, which indicates the high efficiency of the proposed method. Compared to other breast cancer detection methods, there is a significant improvement in the three parameters of accuracy, sensitivity, and specificity. To implement the proposed method, the set of mammography images from the MIAS dataset was used. To evaluate the presented method, the desired method was implemented many times to determine the accuracy level in the detection of samples at each stage. Figure 2 shows some examples of Regions of Interest (RoI) from the MIAS dataset used in the implementation.

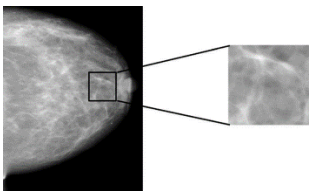


Figure 2. An example of RoI from the MIAS dataset.

These areas of interest belong to different types and severity of pathologies. As stated in the proposed method, the dimensions of these samples were 100x100 pixels, which were extracted from mammography images in the database, and by extracting these pieces, the

proposed method was performed by applying LSD analysis and KNN classification. Table 1 shows the comparison of the values of evaluation criteria for N/A and N/B classes in a diagram. According to the stated content, it can be concluded that the proposed method can meet the objectives of increasing the accuracy of classification and diagnosis for normal/abnormal and benign/malignant classes. According to the obtained results, it can be seen that the proposed method for classifying the normal/abnormal class performed better than the benign/malignant class. The accuracy criterion for the normal/abnormal class was 92%, and this value for the benign/malignant class was 81%.

Table 1. Comparison of evaluation criteria values for N/A and B/M classes by the proposed method.

Output Class / Evaluation criteria	N/A	B/M
Accuracy	92 %	81 %
Specificity	99 %	67 %
Recall	93 %	82 %
Precision	77 %	84 %
F1-Score	84 %	64 %
PPV	98 %	91 %

Conclusion

In this research, considering the necessity of early and timely diagnosis of this disease, a new method based on LSD analysis and KNN classification on mammography was presented. The proposed method was implemented using MATLAB software and the results showed that the proposed method was very effective in reducing human errors in detecting normal and abnormal masses in images with an accuracy of 92%. Many images received from the MIAS mammography database were analyzed by the presented model, the results of which are very acceptable, and in terms of various criteria, they are higher than the models presented in the authoritative papers.

فصلنامه علمی دانشگاه فنی و حرفه‌ای
بهار ۱۴۰۳، دوره ۱، شماره ۱، ۳۴-۱۳
آدرس نشریه: https://jssc.tvu.ac.ir/
doi:10.48301/JSSC.2024.190496

تشخیص زودهنگام سرطان سینه با تحلیل LSD و طبقه‌بند KNN بر روی بانک داده ماموگرافیِ MIAS

فرناز حسینی[1*]، حامد سپهرزاده[2]، معصومه خیری[3]

۱ و ۲- استادیار، گروه مهندسی کامپیوتر، دانشگاه فنی و حرفه‌ای، تهران، ایران.
۳- مربی، گروه مهندسی کامپیوتر، دانشگاه فنی و حرفه‌ای، تهران، ایران.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۲/۰۲/۲۷
بازنگری مقاله: ۱۴۰۲/۰۹/۲۰
پذیرش مقاله: ۱۴۰۲/۱۱/۲۳

کلید واژگان:
سرطان سینه
تصاویر ماموگرافی
موجک گابور
تحلیل LSD
استخراج ویژگی
بانک داده MIAS

***نویسنده مسئول:** فرناز حسینی
پست الکترونیکی:
f-hoseini@tvu.ac.ir

چکیده

نظریهٔ سرطان سینه، رایج‌ترین سرطان در میان زنان به‌خصوص در بین زنان بالای ۵۰ سال است. مطالعات اخیر ثابت کرده‌اند که اگر سرطان سینه در مراحل اولیه تشکیل بافت‌های سرطانی تشخیص داده شود، احتمال حیات به‌طور قابل‌توجهی افزایش یافته و هزینه‌های ناشی از کنترل بیماری به شدت کاهش یافته است بنابراین راه‌حل اصلی، شناسایی زودهنگام سرطان سینه است. تاکنون پژوهش‌های مختلفی برای تشخیص سرطان سینه ارائه شده است اما به دلیل انتخاب ویژگی‌های غیرمؤثر و همچنین استفاده‌نکردن از یک روش تحلیلی مناسب بر روی ویژگی‌ها نتوانستند به دقت کافی برسند. در این مطالعه از تحلیل LSD و استخراج ویژگی‌های مؤثر توسط طبقه‌بند KNN برای تشخیص خودکار سرطان سینه استفاده شده است. هدف از ارائه روش پیشنهادی، افزایش دقت تشخیص برای کلاس‌های نرمال و غیرنرمال می‌باشد. روش پیشنهادی در محیط متلب و بر روی بانک تصاویر ماموگرافیِ MIAS اجرا شده است. نتایج حاصل از خروجی پیاده‌سازی، بیانگر شناسایی سرطان سینه با دقت ۹۲ درصد است. نتایج به‌دست‌آمده با سایر روش‌ها مقایسه شدند که نشان‌دهنده عملکرد بهتر روش پیشنهادی از نظر معیار دقت می‌باشد.

مقدمه

سرطان سینه، نوعی بیماری است که باعث می‌شود سلول‌های قسمتی از بدن به‌طور غیرعادی و بیش‌ازحد تولید شوند (ایوب‌زاده و همکاران، ۲۰۲۳). عموماً سلول‌های تولیدشده در کنار هم جمع می‌شوند و یک توده یا غده را تشکیل می‌دهند. سرطان‌ها می‌توانند از دو نوع خوش‌خیم یا بدخیم باشند که در نوع خوش‌خیم، سلول‌های سرطانی ثابت هستند ولی در نوع بدخیم، این سلول‌ها به قسمت‌های دیگر بدن منتقل می‌شوند و امکان رشد دوباره سلول‌های سرطانی را فراهم می‌سازند (اسکوبر-لینرو و همکاران، ۲۰۲۳)[1]. سرطان سینه یا سرطان پستان یکی از انواع شایع سرطان است که بیشتر در زنان دیده می‌شود (ژنگ و همکاران، ۲۰۲۳)[2]. برای تعیین میزان خطر یک توده سرطانی با استفاده از معیارهایی مانند اندازه توده و مقدار پخش‌شدن و نفوذ آن در ارگان‌های مجاور، مرحله پیشرفت یک توده را تعیین می‌کنند (کامپنلا و همکاران، ۲۰۲۳)[3]. ارتباط مستقیم خطر با میزان پیشرفت یک توده اهمیت تشخیص زودهنگام وجود توده یا احتمال به‌وجودآمدن توده را زیاد می‌کند (چن و همکاران، ۲۰۲۳)[4]. یکی از فاکتورهای مهم در تعیین احتمال ابتلا به سرطان سینه، سن است. علاوه بر سن عوامل دیگری مانند نژاد، محل جغرافیایی زندگی، شرایط زندگی، سوابق خانوادگی و غیره نیز در تعیین احتمال ابتلا به سرطان سینه مؤثر هستند. آمار نشان می‌دهد که در آمریکا از هر ۸ زن یک نفر و در اروپا از هر ۶ زن یک نفر به این بیماری مبتلا می‌شود (یاری و همکاران، ۲۰۲۰)[5]. با وجود افزایش این بیماری، آمار کاهش میزان مرگ‌ومیر ناشی از این بیماری نشان می‌دهد که احتمالاً به‌وجودآمدن روش‌های جدید درمانی و روش‌های جدید تشخیص مانند سیستم‌های ماموگرافی توانسته کمک شایانی در روند بهبود این بیماری انجام دهد (گاردزی و همکاران، ۲۰۱۴)[6]. ماموگرافی نوعی از تصویربرداری پزشکی است که از اشعه X با دز پایین استفاده می‌کند و تصویر خروجی آن بافت‌های درونی سینه را نمایش می‌دهد (تسوچاتزیدیس و همکاران، ۲۰۱۹)[7]. با گسترش استفاده از ماموگرافی امکان ذخیره‌سازی، مشاهده و آنالیز تصاویر ماموگرافی به شکل الکترونیکی فراهم شد. با استفاده از تصاویر ماموگرافی می‌توان ناهنجاری‌های مختلفی مانند سرطان سینه را شناسایی کرد (کویتا و همکاران، ۲۰۲۲)[8]. تصاویر ماموگرافی مانند بسیاری دیگر از تصاویر پزشکی به علت ویژگی‌های خاصی که دارند به‌راحتی قابل‌تحلیل نیستند. همچنین در کارهای تشخیص پزشکی با استفاده از سیستم‌های کامپیوتری و هوش مصنوعی، ویژگی‌های متنوع و متعددی استخراج می‌شوند (زوندرلند و همکاران، ۱۹۹۹)[9]. بدیهی است که از بین این ویژگی‌ها تعدادی از آن‌ها ممکن است جزء ویژگی‌های مازاد یا نامربوط محسوب شوند و بدین ترتیب سبب افت دقت سیستم شوند. به همین دلیل لازم است روشی ارائه شود که علاوه بر تسریع در امر تشخیص، دقت و قابلیت اطمینان را نیز بهبود بخشد (آکای، ۲۰۰۹)[10]. در سال‌های اخیر تحقیقات زیادی در زمینه تشخیص سرطان سینه در تصاویر ماموگرافی صورت گرفته است (چن و همکاران، ۲۰۱۱)[11]. روش‌های هوش مصنوعی سنتی و روش‌های آماری تشخیص الگو امروزه معتبر نیستند و استفاده از تبدیلات موجک می‌تواند برای تشخیص مؤثرتر باشد. این تبدیلات، تکنیک‌های تطابق‌ناپذیری برای نمایش چندمقیاسی شیء هستند. همچنین این تکنیک‌ها

[1] Escobar-Linero
[2] Zheng
[3] Campanella
[4] Chan
[5] Yari
[6] Gardezi
[7] Tsochatzidis
[8] Kavitha
[9] Zonderland
[10] Akay
[11] Chen

در زمینه‌های مشابهی ازجمله پردازش تصویر دارای محبوبیت هستند و کاربرد آن در زمینه تشخیص سرطان سینه مثال‌زدنی است.

با توجه به مطالب بیان‌شده، سرطان سینه جزء سرطان‌های شایع زنان است که تشخیص به‌موقع آن در ادامه حیات و درمان آن، نقش مهمی دارد اما حتی متداول‌ترین تکنیک‌های تشخیصی مثل ماموگرافی نیز نمی‌توانند قابلیت تشخیص صد درصدی داشته باشند بنابراین، ضرورت دارد تکنیک‌های تشخیص بهینه‌تری بررسی شوند. برای این منظور، در این مطالعه یک روش برای تشخیص زودهنگام سرطان سینه مبتنی بر موجک گابور و تحلیل LSD[1] (کمترین تفاوت معنی‌دار) ارائه شده است که هدف اصلی آن تشخیص تومورهای سرطان سینه در تصاویر ماموگرام با دقت بالا می‌باشد. ساختار این مقاله به این شکل است که در بخش دوم تکنیک‌ها و روش‌های مختلف در تشخیص سرطان سینه مربوط به سال‌های اخیر مرور شده است. در بخش سوم روش پیشنهادی مورد بحث قرار گرفته است. در بخش چهارم نتایج به‌دست‌آمده از روش پیشنهادی ارائه شده است. نتیجه‌گیری و پیشنهادهای آتی نیز در بخش پنجم ارائه شده است.

پیشینه تحقیق

در این بخش، مطالعات اخیر انجام‌گرفته در حیطه موضوع موردمطالعه در سال‌های اخیر مورد بحث و بررسی قرار می‌گیرد. با توجه به نتایج به‌دست‌آمده در تمامی روش‌ها مجموعه داده تصاویر مجموعه داده (میدر، ۲۰۱۷)[2] بوده است. در (آوجی و کاراکایا، ۲۰۲۳)[3] از ترکیب بهینه الگوریتم‌های پیش‌پردازش مختلف برای امکان تفسیر و طبقه‌بندی بهتر تصاویر ماموگرافی استفاده شده است زیرا الگوریتم‌های پیش‌پردازش به‌طور قابل‌توجهی بر دقت روش‌های تقسیم‌بندی و طبقه‌بندی تأثیر می‌گذارند. در این مطالعه، تأثیر ترکیب روش‌های مختلف پیش‌پردازش در افتراق ضایعات خوش‌خیم و بدخیم پستان بررسی شده است. تمام الگوریتم‌های پردازش تصویر مورداستفاده برای تشخیص ضایعه از مجموعه داده MIAS استفاده کردند. در مرحله اول، اطلاعات برچسب و عضله سینه‌ای حاصل از گرفتن تصاویر ماموگرافی حذف شده، در مرحله دوم، فیلتر میانه، تساوی هیستوگرام تطبیقی محدود کنتراست و الگوریتم‌های پوشش غیرشارپ با ترکیب‌های مختلف وضوح و دید تصاویر افزایش می‌یابد. در مرحله سوم، مناطق مشکوک با استفاده از تکنیک خوشه‌بندی k-means از ماموگرافی استخراج می‌شوند. در نهایت، مجموعه داده‌های ویژگی با استفاده از الگوریتم‌های یادگیری ماشین به‌عنوان نرمال/غیر طبیعی و خوش‌خیم/بدخیم (طبقه‌بندی دو طبقه) طبقه‌بندی می‌شوند. در (جستی و همکاران، ۲۰۲۲)[4] یک رویکرد تکاملی برای طبقه‌بندی و تشخیص سرطان سینه مورد بحث قرار گرفته که مبتنی بر یادگیری ماشین و پردازش تصویر است. این مدل تکنیک‌های پیش پردازش تصویر، استخراج ویژگی، انتخاب ویژگی و تکنیک‌های یادگیری ماشینی را برای کمک به طبقه بندی و شناسایی بیماری‌های پوستی ترکیب می‌کند. در این مطالعه برای افزایش کیفیت تصویر از فیلتر میانگین هندسی استفاده شده است. AlexNet برای استخراج ویژگی‌ها استفاده شده است. برای طبقه‌بندی و تشخیص بیماری، این مدل از تکنیک‌های یادگیری ماشین مانند ماشین بردار پشتیبان، KNN، جنگل تصادفی و بیزین ساده استفاده کرده. تحقیقات تجربی بر روی داده‌های MIAS اعتبارسنجی شده است. این فناوری پیشنهادی برای شناسایی دقیق بیماری سرطان پستان با استفاده از تجزیه و تحلیل تصویر سودمند است. در (الیامی و همکاران، ۲۰۲۲)[5] تلفیقی از معماری AlexNet و ویژگی‌های GLCM [6] (ماتریس همزمانی سطح خاکستری) برای استخراج ویژگی‌های بافت قابل تشخیص از بافت‌های سینه استفاده شد. در نهایت، برای دستیابی به دقت بالاتر، از مجموعه‌ای از MK-SVM

[1] Lysergic Acid Diethylamide
[2] Mader
[3] Avcı & Karakaya
[4] Jasti
[5] Alyami
[6] Gray-level Cooccurrence Matrix

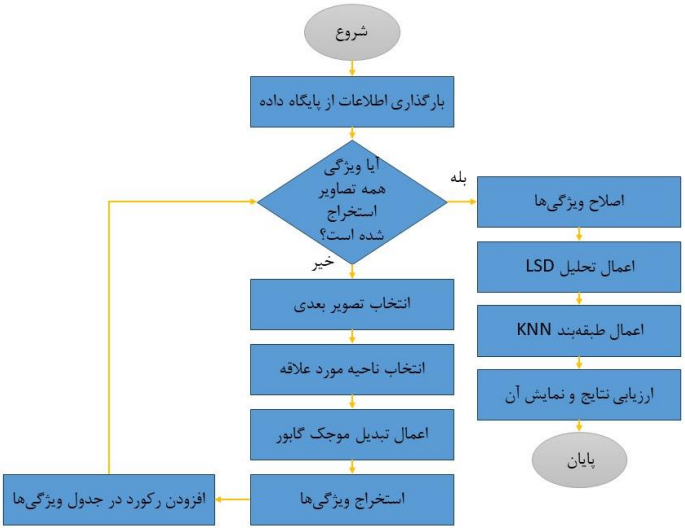
استفاده شده و برای اهداف آزمایشی، مدل پیشنهادی به مجموعه داده MIAS، که یک مجموعه داده تصویر پستان که معمولاً زیاد مورد استفاده قرار می‌گیرد، اعمال شده است. در (تینگ و همکاران، ۲۰۱۹)[1] یک شبکه عصبی کانولوشنی بهبودیافته برای طبقه‌بندی توده‌ها در سرطان پستان (CNNI-BCC) ارائه شده است. روش پیشنهادی از شبکه عصبی کانولوشنی استفاده می‌کند که طبقه‌بندی ضایعات سرطان سینه را بهبود ببخشد تا به متخصصان در تشخیص سرطان سینه کمک کند. در واقع هدف از روشِ ارائه‌شده، کمک به متخصصان حوزه پزشکی برای طبقه‌بندی ضایعات سرطان پستان از طریق اجرای شبکه عصبی کانولوشنی برای طبقه‌بندی سرطان پستان بوده است. نتایج تجربی نشان می‌دهد که روش ارائه‌شده توانسته دقت ۹۰/۵۰ درصد را بر روی مجموعه داده MIAS (انجمن آنالیز عکس ماموگرافیک) برای تخشیص توده‌های خوش خیم/بدخیم به‌دست آورد. در روش PCA-CC (التوخی و همکاران، ۲۰۱۴)[2] و روش پیشنهادی ابعاد تصاویر ناحیه موردعلاقه استخراج‌شده ۲۰۰*۲۰۰ پیکسل است که این ابعاد برای سایر روش‌ها ۱۲۸*۱۲۸ پیکسل است. در روش COCC (گدیک و آتسوی، ۲۰۱۳)[3] روش طبقه‌بندی استفاده‌شده طبقه‌بند لجستیک ساده است که برای روش‌های SCC و PCA-CC (گاردزی و همکاران، ۲۰۱۹)[4] ماشین بردار پشتیبان (SVM)[5] و برای روش پیشنهادی، طبقه‌بند نیو بیزین (NB)[6] درنظر گرفته شده است. تعداد تصاویر استفاده‌شده برای هریک از روش‌ها متفاوت است که این تعداد برای روش پیشنهادی ۳۳۰ و برای روش‌های CNNI-BCC، SCC، PCA-CC و COCC به‌ترتیب ۳۰۰، ۳۰۷، ۲۰۰ و ۳۰۵ تصویر ماموگرافی است. در (لی و همکاران، ۲۰۰۱)[7] روشی برای تشخیص توده‌ها، ۳۲۰ تصویر از مجموعه داده MIAS در انگلستان انجام شد. الگوریتم حاضر برای تشخیص توده‌ها بر اساس یک تکنیک آستانه‌سازی تطبیقی استاندارد توسعه داده شده بود. با این حال، نتایج اولیه به‌خوبی نتایج یک پایگاه داده ژاپنی نبود. در این مطالعه، تفاوت‌های بین MIAS و یک مجموعه داده ژاپنی نیز مورد بحث قرار گرفته است.

روش پیشنهادی

سرطان سینه به‌عنوان یکی از شایع‌ترین علل مرگ‌ومیر در میان زنان است. همان‌طور که تشخیص زودهنگام سرطان سینه احتمال زنده‌ماندن را افزایش می‌دهد، ایجاد یک سیستم با خروجی دقت بالا برای تشخیص توده‌های مشکوک در تصاویر ماموگرافی بسیار مهم است. روش‌های مختلفی برای تشخیص سرطان سینه برای تصاویر ماموگرافی پیشنهاد شده اند اما هیچ‌کدام از این روش‌های نتوانسته‌اند تشخیص دقیقی از دو کلاس نرمال/غیرنرمال و خوش‌خیم/ بدخیم ارائه دهند. در این مطالعه، روشی خودکار برای تشخیص انواع توده خوش‌خیم و بدخیم شده است که می‌تواند با دقت بالایی طبقه‌بندی و تخمین برای کلاس‌های نرمال/غیرنرمال و خوش‌خیم/بدخیم نسبت به سایر روش‌های مقایسه‌شده را انجام دهد. در این بخش از روشی برای تشخیص زودهنگام سرطان سینه مبتنی بر موجک گابور و تحلیل LSD بر روی تصاویر ماموگرافی مطالعه می‌شود. در روش پیشنهادی مطابق آنچه در فلوچارت شکل ۱ نشان داده شده، ابتدا بارگذاری اطلاعات (تصاویر ماموگرافی) از پایگاه داده انجام می‌گردد، سپس با انتخاب یک تصویر از پایگاه داده ناحیه موردعلاقه استخراج می‌شود. با اعمال تبدیل موجک گابور بر روی ناحیه موردعلاقه تصویر، استخراج ویژگی‌ها با درنظر گرفتن ضرایب حاصل از تبدیل صورت می‌گیرد. ویژگی‌های به‌دست‌آمده در جدول ویژگی‌ها درج می‌شوند و در ادامه با استفاده از تحلیل LSD بر روی جدول ویژگی، ویژگی‌های مؤثر انتخاب می‌گردد و در آخر کار با اعمال طبقه‌بند KNN دقت طبقه‌بند کلاس

[1] Ting
[2] Eltoukhy
[3] Gedik & Atasoy
[4] Gardezi
[5] Support Vector Machine
[6] Naive Bayes
[7] Li

نرمال/غیرنرمال مشخص می‌شود. روش‌های یادگیری بدون نظارت، مانند الگوریتم KNN در مرحله یادگیری خود به هیچ دانش ازقبل تعیین‌شده‌ای نیاز ندارند. روش‌های بدون نظارت به‌عنوان روش‌های خوشه‌بندی شناخته شده‌اند. این روش‌ها به یک مجموعه داده آموزش با برچسب نیاز ندارد، گروه‌های نرمال (خوشه‌ها) را در داده‌ها کشف می‌کنند و روی کشف الگوهای موجود در داده‌ها تمرکز دارند. نمونه‌ها را بر اساس میزان شباهت ویژگی‌های آنها که توسط یک رویکرد اندازه‌گیری فاصله تعریف می‌شود مانند فاصله اقلیدسی به گروه‌ها خوشه‌بندی می‌کنند. در انتخاب مقدار K با افزایش مقدار K مرز بین دو کلاس هموارتر می‌شود. برای تعیین میزان درست K دو پارامتر خطای آموزش و خطای اعتبار سنجی را بررسی کردیم. میزان خطای آموزش به مقداری گفته می‌شود که در سیستم بعد از اینکه با داده‌های آموزش، مراحل یادگیری را می‌گذراند با همان داده‌ها یک مرحله آزمون نیز انجام می‌شود. اگر هنگام آزمون با داده‌های آموزش سیستم دچار خطا شود به آن مقدار خطای آموزش گفته می‌شود. مقدار خطای اعتبارسنجی به میزان خطای سیستم در زمان آزمون سیستم با داده‌های جدید یا به عبارتی داده‌های آزمون گفته می‌شود. با بررسی‌های انجام‌شده، بهترین مقدار برای K در این مطالعه ۹ درنظر گرفته شد.



شکل ۱. فلوچارت فرایند کلی روش پیشنهادی.

در روش پیشنهادی تعداد ۳۳۰ تصویر ماموگرافی درنظر گرفته شده است. در پایگاه داده تصاویر علاوه بر تصاویر ماموگرافی اطلاعاتی برای هر تصویر وجود دارد که نشان‌دهنده این است که در تصویر موردنظر عارضه یا توده وجود دارد یا نه. بنابراین براساس اطلاعات موجود، اگر تصویر ورودی نرمال باشد از وسط تصویر و اگر تصویر غیرنرمال باشد از محل عارضه قطعه‌ای به ابعاد ۱۰۰ * ۱۰۰ پیکسل استخراج می‌شود (استخراج ناحیه موردعلاقه). بنابراین برای هر تصویر (نرمال و غیرنرمال) از پایگاه داده یک قطعه به ابعاد ۱۰۰*۱۰۰ پیکسل استخراج می‌شود. پس تمامی ۳۳۰ تصویر ماموگرافی تبدیل به تصاویری با ابعاد ۱۰۰*۱۰۰ پیکسل می‌شوند. شکل ۲ نمونه‌ای از تصاویر ماموگرافی با ابعاد ۱۰۰*۱۰۰ پیکسل که از تصویر اصلی استخراج شده‌اند را نمایش می‌دهد.

شکل ۱. نمونه تصاویر استخراج‌شده از پایگاه داده که به قطعه‌های با ابعاد ۱۰۰*۱۰۰ تبدیل شده‌اند تصویر سمت راست) تصویر نرمال / تصویر سمت چپ) تصویر غیرنرمال.

ویژگی‌ها خصوصیات شی‌هایی هستند که به‌عنوان ورودی طبقه‌بندی‌کننده، به‌کار می‌روند و طبقه‌های مختلفی را تشکیل می‌دهند. ویژگی یک شی در واقع یک الگوی ورودی را از الگوی دیگر تفکیک می‌کند. بنابراین استخراج ویژگی‌ها یکی از پرکاربردترین و چالش‌برانگیزترین زمینه‌ها برای بهبود نتایج طبقه‌بندی داده‌های مختلف به‌حساب می‌آید. هنگام حل بسیاری از مشکلات در یادگیری ماشین، ویژگی‌های ورودی بسیار زیادی وجود دارند. با این حال همه آن ویژگی‌ها برای حل مشکل مناسب نیستند و در بسیاری از موارد، انتخاب ویژگی‌های بی‌ربط باعث ازبین‌رفتن کارایی مدل آموزشی می‌شود. استخراج و انتخاب ویژگی، یکی از مراحل حساس و مهم در ایجاد یک مدل آموزشی مناسب می‌باشد. در روش پیشنهادی، پس از استخراج ناحیه موردعلاقه، دو مرحله استخراج ویژگی یکی برای خود تصویر (تصویر موردعلاقه) و دیگری ضرایب حاصل از موجک گابور انجام می‌شود. تعداد ۱۳ ویژگی (انرژی، آنتروپی، اینرسی و گشتاورهای μ, M, K, S) برای هر مورد استخراج می‌شود که توسط عبارت‌های (۱) تا (۴) نمایش داده شده است.

محاسبه M_p در عبارت (۱) نمایش داده‌شده که در آن $X = (X_i)_{1\le i\le N}$ یک توزیع از N ضرایب تبدیل است.

$$M_p = \frac{1}{N}\sum_{i=1}^{N}(X_i)^p \qquad p = 1, 2, 3, 4 \tag{۱}$$

محاسبه μ_p در عبارت (۲) نمایش داده شده که میانگین μ_p نشان‌دهنده برآورد از محل که در آن مرکز خوشه‌بندی رخ می‌دهد و برابر است با اولین مومنت.

$$\mu_p = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^p \qquad p = 1, 2, 3\ 4 \tag{۲}$$

محاسبه K در عبارت (۳) نمایش داده شده که Kurtosis ضریب کشیدگی نسبی یا همواری توزیع برای یک را به‌طور نرمال اندازه‌گیری می‌کند.

$$kurtosis = \frac{\mu_4}{(\mu_2)^2} - 3 \tag{۳}$$

محاسبه S در عبارت (۴) نمایش داده‌شده که Skewness نشان‌دهنده درجه نداشتن تقارن توزیع می‌باشد.

$$skewness = \frac{\mu_3}{(\mu_2)^{3/2}} \quad (۴)$$

براساس ویژگی‌های به‌دست‌آمده جدولی به نام بردار ویژگی ایجاد می‌شود (جدول ۱) که تعداد سطرهای این بردار ۳۳۰ سطر است که مطابق با تعداد تصاویر موجود در پایگاه داده است. تعداد ۲۶ ستون با توجه به تعداد کل ویژگی‌های استخراج‌شده برای هر تصویر درنظر گرفته می‌شود. دو ستون آخر بردار موردنظر مربوط به T_1 (نرمال/غیرنرمال) و T_2 (خوش‌خیم/بدخیم) می‌باشد. جدول ۱ نشان‌دهنده تعداد ویژگی‌ها می‌باشد، نه مقادیر ویژگی‌ها، از این رو نمایش بصری مربوط به تعداد ویژگی‌ها، تعداد تصاویر و کلاس‌های خروجی در این جدول نمایش داده شده است.

جدول ۱. ویژگی‌های تصاویر با کلاس‌های نرمال/غیرنرمال و خوش‌خیم/بدخیم.

ویژگی / تصویر	۱	۲	۳	...	۲۶	T_1	T_2
تصویر ۱				...			
تصویر ۲				...			
تصویر ۳				...			
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
تصویر ۳۳۰							

پس از ایجاد بردار ویژگی، مرحله مربوط به انتخاب ویژگی انجام می‌شود تا بررسی شود که آیا نیاز به کاهش ویژگی است یا نه. مسئله انتخاب ویژگی در واقع برگزیدن ویژگی‌هایی است که حداکثر توان را در پیشگویی خروجی دارا باشند. در روش پیشنهادی از تحلیل LSD جهت کاهش ویژگی استفاده شده است (به عبارتی با تحلیل LSD بررسی می‌شود که کدام ویژگی حذف شود بهتر است). بنابراین با انجام تحلیل LSD بر روی بردار ویژگی، اگر نیاز به کاهش ویژگی باشد با استفاده از روش t-test حذف ویژگی صورت می‌گیرد. معادله LSD به‌صورت (۵) محاسبه می‌شود:

$$\mathrm{LSD} = \frac{\mathrm{t}(\mathrm{S}\sqrt{2})}{\sqrt{\mathrm{n}}} \quad (۵)$$

در معادله ۵، n تعداد ویژگی و مقدار t براساس دو پارامتر (α و df) به‌دست می‌آید. این نکته را باید درنظر گرفت که برای کاربردهای مختلف هر دو مقادیر معلوم هستند. طبق جدول توزیع در ماموگرافی مقادیر پارامترهای $\alpha = 0.01$ و $\mathrm{df} = 10\%$ است و با درنظرگرفتن این مقادیر، مقدار t=۳.۱۶۹ به‌دست می‌آید. مقدار پارامتر S به‌صورت معادله (۶) به‌دست می‌آید:

$$S = \sqrt{MSE}$$

$$MSE = \frac{SSE}{(n-1)(v-1)} \quad (۶)$$

$$SSE = SS - SSB - SST$$

در معادله (۶) n برابر تعداد ویژگی و v تعداد رکورد است و مجموعه مقادیر مربوط به متغیر SSE به شکلی که در عبارت (۷) نشان داده شده است محاسبه می‌شود:

$$SS = \text{مجموع مربعات کل مقادیر} - \frac{(gt)^2}{\text{تعداد رکوردها}}$$

$$SSB = \frac{\text{مجموع مربعات مجموع ویژگی ها}}{v} - \frac{(gt)^2}{\text{تعداد کل مقادیر}} \quad (۷)$$

$$SST = \frac{\text{مجموع مربعات مجموع رکوردها}}{\text{تعداد ویژگی}} - \frac{(gt)^2}{\text{تعداد کل رکوردها}}$$

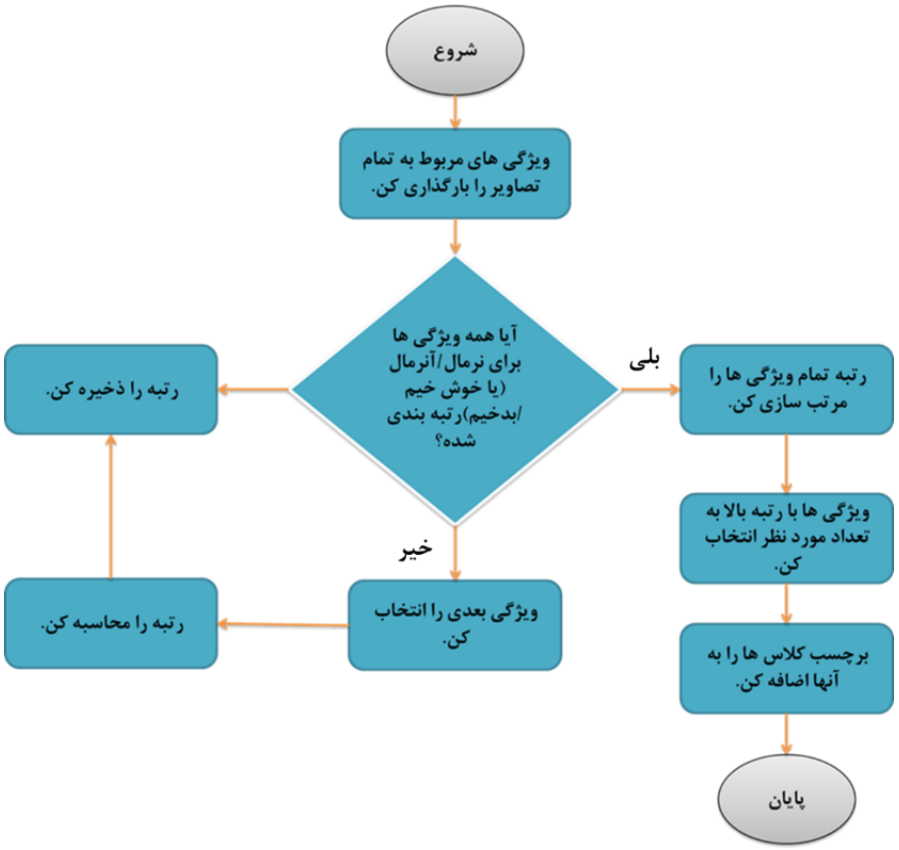
به‌منظور بیان دقیق محاسبه S با در نظر گرفتن شکل ۳ در قالب یک مثال توضیح داده می‌شود.

	X1	X2	X3	X4
R1	2	3	5	1
R2	3	2	1	5
R3	2	1	3	3

شکل ۲. مثالی برای نمایش نحوه محاسبه دقیق پارامتر S.

در شکل (۳) X بیان‌کننده تعداد ویژگی و R بیان‌کننده تعداد رکورد است (X=۴ و R=۳). با توجه به مثال درنظر گرفته‌شده، مجموع مربعات کل مقادیر، بیان‌کننده مجموع تمام مقادیری هستند که به توان دو رسیده است. معیار gt برابر جمع مقادیر هر سطر و ستون است (که در مثال مقدار ۶۲ است). مجموع مربعات مجموع ویژگی‌ها برابر مجموع ویژگی‌هایی که به توان ۲ می‌رسد و جمع می‌شوند و همچنین مجموع مربعات مجموع رکوردها برابر مجموع رکوردهایی که به توان ۲ می‌رسد و جمع می‌شوند. با مشخص‌شدن مقدار LSD، در مرحله بعد تمامی ویژگی‌های هر سطر بردار ویژگی را جمع می‌کند و مقادیر به‌دست‌آمده به‌صورت صعودی مرتب می‌شود. در ادامه با بررسی اینکه اختلاف دو مقدار پشت سرهم از LSD بیشتر است یا نه، اگر اختلافی پیدا نشود که از مقدار LSD کمتر باشد ویژگی‌ها نیاز به کاهش ندارد و لازم نیست حذف شوند اما اگر اختلافی پیدا شود که از مقدار LSD بیشتر باشد بنابراین یک ویژگی از بردار ویژگی حذف می‌شود همان‌طور که پیش‌تر بیان شد برای کاهش ویژگی (حذف ویژگی) از روش t-test استفاده می‌شود

بنابراین برای تمامی ویژگی‌های بردار مطابق مراحل نمایش داده‌شده در شکل ۴ روش t-test اعمال می‌شود و با رتبه‌بندی هر ویژگی، ویژگی‌ای که کمترین رتبه را داشته باشد از بردار حذف می‌شود. با حذف ویژگی، دوباره فرایند LSD برای بردار ویژگی کاهش‌یافته اعمال می‌گردد و تحلیل می‌شود که آیا ویژگی‌ای وجود دارد که باید حذف شود یا نه؛ بنابراین این نکته را باید درنظر گرفت که در هر مرحله یک ویژگی حذف می‌شود. با حذف ویژگی‌های غیرمؤثر، بردار ویژگی نهایی برای اعمال طبقه‌بندی درنظر گرفته می‌شود.



شکل ۴. فلوچارت مربوط به روش T-test و محاسبه بردار ویژگی.

برای توضیح دقیق انتخاب ویژگی‌های بهتر و حذف ویژگی غیرمؤثر جدول ۲ برای مثال (تعداد سه ویژگی و کلاس نرمال/غیرنرمال) را درنظر می‌گیرد. در این جدول برای هر ویژگی مشخص می‌شود که این ویژگی برای جداکردن بافت نرمال از غیرنرمال یا خوش‌خیم از بدخیم چقدر ارزش دارد. با درنظرگرفتن ستون آخر (نرمال/غیر نرمال با مقدار صفر یا یک) معادله (۸) را برای هر ستون از جدول اعمال می‌کنیم و مقدار t را به‌دست می‌آوریم، ابتدا تصاویری که نرمال هستند مقادیر آنها باهم، جمع و متوسط‌گیری انجام می‌شود که مقدار به‌دست‌آمده X_1 است همچنین تصاویری که غیرنرمال هستند مقادیر آنها با هم، جمع و متوسط‌گیری انجام می‌شود که مقادیر به‌دست‌آمده X_2 می‌باشد. S_0^2 انحراف معیار مقادیر تصاویری که صفر یا نرمال هستند و S_1^2 انحراف معیار مقادیری که صفر یا یک هستند می‌باشد و n تعداد کلاس‌های نرمال و غیرنرمال را مشخص می‌کند. با توضیحاتی که بیان شد یک مقدار به‌دست می‌آید؛ این مقدار

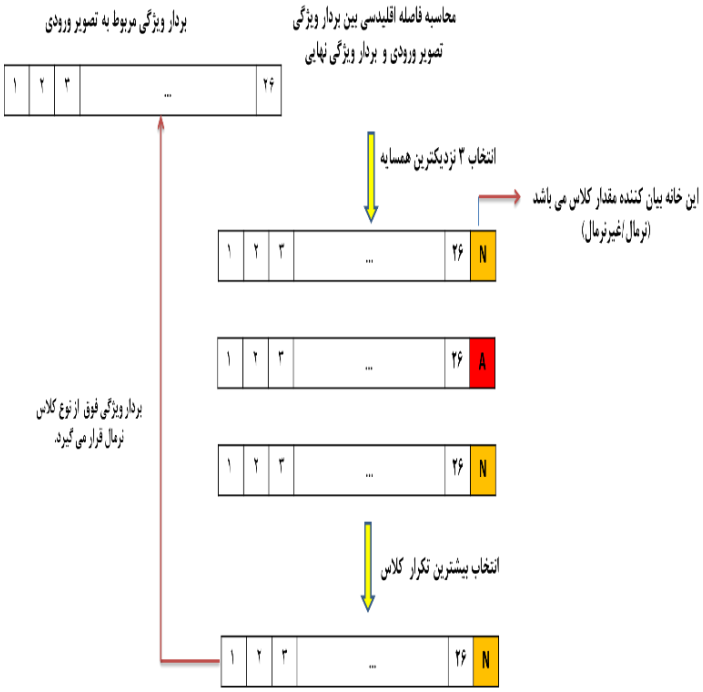
نشان‌دهنده ارزش هر ویژگی را بیان می‌کند همین روال برای سایر ستون‌های جدول محاسبه می‌شود و ارزش هر ویژگی مشخص می‌شود.

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{(S_0^2/n_0) + (S_1^2/n_1)}} \quad (۸)$$

جدول ۲. انتخاب ویژگی مناسب با درنظرگرفتن کلاس نرمال/غیرنرمال.

ویژگی تصویر	۱	۲	۳	(نرمال/غیر نرمال)
تصویر ۱	۵۰	-	-	۰
تصویر ۲	۲۰	-	-	۱
تصویر ۳	۱۰	-	-	۱
تصویر ۴	۶۰	-	-	۰
تصویر ۵	۳۰	-	-	۱

پس از انتخاب ویژگی‌های مؤثر مرحله مربوط به طبقه‌بندی صورت می‌گیرد. طبقه‌بند استفاده‌شده در روش پیشنهادی KNN می‌باشد که مقدار K=۹ درنظر گرفته شده است. برای برآوردن معیارهای دقت تشخیص فرض می‌شود که تصویر ورودی (که مرحله استخراج ویژگی و انتخاب ویژگی صورت‌گرفته) و بردار ویژگی نهایی مشخص است. بردار ویژگی تصویر ورودی طبق فاصله اقلیدسی، ۱۵ کوتاه‌ترین فاصله از بین بردار ویژگی نهایی پیدا می‌شود و با بررسی کلاس‌های (در اینجا فرض می‌کنیم که کلاس نرمال/غیرنرمال مدنظر است) مربوط به ۱۵ کوتاه‌ترین فاصله، کلاس‌هایی که بیشترین تکرار را داشته باشند مشخص می‌شود که تصویر ورودی درنظر گرفته‌شده مربوط به کدام کلاس (نرمال/غیرنرمال) است. با انجام مراحل فوق برای تمامی تصاویر ماموگرافی موجود در مجموعه تست، دقت کل قسمت تست مشخص می‌شود. برای هرکدام از قسمت‌های تست دقت به‌دست می‌آید و با میانگین‌گیری دقت‌های به‌دست‌آمده نهایی مشخص می‌شوند. نکته‌ای که باید درنظر گرفت این است که کلاس مربوط به بردار ویژگی نهایی از قبل مشخص است. شکل ۵ مثالی از طبقه‌بندی تصاویر ماموگرافی با مقدار K=۳ را نشان می‌دهد.



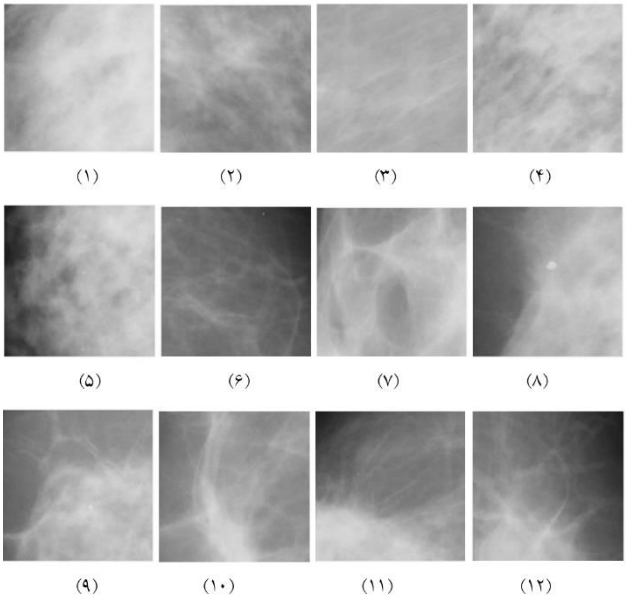
شکل ۳. مثالی از طبقه‌بندی تصاویر ماموگرافی با مقدار K=۳.

نتایج تجربی

در این بخش به مطالعه و ارزیابی نتایج حاصل از روش پیشنهادی پرداخته می‌شود. روش پیشنهادی به کمک نرم‌افزار متلب نسخه ۲۰۱۹ پیاده‌سازی گردید. در نتایج به‌دست‌آمده، میزان دقت تشخیص در مجموعه داده‌های آموزشی سرطان سینه ۹۲ درصد برای کلاس نرمال/غیرنرمال و ۸۱ درصد برای کلاس خوش‌خیم/بدخیم می‌باشد که نشان‌دهنده کارایی بالا برای روش پیشنهادی است. در مقایسه با سایر روش‌های شناسایی سرطان سینه، بهبود قابل‌ملاحظه‌ای در بهبود سه پارامتر دقت حساسیت و اختصاصی‌بودن دارد. برای پیاده‌سازی روش پیشنهادی از مجموعه تصاویر ماموگرافی از مجموعه داده MIAS استفاده شده است. مجموعه داده MIAS شامل ۳۲۲ تصویر ماموگرافی از پستان‌های چپ و راست ۱۶۱ خانم مختلف می‌باشد. ابعاد هریک از تصاویر پایگاه داده ۱۰۲۴ در ۱۰۲۴ بوده که با رزولوشن micron pixel edge ۲۰۰ دیجیتایز شده‌اند. در واقع تصاویر ماموگرافی دیجیتال، از نوع سطح خاکستری با عمق ۸ بیت هستند. این تصاویر از نظر آسیب، شامل پستان‌های نرمال، حاوی توده، حاوی خوشه میکروکلسیفیکیشن، نامتقارن و دچار درهم‌دریختگی ساختاری می‌باشند. این پایگاه شامل ۲۰۹ پستان نرمال، ۶۷ ناحیه موردعلاقه (RoI[1]) با آسیب‌های خوش‌خیم و ۵۴ RoI با آسیب‌های بدخیم است. برای ارزیابی روش ارائه‌شده، روش موردنظر را بارها اجرا و در هر مرحله میزان دقت در تشخیص نمونه‌ها را مشخص می‌کنیم. شکل ۶ برخی از نمونه‌های مناطق موردعلاقه از مجموعه داده MIAS که در پیاده‌سازی استفاده‌شده را نشان می‌دهد. این مناطق موردعلاقه به انواع مختلف و شدت آسیب‌شناسی‌های متفاوتی تعلق دارند. همان‌طور که در روش پیشنهادی بیان شد ابعاد این نمونه‌ها ۱۰۰*۱۰۰ پیکسل می‌باشد که از تصاویر ماموگرافی

[1] Region of Interest

در پایگاه داده استخراج شده است و با استخراج این قطعه‌ها، روش پیشنهادی با اعمال تحلیل LSD و طبقه‌بند KNN انجام شده است.



شکل ۴. نمونه‌ای از مناطق موردعلاقه استفاده‌شده از مجموعه داده MIAS.

میزان صحت یک آزمون به‌خصوص در حوزه شناسایی سرطان سینه را با استفاده از شاخص‌های اصلی دقت[1]، اختصاصی‌بودن[2]، بازخوانی[3]، درستی[4] و میانگین هارمونیک f1-Score بیان می‌کنند. به‌منظور ارزیابی روش پیشنهادی پس از پیاده‌سازی از علاوه بر شاخص‌های فوق، PPV و NPV نیز استفاده شده است.

- **دقت:** عبارت است از تعداد نمونه‌هایی که به‌درستی طبقه‌بندی شده‌اند، نسبت به کل نمونه‌ها.
- **اختصاصی‌بودن:** (که نرخ منفی واقعی نیز نامیده می‌شود) به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به‌درستی به عنوان منفی علامت‌گذاری می‌کند (برای مثال درصد افراد سالم که به‌درستی شناسایی شده‌اند و این افراد واقعاً بیمار نبوده‌اند).
- **درستی:** این معیار بیان می‌کند که چند درصد از خروجی‌های درست تشخیص داده‌شده، واقعاً درست هستند. به عبارت دیگر به حاصل تقسیم «تعداد نمونه‌هایی از کلاس X که به‌درستی پیش‌بینی شدند» بر «تعداد کل نمونه‌هایی که با نام کلاس X پیش‌بینی شده‌اند» گفته می‌شود.

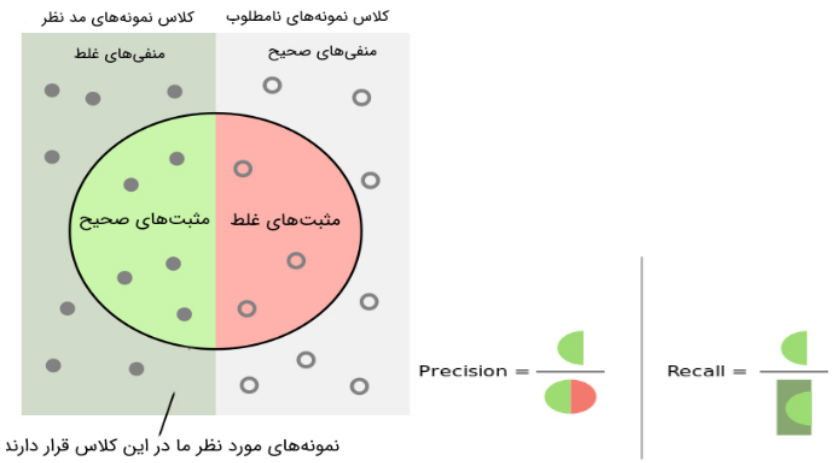
[1] Accuracy
[2] Specificity
[3] Sensitivity
[4] Precision

- **بازخوانی:** (که در برخی از علوم نرخ مثبت واقعی یا احتمال تشخیص صحیح نیز نامیده می‌شود) نسبتی از موارد مثبت است که آزمایش آن‌ها را به‌درستی به‌عنوان مثبت علامت‌گذاری می‌کند (برای مثال درصد افراد بیمار که به‌درستی شناخته شده‌اند و این افراد واقعاً سالم نیستند).
- **معیار f1-score**: زمانی که می‌خواهیم معیار ارزیابی ما میانگینی از دو مورد قبلی باشد یعنی همان درستی و بازخوانی؛ می‌توان از میانگین هارمونیک این دو معیار استفاده کرد که به آن معیار f1-score می‌گویند.

علاوه بر شاخص‌های ارزیابی فوق دو معیار به نام مقدار پیش‌بینی درست (PPV) و مقدار پیش‌بینی غلط (NPV) نیز برای ارزیابی نتایج درنظر گرفته شده است.

ماتریس درهم‌ریختگی، یکی از ابزارهای مفید برای ارزیابی عملکرد استفاده روش‌های دسته‌بندی است. اگر تعداد دسته‌های موجود m باشد، ماتریس درهم‌ریختگی جدولی به اندازه n*m خواهد بود. اگر i شماره سطر و j شماره ستون باشد، c_ij تعداد مشاهداتی از دسته i است که توسط الگوریتم دسته‌بندی j تشخیص داده شده است. این ماتریس چگونگی عملکرد الگوریتم دسته‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مسئله دسته‌بندی، نمایش می‌دهد. شکل ۷ نمایش بصری ماتریس درهم‌ریختگی به همراه معیارهای ارزیابی را بیان می‌کند.

- **TP (مثبت صحیح):** نشان‌دهنده تعداد پیش‌بینی‌های صحیح مربوط به همان کلاس
- **TN (منفی صحیح):** نشان‌دهنده تعداد درست پیش‌بینی‌شده به‌غیر از کلاس فعلی
- **FP (مثبت کاذب):** نشان‌دهنده تعداد پیش‌بینی نادرست مربوط به کلاس‌های دیگر
- **FN (منفی کاذب):** نشان‌دهنده تعداد پیش‌بینی نادرست مربوط به همان کلاس.



شکل ۷. نمایش بصری ماتریس درهم‌ریختگی.

نحوه محاسبه معیارهای ارزیابی برای کلاس نرمال/غیرنرمال:

```
Accuracy NA = (NA_TP+NA_TN)/(NA_TP+NA_TN+NA_FP+NA_FN);
Specificity NA = NA_TN/(NA_FP+NA_TN);
Recall NA = NA_TP/(NA_TP+NA_FN);
```

Precision NA = NA_TP/(NA_TP+NA_FP);
F1-Score NA = 2. ((Precision × Recall) / (Precision + Recall))
PPV_NA = NA_TP/(NA_TP+NA_FP);
NPV_NA = NA_TN/(NA_FN+NA_TN)

نحوه محاسبه معیارهای ارزیابی برای کلاس خوش‌خیم/بدخیم:

Accuracy MB = (MB_TP+MB_TN)/(MB_TP+MB_TN+MB_FP+MB_FN);
Specificity MB = MB_TN/(MB_FP+MB_TN);
Recall MB = MB_TP/(MB_TP+MB_FN);
Precision MB = MB_TP/(MB_TP+MB_FP);
F1-Score MB = 2. ((Precision × Recall) / (Precision + Recall))
PPV_MB = MB_TP/(MB_TP+MB_FP);
NPV_MB = MB_TN/(MB_FN+MB_TN)

جدول ۳ مقادیر دقت کلاس نرمال/ غیرنرمال(N/A) و خوش‌خیم/ بدخیم (M/B) برای روش پیشنهادی را نشان می‌دهد. مقادیر به‌دست‌آمده نشان‌دهنده دقت بالای روش پیشنهادی است. با توجه به نتایج مقدار دقت نرمال/ غیرنرمال نسبت به خوش‌خیم/بدخیم بیشتر می‌باشد.

جدول ۳. مقادیر دقت برای کلاس‌های نرمال/غیرنرمال و خوش‌خیم/ بدخیم توسط روش پیشنهادی.

کلاس	دقت Accuracy (٪)
نرمال / غیرنرمال	۹۲
خوش‌خیم / بدخیم	۸۱

جدول ۴ نتایج مربوط به معیارهای ارزیابی برای کلاس‌های نرمال/ غیرنرمال و خوش‌خیم/بدخیم را نشان می‌دهد. با توجه به نتایج به‌دست‌آمده می‌توان مشاهده کرد که روش پیشنهادی برای طبقه‌بندی کلاس نرمال/غیرنرمال بهتر از کلاس خوش‌خیم/بدخیم عمل کرده است. معیار دقت برای کلاس نرمال/غیرنرمال مقدار ۹۲ درصد است که این مقدار برای کلاس خوش‌خیم/بدخیم ۸۱ درصد را نشان می‌دهد. با توجه به مطالب بیان‌شده می‌توان این نتیجه را نوشت که روش پیشنهادی توانسته اهداف مربوط به افزایش دقت طبقه‌بند و تشخیص برای کلاس‌های نرمال/غیرنرمال و خوش‌خیم/بدخیم را به‌خوبی برآورده کند.

جدول ۴. نتایج حاصل از معیارهای ارزیابی برای کلاس‌های N/A و B/M توسط روش پیشنهادی.

معیار ارزیابی	کلاس	نرمال / غیرنرمال (N/A)	خوش‌خیم / بدخیم (B/M)
Accuracy		۹۲ درصد	۸۱ درصد
Specificity		۹۹ درصد	۶۷ درصد
Recall		۹۳ درصد	۸۲ درصد

معیار ارزیابی	کلاس	نرمال / غیرنرمال (N/A)	خوش‌خیم / بدخیم (B/M)
Precision		۷۷ درصد	۸۴ درصد
F1-Score		۸۴ درصد	۶۴ درصد
PPV		۹۸ درصد	۹۱ درصد
NPV		۸۸ درصد	۵۲ درصد

همچنین جدول ۵ نتایج مربوط به آزمون تشخیص سرطان سینه با استفاده از انواع طبقه‌بند را نشان می‌دهد. با توجه به نتایج نشان داده‌شده برای طبقه‌بندی روش پیشنهادی، انواع طبقه‌ها از جمله شبکه‌های عصبی کانولوشنی (CNN)، ماشین بردار پشتیبان (SVM)، طبقه‌بند نیو بیزین (NB) و طبقه‌بند نزدیک‌ترین همسایه (KNN) پیاده‌سازی و آزمایش شد. از میان این طبقه‌بندها، طبقه‌بند نزدیک‌ترین همسایه توانست بر روی روش پیشنهادی با استفاده از تعداد کمتری از ویژگی‌ها دقت شناسایی سیستم تشخیص سرطان سینه را بهبود بخشد. این نکته را باید درنظر گرفت که علاوه بر عملکرد بهتر طبقه‌بند، مواردی همچون نحوه استخراج ویژگی و انتخاب ویژگی‌های مؤثر نیز در راستای افزایش بهینه‌گی معیارهای ارزیابی نقش دارند. بنابراین روش پیشنهادی با درنظرگرفتن ویژگی‌های کارآمد و حذف ویژگی‌های غیرمؤثر و با انتخاب طبقه‌بند مناسب توانسته به مقادیر بالایی برای معیارهای ارزیابی دست یابد.

جدول ۵. مقایسه طبقه‌بند استفاده شده در روش پیشنهادی با طبقه بندهای مختلف.

طبقه‌بند	Accuracy (درصد)	Specificity (درصد)	Precision (درصد)
CNN	۷۶	۷۶	۷۰
SVM	۷۴	۷۵	۷۳
NB	۸۱	۹۱	۷۴
KNN پیشنهادی	۹۲	۹۹	۷۷

جدول ۶ و ۷ مقایسه عملکرد روش پیشنهادی برای شناسایی کلاس‌های نرمال/ غیرنرمال و خوش‌خیم/بدخیم با سایر روش‌های موردبررسی در بخش دوم را نشان می‌دهد. موارد بیان‌شده و نتایج حاصل، نشان‌دهنده دقت بالای روش پیشنهادی برای تشخیص برای هر دو کلاس نرمال/غیرنرمال و خوش‌خیم/بدخیم نسبت به سایر روش‌ها است. در روش پیشنهادی با اندازه مجموعه ویژگی ۲۶ و تعداد تصاویر ۳۳۰، دقت به‌دست‌آمده ۹۲ درصد شده است.

جدول ۶. مقایسه دقت روش پیشنهادی برای تشخیص کلاس‌های نرمال/غیرنرمال (N/A) با سایر روش‌ها.

روش	طبقه‌بندی	اندازه ناحیه موردعلاقه	تعداد تصاویر	دقت Accuracy (٪)
روش CNNI-BCC (تینگ و همکاران، ۲۰۱۹)	NB	۲۰۰*۲۰۰	۳۰۰	۹۰
روش PCA-CC (التوخی و همکاران، ۲۰۱۴)	SVM	۲۰۰*۲۰۰	۱۰۲۴	۹۰
روش COCC (گدیک و آتسوی، ۲۰۱۳)	لجستیک ساده	۱۲۸*۱۲۸	۱۰۵۶	۸۳
روش SCC (گاردزی و همکاران، ۲۰۱۹)	SVM	۱۲۸*۱۲۸	۴۰۰	۸۵
روش پیشنهادی	KNN	۱۰۰*۱۰۰	۳۳۰	۹۲

جدول ۱. مقایسه دقت روش پیشنهادی برای تشخیص کلاس‌های خوش‌خیم/ بدخیم (M/B) با سایر روش‌ها

روش	طبقه‌بندی	اندازه ناحیه موردعلاقه	تعداد تصاویر	دقت Accuracy (٪)
روش CNNI-BCC (تینگ و همکاران، ۲۰۱۹)	NB	۲۰۰*۲۰۰	۳۰۰	۸۹
روش PCA-CC (التوخی و همکاران، ۲۰۱۴)	SVM	۲۰۰*۲۰۰	۱۰۲۴	۷۲
روش COCC (گدیک و آتسوی، ۲۰۱۳)	لجستیک ساده	۱۲۸*۱۲۸	۱۰۵۶	۷۵
روش SCC (گاردزی و همکاران، ۲۰۱۹)	SVM	۱۲۸*۱۲۸	۴۰۰	۶۹
روش پیشنهادی	KNN	۱۰۰*۱۰۰	۳۳۰	۸۱

نتیجه‌گیری و پیشنهادها

سرطان پستان، یکی از شایع‌ترین انواع سرطان است که هرساله باعث مرگ‌ومیر فراوانی در بین زنان و مردان می‌شود و علی‌رغم پیشرفت‌های بسیاری که در مورد تشخیص زودهنگام و درمان مناسب این بیماری صورت گرفته است، کماکان یک از علل اصلی مرگ با سرطان در بین زنان است. با اینکه هر روز راهکارهای جدیدتری در برخورد با سرطان پستان معرفی می‌شود، هنوز هم این بیماری، جان عده زیادی را در معرض خطر قرار داده است، شاید توجه بیشتر به ساختار ریزملکولی و اساس زیست‌شناسی این بیماری باعث شود تا بتوان با دانسته‌های بیشتری در مورد چگونگی پیدایش این بیماری از سطح سلول‌ها، اطلاعات گسترده‌تری در مورد ایجاد رشد این بیماری مهلک به‌دست آورد. بنابراین سرطان سینه، دومین مورد از سرطان شایع در میان زنان و دومین سرطان منجر به مرگ در جهان می‌باشد. سالیانه تعدادی زیادی زن در اثر ابتلا به این بیماری جان خود را ازدست می‌دهند. براساس آمار مرکز ملی سرطان، فقط در کشور آمریکا از هر هشت زن یک نفر به بیماری سرطان سینه مبتلا می‌شود. بنابراین شش درصد از کل مرگ‌ومیرهای جهان ناشی از ابتلای به این بیماری است. در کشور ایران نیز سرطان سینه سومین عامل مرگ‌ومیر در بین زنان است. تحقیقات زیادی در ارتباط با سرطان سینه نشان می‌دهند که پیشگیری از این بیماری به دلیل ناشناخته‌بودن عوامل آن تقریباً غیرممکن به‌نظر می‌رسد. بنابراین تشخیص به‌موقع یکی از عوامل مهم و اساسی در درمان این بیماری است. اتخاذ تکنیک‌های تشخیصی مؤثر و کارآمد در مراحل اولیه، بسیار حائز اهمیت می‌باشد و باید به‌عنوان یکی از اصلی استراتژی‌هایی که هدف آنها ارتقای سلامت زنان و کاهش میزان ابتلا و مرگ‌ومیر ناشی از سرطان پستان است، لحاظ گردد. یکی از این تکنیک‌های ماموگرافی است. ماموگرافی اشعه ایکس متداول‌ترین تکنیک مورداستفاده رادیولوژیست‌ها در تشخیص و غربالگری سرطان پستان است. استفاده از روش ماموگرافی در حال حاضر رایج‌ترین راه تشخیص زودهنگام این بیماری است و درصد مرگ‌ومیر را تا ۲۵ درصد کاهش داده است ولی با این حال تفسیر و تشریح تصاویر حاصل از ماموگرافی بسیار دشوار می‌باشد و براساس آمار رسمی مرکز ملی سرطان در آمریکا ۱۰ تا ۳۰ درصد غدد موجود در پستان بیمار در تصاویر ماموگرافی توسط رادیولوژیست قابل تشخیص نیستند. در دهه اخیر، تحقیقات گسترده‌ای برای کاهش خطای تشخیص سرطان پستان و همچنین افزایش سرعت تشخیص انجام گرفته است. نتایج این تحقیقات می‌تواند به رادیولوژیست‌ها و متخصصان در تشخیص سریع و مطمئن کمک کنند. استفاده از روش‌ها و تکنیک‌های پردازش تصویر و شناسایی الگوها در تشخیص و تعیین خودکار سرطان پستان از روی تصاویر ماموگرافی باعث کم‌شدن خطاهای انسانی و افزایش سرعت تشخیص می‌شود. با درنظرگرفتن این مورد در این تحقیق با توجه به ضرورت تشخیص زودهنگام و به‌موقع این بیماری، روشی نوین مبتنی بر تحلیل LSD و طبقه‌بند KNN بر روی ماموگرافی ارائه شد. روش پیشنهادی با استفاده از نرم‌افزار متلب اجرا شد و نتایج نشان‌دهنده این است که روش ارائه‌شده در کم‌کردن خطاهای انسانی در تشخیص توده‌های نرمال و غیرنرمال در تصاویر با دقت ۹۲ درصد بسیار مؤثر است. تصاویر متعدد دریافت‌شده از مجموعه داده ماموگرافی MIAS توسط مدل ارائه‌شده، بررسی و تحلیل شدند که نتیجه حاصل از آنها بسیار قابل‌قبول می‌باشد و

از لحاظ معیارهای مختلف نسبت به مدل‌های ارائه‌شده در مقالات معتبر بالاتر می‌باشد. با توجه به دسترسی به داده‌های موردنظر و به‌روزبودن موضوع و ارائه راهکارهای مختلف در سایت‌ها و تحقیقات گسترده در کشورهای مختلف برای دستیابی به بهترین روند تشخیص، می‌توان تحقیقات صورت‌گرفته در کشورهای مختلف را جمع‌آوری و میزان موفقیت آن‌ها را ارزیابی کرد. بنابراین با وجود پیشرفت چشمگیر که در سال‌های اخیر صورت گرفته هنوز نیاز است که کار بیشتری برای گسترش سیستم‌های شناسایی سرطان سینه و استفاده از روش‌های دقیق انجام شود. استفاده از روش مناسب کارا و مؤثر باید به کشف زودهنگام بیماری و پیش‌بینی پیشرفته برای بیماری منجر شود بنابراین در کارهای آینده برای افزایش دقت تشخیص سرطان سینه می‌توان مواردی همچون کاهش تعداد ویژگی در راستای سرعت کارکرد سیستم پیشنهادی، استفاده از الگوریتم شبکه عصبی مصنوعی برای انتخاب ویژگی‌های مناسب، ارتقای طبقه‌بندی با تصفیه ویژگی‌ها و طبقه‌بندی با استفاده از انواع مختلف الگوریتم‌های آموزشی را لحاظ کرد.

References

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *36*(2), 3240-3247. https://doi.org/10.1016/j.eswa.2008.01.009

Alyami, J., Sadad, T., Rehman, A., Almutairi, F., Saba ,T., Bahaj, S. A., & Alkhurim, A. (2022). Cloud Computing-Based Framework for Breast Tumor Image Classification Using Fusion of AlexNet and GLCM Texture Features with Ensemble Multi-Kernel Support Vector Machine (MK-SVM). *Computational Intelligence and Neuroscience*, *2022*(11), 1-9. https://doi.org/10.1155/2022/7403302

Avcı, H., & Karakaya, J. (2023). A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning. *Diagnostics*, *13*(3), 348. https://doi.org/10.3390/diagnostics13030348

Ayyoubzadeh, S. M., Baniasadi, T., Shirkhoda, M., Rostam Niakan Kalhori, S., Mohammadzadeh, N., Roudini, K., Ghalehtaki, R., Memari, F., & Jalaeefar, A. (2023). Remote Monitoring of Colorectal Cancer Survivors Using a Smartphone App and Internet of Things–Based Device: Development and Usability Study. *Journal of Medical Internet Research cancer*, *9*, e42250. https://doi.org/10.2196/42250

Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., & Palma, L. (2023). A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors*, *23*(7), 3565. https://doi.org/10.3390/s23073565

Chan, R. C., To, C. K. C., Cheng, K. C. T., Yoshikazu, T., Yan, L. L. A., & Tse, G. M. (2023). Artificial intelligence in breast cancer histopathology. *Histopathology*, *82*(1), 198-210. https://doi.org/10.1111/his.14820

Chen, H-L., Yang, B., Liu, J., & Liu, D.-Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014-9022. https://doi.org/10.1016/j.eswa.2011.01.120

Eltoukhy, M. M., Gardezi, S. J. S., & Faye, I. (2014, April 14-16). *A method to reduce curvelet coefficients for mammogram classification.* 2014 Institute of Electrical and Electronics Engineers Region 10 Symposium, Kuala Lumpur, Malaysia. https://doi.org/10.1109/TENCONSpring.2014.6863116

Escobar-Linero, E., Muñoz-Saavedra, L., Luna-Perejón, F., Civit-Masot, J., Rivas-Pérez, M., Domínguez-Morales, M., & Balcells, A. C. (2023). Evolution and Tendency on the Feature Extraction Process for Diagnostic Aid in Healthcare. In F. Zeshan & A. Ahmad (Eds.), *Recent Advancements in Smart Remote Patient Monitorling, Wearabe Devices*,

and Diagnostics Systems (pp. 109-153). IGI Global. https://doi.org/10.4018/978-1-6684-6434-2.ch006

Gardezi, S. J. S., Elazab, A., Lei, B., & Wang, T. (2019). Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review. *Journal of Medical Internet Research 21*(7), e14464. https://doi.org/10.2196/14464

Gardezi, S. J. S., Faye, I., & Eltoukhy, M. M. (2014, October 26-27). *Analysis of mammogram images based on texture features of curvelet Sub-bands.* Fifth International Conference on Graphic and Image Processing, Hong Kong, China. https://doi.org/10.1117/12.2054183

Gedik, N., & Atasoy, A. (2013). A computer-aided diagnosis system for breast cancer detection by using a curvelet transform. *Turkish Journal of Electrical Engineering and Computer Sciences*, *21*(4), 1002-1014. https://doi.org/10.3906/elk-1201-8

Jasti, V. D. P., Zamani, A. S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., Raghuvanshi, A., & Kaliyaperumal, K. (2022). Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. *Security and Communication Networks*, *2022*(1918379), 1-7. https://doi.org/10.1155/2022/1918379

Kavitha, T., Mathai, P. P., Karthikeyan, C., Ashok, M., Kohar, R., Avanija, J., & Neelakandan, S. (2022). Deep Learning Based Capsule Neural Network Model for Breast Cancer Diagnosis Using Mammogram Images. *Interdisciplinary Sciences: Computational Life Sciences*, *14*(1), 113-129. https://doi.org/10.1007/s12539-021-00467-y

Li, S., Hara, T., Hatanaka, Y., Fujita, H., Endo, T., & Iwase, T. (2001). Performance Evaluation of a CAD System for Detecting Masses on Mammograms by Using the MIAS Database. *Medical Imaging and Information Sciences*, *18*(3), 144-153. https://doi.org/10.11318/mii1984.18.144

Mader, K. S. (2017). *MIAS Mammography* [Dataset]. Kaggle. https://www.kaggle.com/datasets/kmader/mias-mammography/data

Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. *Expert systems with applications*, *120*(6), 103-115. https://doi.org/10.1016/j.eswa.2018.11.008

Tsochatzidis, L., Costaridou, L., & Pratikakis, I. (2019). Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study. *Journal of Imaging*, *5*(3), 37. https://doi.org/10.3390/jimaging5030037

Yari, Y., Nguyen, T. V., & Nguyen, H. T. (2020). Deep Learning Applied for Histological Diagnosis of Breast Cancer. *Institute of Electrical and Electronics Engineers Access*, *8*, 162432-162448. https://doi.org/10.1109/ACCESS.2020.3021557

Zheng, D., He, X., & Jing, J. (2023). Overview of Artificial Intelligence in Breast Cancer Medical Imaging. *Journal of Clinical Medicine*, *12*(2), 419. https://doi.org/10.3390/jcm12020419

Zonderland, H. M., Coerkamp, E. G., Hermans, J., Vijver, M. J. V. D., & Voorthuisen, A. E. V. (1999). Diagnosis of Breast Cancer: Contribution of US as an Adjunct to Mammography. *Radiology*, *213*(2), 413-422. https://doi.org/10.1148/radiology.213.2.r99nv05413