Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 210-220, E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026931

Modeling and classification of breast cancer using machine learning techniques: A comparative analysis of Bayesian Networks and XGBoost

Suzan S. Haydar [1*], Farhad A. Ahmed[2]
[1,2] Department of Statistics and Informatics, College of Administration and Economics, Sulamani University, Sulamani, Iraq
sozan.haider@univsul.edu.iq, farhad.ahmed@univsul.edu.iq

Article information:
Published: 25–08– 2026

[*]Corresponding author:
Suzan Saber Haydar
sozan.haider@univsul.edu.iq

This work is licensed under a Creative Commons Attribution 4.0 International License.

Abstract:

Breast cancer is one of the most prevalent and life-threatening diseases worldwide, making early and accurate diagnosis essential for improving patient outcomes. This study aims to evaluate the performance of two machine learning models, Extreme Gradient Boosting (XGBoost) and Gaussian Naive Bayes, for breast cancer classification using the Breast Cancer Wisconsin Diagnostic dataset. A comparative analysis was conducted based on several performance metrics, including Accuracy, Precision, Recall, F1-score, Specificity, Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), Balanced Accuracy, and Confusion Matrix. In addition, feature importance analysis was performed to identify the most influential variables contributing to the classification process. The experimental results demonstrated that both models achieved high classification performance. However, the Gaussian Naive Bayes model showed slightly superior performance, achieving an Accuracy of 96.5% and an AUC of 0.997, compared with 96.0% and 0.993, respectively, for the XGBoost model. Furthermore, the analysis revealed that features related to tumor morphology, particularly concave points mean, perimeter worst, and radius worst, were among the most influential predictors of breast cancer classification. The findings indicate that both machine learning models provide reliable and effective tools for supporting early breast cancer diagnosis, with Gaussian Naive Bayes demonstrating a slight advantage for the dataset used in this study. These results highlight the potential of machine learning techniques to enhance diagnostic accuracy and support clinical decision-making.

Keywords: Breast cancer classification, machine learning, XGBoost, gaussian naive bayes, medical diagnosis, ROC-AUC, feature importance.

Conclusions:

This study compared the performance of XGBoost and Gaussian Naive Bayes for breast cancer classification using the Breast Cancer Wisconsin Diagnostic dataset. The results showed that both models achieved excellent classification performance, demonstrating the effectiveness of machine learning in supporting early breast cancer diagnosis. However, Gaussian Naive Bayes slightly outperformed XGBoost in terms of Accuracy and ROC-AUC. Overall, the findings indicate that both models are reliable tools for medical decision support, with Gaussian Naive Bayes showing a slight advantage for the dataset used in this study.

مجلة الريادة للمال والاعمال
2026، المجلد السابع، العدد الخاص، الصفحات 210- 220

نمذجة وتصنيف سرطان الثدي باستخدام تقنيات التعلم الآلي: دراسة مقارنة بين الشبكات البايزية وخوارزمية XGBoost

سوزان صابر حيدر*[1]، فرهاد على احمد[2]
[1،2] قسم الاحصاء والمعلوماتية، كلية الادارة و الاقتصاد، جامعة السليمانية، السليمانية، العراق
sozan.haider@univsul.edu.iq, farhad.ahmed@univsul.edu.iq

معلومات البحث:
- تاريخ النشر: 2026 –08– 25

المؤلف المراسل:
سوزان صابر حيدر
sozan.haider@univsul.edu.iq

هذا العمل مرخص بموجب المشاع الابداعي نسب المصنف 4.0 دولي (CC BY 4.0)

المستخلص:

يُعد سرطان الثدي من أكثر الأمراض انتشارًا وخطورة على مستوى العالم، حيث يلعب التشخيص المبكر والدقيق دورًا مهمًا في تحسين فرص بقاء المرضى على قيد الحياة. تهدف هذه الدراسة إلى استخدام تقنيات تعلم الآلة لتصنيف حالات سرطان الثدي اعتمادًا على مجموعة بيانات Breast Cancer Wisconsin Diagnostic. تم تطبيق وتقييم نموذجي (XGBoost) و (Naive Bayes) للتمييز بين الحالات الحميدة والخبيثة. وتم تقييم أداء النماذج باستخدام مجموعة من المقاييس الإحصائية، بما في ذلك الدقة (Accuracy)، والدقة الإيجابية(Precision) ، والاسترجاع (Recall)، ودرجة F1، بالإضافة إلى منحنى ROC والمساحة تحت المنحنى (AUC)، إلى جانب تحليل مصفوفة الالتباس لتقييم الأخطاء بشكل تفصيلي. حيث أظهرت النتائج التجريبية أن كلا النموذجين حققا أداءً عاليًا في التصنيف، مع تفوق طفيف لنموذج Naive Bayes من حيث قيمة AUC التي بلغت (0.997) مقارنة بـ (0.993) لنموذج XGBoost، مما يدل على قدرة تمييزية قوية لكلا النموذجين. كما بين تحليل أهمية الخصائص أن المتغيرات المرتبطة بالشكل والحجم مثل concave points mean وperimeter worst وradius worst تُعد من أهم العوامل المؤثرة في عملية التنبؤ.

بشكل عام، تؤكد النتائج أن نماذج تعلم الآلة يمكن أن تُستخدم كأدوات فعالة لدعم القرار الطبي في التشخيص المبكر لسرطان الثدي، مما يساهم في تحسين دقة التشخيص وتقليل الأخطاء الطبية.

الكلمات المفتاحية: تعلم الآلة، تصنيف سرطان الثدي، خوارزمية XGBoost، نموذج بايزNaive Bayes ، منحنى (AUC)، مصفوفة الالتباس.

المقدمة:

يُعدّ سرطان الثدي من أكثر الأمراض شيوعًا وخطورةً بين النساء في مختلف أنحاء العالم، حيث يلعب التشخيص المبكر والدقيق دورًا حاسمًا في زيادة فرص البقاء على قيد الحياة وتحسين فعالية العلاج. ومع ذلك، فإن الطرق التقليدية للتشخيص الطبي تعتمد بشكل كبير على التفسير البشري، مما يجعلها عرضة للأخطاء الفردية والاختلافات في التقييم السريري. في السنوات الأخيرة، أصبحت تقنيات تعلم الآلة من الأدوات القوية في مجال التشخيص الطبي وتصنيف الأمراض، حيث تتميز بقدرتها على استخراج الأنماط المعقدة من البيانات الطبية عالية الأبعاد، وغالبًا ما تقدم أداءً أفضل مقارنة بالأساليب الإحصائية التقليدية. ومن بين هذه التقنيات، برزت النماذج المعتمدة على التعلم التجميعي مثل خوارزمية XGBoost، والنماذج الاحتمالية مثل الشبكات البايزية، وذلك لما تتمتع به من دقة عالية وقدرة على التفسير. تُعد خوارزمية XGBoost من تقنيات التعلم المتقدمة التي تعتمد على تعزيز التدرج، حيث تتميز بقدرتها على نمذجة العلاقات غير الخطية والتفاعلات بين المتغيرات، مما يجعلها فعالة في مهام التصنيف. في المقابل، تعتمد الشبكات البايزية على تمثيل العلاقات الشرطية بين المتغيرات في شكل احتمالي، مما يتيح التعامل مع عدم اليقين في البيانات الطبية ودعم اتخاذ القرار. تهدف هذه الدراسة إلى مقارنة أداء كل من XGBoost والشبكات البايزية في تصنيف سرطان الثدي باستخدام مجموعة بيانات Breast Cancer Wisconsin Diagnostic، وذلك بالاعتماد على عدة مقاييس تقييم مثل الدقة، الاسترجاع، الدقة الإيجابية، مقياسF1 ، ومساحة تحت منحنىROC-AUC ، بالإضافة إلى استخدام التحقق المتقاطع (Cross Validation). كما تتضمن الدراسة تحليل أهمية الخصائص لتحديد المتغيرات الأكثر تأثيرًا في عملية التنبؤ بالمرض.

شهدت السنوات الأخيرة اهتمامًا متزايدًا بتطبيق تعلم الآلة في مجال تشخيص سرطان الثدي، حيث أظهرت العديد من الدراسات فعالية عالية لهذه النماذج في تحليل البيانات الطبية. فقد بيّنت دراسة Wang et al. (2024) أن النماذج المعتمدة على تعزيز التدرج (Gradient Boosting)تحقق أداءً أفضل من النماذج التقليدية مثل الانحدار اللوجستي في تصنيف سرطان الثدي. كما استخدم Li et al. (2024)الشبكات البايزية لنمذجة العلاقات الاحتمالية بين المتغيرات السريرية، وأكدوا أن هذه النماذج أكثر موثوقية في التعامل مع

حالات عدم اليقين في البيانات. وفي دراسة أخرى، أظهر Khan et al. (2025) أن خوارزمية XGBoost حققت دقة عالية تتجاوز 96% في التنبؤ بسرطان الثدي باستخدام بيانات جينية. بالإضافة إلى ذلك، أجرى Alvarez et al. (2024) دراسة مقارنة بين عدة خوارزميات مثل Random Forest وSVM وXGBoost، ووجدوا أن XGBoost يقدم أفضل أداء من حيث الدقة والاستقرار. كما اقترح Chen et al. (2025) دمج النماذج البايزية مع تقنيات التعلم العميق لتعزيز قابلية التفسير في أنظمة التشخيص الطبي، وأظهرت النتائج أن النماذج الهجينة يمكن أن تحقق أداءً أفضل من النماذج الفردية. وتشير هذه الدراسات إلى أنه رغم القوة العالية لنماذج تعلم الآلة الحديثة، لا يزال هناك حاجة إلى دراسات مقارنة دقيقة بين النماذج الاحتمالية والنماذج المعتمدة على التعزيز.

1. مشكلة البحث:

رغم التطور الكبير في تطبيقات تعلم الآلة في تشخيص سرطان الثدي، لا يزال هناك تباين في نتائج الدراسات المتعلقة بأداء خوارزميات التصنيف المختلفة، ولا سيما بين النماذج الاحتمالية مثل Naive Bayes والنماذج المعتمدة على التعلم التجميعي مثل XGBoost. كما أن العديد من الدراسات ركزت على تحقيق أعلى دقة تصنيف دون إجراء مقارنة شاملة باستخدام مؤشرات تقييم متعددة توضح نقاط القوة والضعف لكل نموذج في التطبيقات الطبية الحساسة. لذلك، تتمثل مشكلة هذا البحث في عدم وضوح أيٍ من النموذجين يوفر أداءً أكثر كفاءة وموثوقية في تصنيف حالات سرطان الثدي عند تطبيقهما على مجموعة بيانات Breast Cancer Wisconsin Diagnostic، مع الأخذ في الاعتبار مجموعة متكاملة من مؤشرات الأداء الإحصائية وقابلية تفسير النتائج.

2. أهمية البحث:

تكمن أهمية هذا البحث في تقييم ومقارنة أداء خوارزمية XGBoost ونموذج Gaussian Naive Bayes في تصنيف سرطان الثدي، بما يسهم في تحديد النموذج الأكثر كفاءة وموثوقية، ودعم تطوير أنظمة تشخيص ذكية تساعد في تحسين دقة التشخيص المبكر وتقليل الأخطاء الطبية.

3. أهداف البحث:

يهدف هذا البحث إلى:

أ. بناء نموذجين لتعلم الآلة باستخدام خوارزمية XGBoostونموذج Gaussian Naive Bayesلتصنيف حالات سرطان الثدي .

ب. مقارنة أداء النموذجين باستخدام مجموعة من مؤشرات التقييم الإحصائية، مثل Accuracy وPrecision و Recall وF1-score و ROC-AUC.

ج. تحديد النموذج الأكثر كفاءة وموثوقية في تصنيف حالات سرطان الثدي اعتمادًا على نتائج المقارنة .

د. تحليل أهمية الخصائص (Feature Importance) لتحديد المتغيرات الأكثر تأثيرًا في عملية تصنيف سرطان الثدي وإبراز إمكانية توظيف تقنيات تعلم الآلة في دعم التشخيص المبكر وتحسين دقة اتخاذ القرار الطبي.

4. الفرضيات:

ينطلق هذا البحث من فرضية مفادها أن هناك اختلافًا في أداء خوارزمية XGBoost ونموذج Gaussian Naive Bayes في تصنيف حالات سرطان الثدي، وأن هذا الاختلاف ينعكس في مؤشرات تقييم الأداء المختلفة، مثل الدقة (Accuracy) ، والدقة الإيجابية (Precision)، والاسترجاع (Recall)، ودرجة F1، والمساحة تحت منحنى ROC (AUC). كما يفترض البحث أن تحليل أهمية الخصائص (Feature Importance) سيمكن من تحديد المتغيرات الأكثر تأثيرًا في عملية تصنيف سرطان الثدي، بما يسهم في اختيار النموذج الأكثر كفاءة وموثوقية لدعم التشخيص الطبي المبكر

5. مجتمع وعينة البحث:

تمثل مجتمع الدراسة بمجموعة بيانات Breast Cancer Wisconsin Diagnostic، وتم اختيارها عينةً للدراسة، إذ تضم (569) عينة، تحتوي كل منها على (30) خاصية عددية، بالإضافة إلى متغير الهدف الذي يصنف الحالات إلى حميدة (Benign) وخبيثة (Malignant)، وقد تم اعتمادها ضمن حدود هذه الدراسة.

6. المواد والطرق (Materials and Methods)

تعتمد هذه الدراسة على إطار عمل قائم على تعلم الآلة من أجل نمذجة وتصنيف سرطان الثدي باستخدام مجموعة بيانات Breast Cancer Wisconsin Diagnostic. ويهدف هذا القسم إلى توضيح المنهجية المتبعة في بناء النماذج، وآلية التدريب، ومعايير التقييم المستخدمة في مقارنة الأداء بين النماذج المختلفة [7].

المحور الأول: الجانب النظري

أولاً: نمذجة وتصنيف باستخدام تقنيات تعلم الآلة

تمت صياغة المشكلة في هذه الدراسة كمسألة تصنيف(Classification Problem) ، حيث يتم التنبؤ بكون الورم سرطانيًا خبيثًا أو حميدًا اعتمادًا على مجموعة من الخصائص العددية المستخرجة من بيانات التشخيص. وقد تم اعتماد نموذجين مختلفين في مجال تعلم الآلة لتحقيق مقارنة شاملة بين أسلوبين مختلفين في النمذجة [8].

النموذج الأول هو خوارزمية XGBoost، وهي خوارزمية تعتمد على التعلم التجميعي باستخدام تعزيز التدرج (Gradient Boosting)، حيث يتم بناء مجموعة من أشجار القرار بشكل تدريجي بهدف تقليل الخطأ وتحسين دقة التنبؤ. وتتميز هذه الخوارزمية بقدرتها العالية على التعامل مع العلاقات غير الخطية والتفاعلات المعقدة بين المتغيرات، مما يجعلها مناسبة جدًا للتطبيقات الطبية.

أما النموذج الثاني فهو المصنف البايزي (Bayesian Classifier) وتحديدًا نموذج Gaussian Naive Bayes، والذي يعتمد على نظرية بايز مع افتراض الاستقلال الشرطي بين المتغيرات. يقوم هذا النموذج بحساب الاحتمالات الشرطية لكل فئة ومن ثم اختيار الفئة ذات الاحتمال الأعلى، ويتميز ببساطته وسرعته وقدرته على العمل بكفاءة مع البيانات عالية الأبعاد [9].

خوارزمية XGBoost

تُعد خوارزمية XGBoost (Extreme Gradient Boosting) واحدة من أكثر خوارزميات التعلم التجميعي تقدمًا وفعالية في مجال تعلم الآلة، حيث تعتمد على مبدأ التعلم التجميعي القائم على بناء نماذج متتابعة من أشجار القرار، بحيث يتم إنشاء سلسلة من النماذج بشكل تدريجي، ويقوم كل نموذج جديد بمحاولة تصحيح الأخطاء التي ارتكبها النموذج السابق.

تعتمد هذه الخوارزمية على تحسين دالة الهدف(Objective Function) التي تتكون من جزأين رئيسيين، هما دالة الخسارة (Loss Function)ومصطلح التنظيم(Regularization Term) ، وذلك بهدف تقليل خطأ التنبؤ وتحسين قدرة النموذج على التعميم وتقليل مشكلة فرط التعلّم [10].

يمكن تمثيل دالة الهدف في خوارزمية XGBoost بالشكل التالي:

$$l(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=0}^{K} \Omega(f_k)$$

حيث $l(y_i, \hat{y}_i)$ تمثل دالة الخسارة التي تقيس الفرق بين القيم الحقيقية والقيم المتوقعة، بينما تمثل $\Omega(f_k)$ مصطلح التنظيم الذي يتحكم في تعقيد النموذج، ويمثل K عدد أشجار القرار المستخدمة في النموذج.

تعتمد الخوارزمية أيضًا على مبدأ الانحدار التدرجي المعزز(Gradient Boosting) ، حيث يتم تحسين النموذج بشكل تدريجي من خلال تقليل دالة الخطأ باستخدام المشتقات الأولى والثانية، مما يساعد على تحسين دقة التنبؤ في كل خطوة تدريبية.

تتميز خوارزمية XGBoost بعدة خصائص مهمة تجعلها مناسبة للتطبيقات الطبية، من بينها قدرتها العالية على التعامل مع العلاقات غير الخطية بين المتغيرات، وكفاءتها في معالجة البيانات عالية الأبعاد، إضافة إلى قدرتها على تقليل مشكلة فرط التعلّم من خلال استخدام تقنيات التنظيم الداخلي. وبناءً على هذه المزايا، تُستخدم خوارزمية XGBoost بشكل واسع في مجالات التصنيف الطبي، وخاصة في تشخيص سرطان الثدي، حيث أثبتت فعاليتها في تحقيق أداء دقيق وموثوق مقارنة بالعديد من النماذج التقليدية [12] [11].

النموذج البايزي(Bayesian Model)

يعتمد نموذج Gaussian Naive Bayes على مبدأ الاحتمالات الإحصائية وفقًا لنظرية بايز، حيث يتم تقدير احتمال انتماء كل عينة إلى فئة معينة بناءً على القيم المرصودة للخصائص. ويُفترض في هذا النموذج أن جميع المتغيرات مستقلة شرطيًا، أي أن تأثير كل خاصية يُحسب بشكل مستقل عن الخصائص الأخرى عند إعطاء الفئة.

يتم حساب الاحتمال الخلفي (Posterior Probability) لكل فئة باستخدام قاعدة بايز كما يلي:

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)}$$

حيث تمثل:

$P(C/X)$ احتمال انتماء العينة X إلى الفئة C و $P(X/C)$ الاحتمال الشرطي لظهور الخصائص X عند الفئة C، $P(C)$ الاحتمال القبلي للفئة و $P(X)$ احتمال ظهور البيانات.

وفي نموذج Gaussian Naive Bayes، يتم افتراض أن توزيع كل خاصية يتبع التوزيع الطبيعي(Gaussian Distribution) ، وبالتالي يتم حساب الاحتمال الشرطي لكل خاصية باستخدام الدالة التالية:

$$P(x_i/C) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}$$

حيث:

μ_c: المتوسط الحسابي للخاصية داخل الفئة C

σ_c^2: التباين داخل الفئة C

x_i قيمة الخاصية

بعد حساب الاحتمالات لكل الخصائص، يتم دمجها تحت فرضية الاستقلال الشرطي للحصول على الاحتمال النهائي لكل فئة، ثم يتم تصنيف العينة إلى الفئة التي تمتلك أعلى احتمالposterior . يتميز هذا النموذج بالبساطة الحسابية العالية، وسرعة التنفيذ، إضافة إلى فعاليته في التعامل مع البيانات عالية الأبعاد، مما يجعله مناسبًا للعديد من التطبيقات الطبية مثل تصنيف الأمراض، بما في ذلك تشخيص سرطان الثدي، خاصة عند توفر بيانات كبيرة أو عندما تكون العلاقات بين المتغيرات غير معقدة بشكل كبير [14][13].

معايير التقييم والمقارنة

تم تقييم أداء نماذج التعلم الآلي المستخدمة في هذه الدراسة باستخدام مجموعة شاملة من المقاييس الإحصائية ومعايير التقييم الحديثة، وذلك لضمان دقة النتائج وموثوقيتها في مهمة تصنيف سرطان الثدي. تعتمد عملية التقييم على عدة مقاييس أساسية تُستخدم على نطاق واسع في مسائل التصنيف الثنائي، بالإضافة إلى مقاييس متقدمة تُستخدم خصيصًا في التطبيقات الطبية الحساسة، بهدف تحقيق تقييم شامل لأداء النماذج من مختلف الجوانب [16][15] .

١. الدقة الكلية Accuracy

تقيس الدقة الكلية نسبة العينات التي تم تصنيفها بشكل صحيح من إجمالي العينات، وتُعد من أبسط وأهم مقاييس التقييم.

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

TP الإيجابي الصحيح (True Positive): هي الحالات التي تم تصنيفها بشكل صحيح على أنها إيجابية (خبيثة)، بينما هي في الواقع إيجابية .

TNالسلبي الصحيح (True Negative): هي الحالات التي تم تصنيفها بشكل صحيح على أنها سلبية (حميدة)، بينما هي في الواقع سلبية .

FPالإيجابي الكاذب (False Positive): هي الحالات التي تم تصنيفها بشكل خاطئ على أنها إيجابية، بينما هي في الواقع سلبية .

FNالسلبي الكاذب (False Negative): هي الحالات التي تم تصنيفها بشكل خاطئ على أنها سلبية، بينما هي في الواقع إيجابية.

٢. الدقة الإيجابية Precision

تقيس Precision مدى دقة التنبؤات الإيجابية، أي نسبة الحالات الإيجابية الصحيحة من إجمالي الحالات التي تم التنبؤ بها كإيجابية.

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

٣. الاسترجاع Recall

يقيس هذا المقياس قدرة النموذج على اكتشاف الحالات الإيجابية الفعلية بشكل صحيح.

$$\text{Recall} = \frac{TP}{TP + FN}$$

٤. المتوسط التوافقي F1-score

يمثل F1-score التوازن بين Precision وRecall، ويُستخدم عندما يكون هناك حاجة إلى موازنة الأداء بين المقياسين.

$$F1 = 2 * \frac{precision * Recall}{precision + Recall}$$

٥. النوعية Specificity

تقيس Specificity قدرة النموذج على التعرف الصحيح على الحالات السلبية (الحميدة).

$$\text{Specificity} = \frac{TN}{TN + FP}$$

٦. معدل الإيجابيات الكاذبة False Positive Rate

يقيس هذا المقياس نسبة الحالات السلبية التي تم تصنيفها بشكل خاطئ على أنها إيجابية.

$$\text{FPR} = \frac{FP}{FP + TN}$$

٧. معدل السلبيات الكاذبة False Negative Rate

يقيس هذا المقياس نسبة الحالات الإيجابية التي لم يتم اكتشافها من قبل النموذج.

$$\text{FNR} = \frac{FN}{FN + TP}$$

٨. ROC Curve and AUC

يتم استخدام منحنى ROC لتقييم قدرة النموذج على التمييز بين الفئات المختلفة عبر مختلف عتبات القرار، بينما تمثل قيمة AUC المساحة تحت هذا المنحنى، والتي تعكس الأداء العام للنموذج في التصنيف.

٩. مصفوفة الالتباس Confusion Matrix

تُستخدم مصفوفة الالتباس لتحليل تفاصيل الأخطاء في التصنيف، من خلال توضيح عدد الحالات المصنفة بشكل صحيح أو خاطئ لكل فئة.(True Positive, True Negative, False Positive, False Negative)

١٠. التحقق المتقاطع Cross-Validation

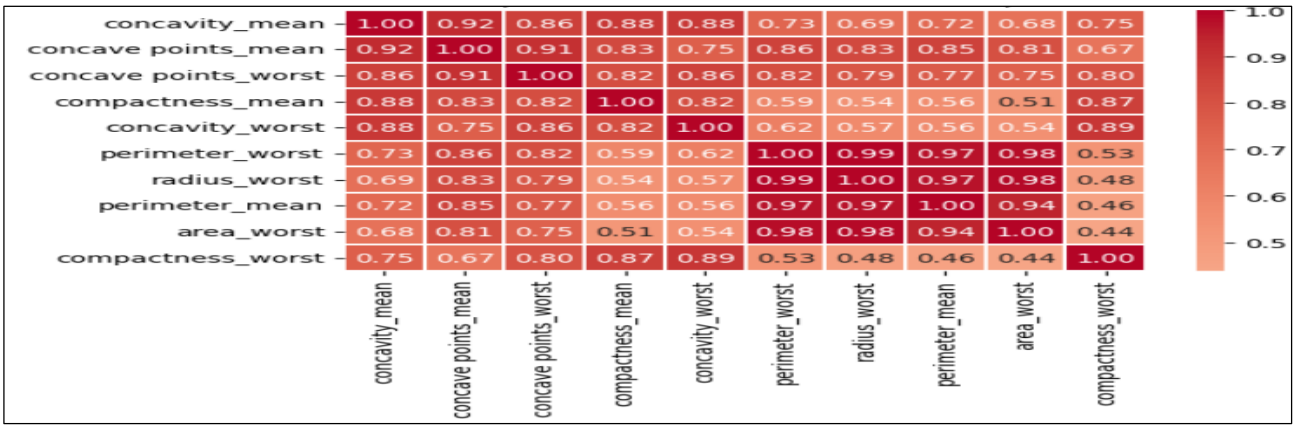
تم استخدام التحقق المتقاطع لضمان استقرار النتائج وتقليل مشكلة فرط التكيّف (Overfitting) ، مما يساعد على تحسين قدرة النموذج على التعميم على بيانات غير مرئية.

المحور الثاني: الجانب التطبيقي

اولاً. نظرة عامة على البيانات

تستخدم هذه الدراسة مجموعة بيانات Breast Cancer Wisconsin Diagnostic، والتي تُعد من أشهر مجموعات البيانات المستخدمة في أبحاث تصنيف سرطان الثدي في مجال تعلم الآلة. تم جمع هذه البيانات من خلال فحوصات تشخيصية لعينات من خلايا كتلة الثدي، حيث تحتوي على خصائص عددية مستخرجة من صور رقمية تمثل نواة الخلايا. تتكون مجموعة البيانات من 569 عينة وكل عينة تحتوي على 30خاصية عددية (Features)بالإضافة إلى متغير الهدف (Target) الذي يمثل التشخيص الطبي، حيث يتم تصنيف الحالات إلى نوعين: حميد (Benign)وخبيث (Malignant) تُستخدم هذه البيانات على نطاق واسع في الأبحاث الطبية والذكاء الاصطناعي نظرًا لجودتها العالية وخلوها من القيم المفقودة، مما يجعلها مناسبة لتطبيق نماذج تعلم الآلة ومقارنة أدائها.

يوضح الشكل رقم (1) خريطة حرارية (Heatmap) تمثل العلاقات والارتباطات بين المتغيرات (الخصائص) المختلفة في مجموعة بيانات سرطان الثدي. حيث تعكس الألوان وشدة القيم مدى قوة العلاقة بين كل زوج من المتغيرات، مما يساعد على فهم البنية العامة للبيانات واكتشاف العلاقات المهمة بينها. ونظرًا لكثرة عدد المتغيرات في مجموعة البيانات، فقد تم عرض جزء تمثيلي فقط من الخصائص لتبسيط العرض وتسهيل التفسير البصري، مع الحفاظ على وضوح الاتجاهات العامة للعلاقات بين المتغيرات



الشكل (1): التمثيل الحراري للعلاقات بين الخصائص في مجموعة بيانات سرطان الثدي

تم الحصول على هذه البيانات من مستودع UCI Machine Learning Repository، وهو أحد المصادر المعتمدة عالميًا في مجال تعلم الآلة والبحوث العلمية.

ثانياً: الإحصاءات الوصفية للخصائص

تم إجراء تحليل إحصائي وصفي لمجموعة البيانات بهدف فهم الخصائص الأساسية للمتغيرات المستخدمة في الدراسة. يوضح هذا التحليل القيم الإحصائية الرئيسية مثل المتوسط (Mean)، والانحراف المعياري (Standard Deviation) ، والقيم الدنيا (Min) ، والقيم العليا (Max) لكل متغير من متغيرات البيانات. تُظهر النتائج أن خصائص الخلايا المرتبطة بحجم النواة مثل radius mean و perimeter mean و area mean تمتلك تباينًا واضحًا بين العينات، مما يشير إلى وجود اختلافات مهمة بين الحالات الحميدة والخبيثة. كما تعكس هذه التباينات أهمية هذه المتغيرات في عملية التصنيف والتنبؤ. بشكل عام، يساعد هذا التحليل في فهم طبيعة البيانات قبل تطبيق نماذج تعلم الآلة، ويُعد خطوة أساسية لضمان جودة النمذجة ودقة النتائج.

يعرض الجدول (1) الإحصاءات الوصفية لأهم الخصائص المستخرجة من مجموعة بيانات سرطان الثدي، بما في ذلك المتوسط والانحراف المعياري والقيم الدنيا والعليا. وتوفر هذه الإحصاءات تصورًا عامًا حول توزيع المتغيرات وتباينها. نظرًا لكثرة عدد الخصائص (30 الخصائص)، تم عرض مجموعة تمثيلية من النتائج فقط في الجدول بهدف التوضيح وتقليل حجم العرض، بينما يمكن توفير النتائج الكاملة عند الحاجة. وتشير النتائج إلى وجود تباين ملحوظ بين الخصائص، خاصة تلك المرتبطة بحجم وشكل الورم مثل نصف القطر والمحيط والمساحة، مما يؤكد أهميتها في التمييز بين الحالات الحميدة والخبيثة

الجدول (1): الإحصاءات الوصفية للخصائص في مجموعة بيانات سرطان الثدي

	Feature	Mean	Std	Min	Max
1	Radius mean	14.127	3.524	6.98	28.1
2	Texture means	19.29	4.301	9.71	39.3
3	Perimeter mean	91.969	24.299	43.8	189
4	Area means	654.889	351.914	144	2501
5	Smoothness means	0.096	0.014	0.05	0.16

6	Compactness mean	0.104	0.053	0.02	0.35
7	Concavity mean				
	0.089	0.08	0	0.43	
⋮	⋮	⋮	⋮	⋮	⋮
25	Smoothness worst	0.132	0.023	0.07	0.22
26	Compactness worst	0.254	0.157	0.03	1.06
27	Concavity worst	0.272	0.209	0	1.25
28	concave points worst	0.115	0.066	0	0.29
29	Symmetry worst	0.29	0.062	0.16	0.66
30	Fractal dimension worst	0.084	0.018	0.06	0.21

نتائج نماذج التعلم الآلي

تم تقييم أداء نموذجي التعلم الآلي المستخدمين في هذه الدراسة، وهما نموذج XGBoost والنموذج البايزي (Naive Bayes)، باستخدام مجموعة بيانات سرطان الثدي. وقد تم الاعتماد على مجموعة من مقاييس التقييم القياسية لتحديد جودة النماذج في مهمة التصنيف الثنائي.

نتائج نموذج XGBoost

يُظهر الجدول (2) أداء نموذج XGBoost في مهمة تصنيف سرطان الثدي باستخدام مجموعة من مقاييس التقييم القياسية. حيث بلغت قيمة الدقة 0.956 (Accuracy) ، مما يشير إلى قدرة النموذج على تصنيف معظم العينات بشكل صحيح. كما حقق النموذج قيمة Precision مقدارها 0.952، وهو ما يدل على أن نسبة قليلة فقط من الحالات التي تم تصنيفها على أنها إيجابية كانت في الواقع سلبية، مما يعكس انخفاض معدل الإيجابيات الكاذبة. أما قيمة Recall فقد بلغت 0.930، مما يشير إلى قدرة جيدة للنموذج على اكتشاف الحالات الإيجابية الحقيقية (الأورام الخبيثة). وبالنسبة لمقياس F1-score الذي يجمع بين Precision وRecall، فقد سجل قيمة 0.941، مما يعكس توازنًا جيدًا بين الدقة والاسترجاع. كما حقق النموذج قيمة مرتفعة جدًا في AUC بلغت 0.993، مما يدل على قدرة قوية على التمييز بين الفئتين (حميد وخبيث).

بشكل عام، يعكس هذا الأداء قدرة نموذج XGBoost على التعامل بكفاءة مع بيانات التصنيف الطبي وتحقيق نتائج دقيقة ومستقرة.

الجدول (2): أداء نموذج XGBoost في تصنيف سرطان الثدي

المقياس	Accuracy	Precision	Recall	F1-score	AUC
القيمة	0.96	1	0.93	0.941	0.993

نتائج النموذج البايزي(Naive Bayes)

يوضح الجدول (3) نتائج أداء نموذج Naive Bayes (Bayesian Model) في تصنيف حالات سرطان الثدي. حيث حقق النموذج دقة (Accuracy) عالية بلغت 0.965، مما يدل على أداء عام قوي في تصنيف العينات بشكل صحيح.

كما سجل النموذج قيمة Precision مرتفعة بلغت 0.976، وهو ما يشير إلى انخفاض معدل الإيجابيات الكاذبة بشكل ملحوظ، مما يعكس دقة عالية في التنبؤ بالحالات الإيجابية. أما قيمة Recall فقد بلغت 0.930، وهي مماثلة تقريبًا لنموذج XGBoost، مما يدل على قدرة جيدة على اكتشاف الحالات الإيجابية الحقيقية.

وقد حقق النموذج قيمة F1-score بلغت 0.952، وهو ما يعكس توازنًا جيدًا بين الدقة والاسترجاع، بينما وصلت قيمة AUCإلى 0.997، وهي قيمة مرتفعة جدًا تشير إلى قدرة ممتازة على التمييز بين الفئات المختلفة.

بشكل عام، يُظهر النموذج البايزي أداءً قويًا ومتفوقًا قليلاً على XGBoost في بعض المقاييس، خصوصًا في الدقة الكلية وAUC، مما يجعله مناسبًا جدًا لهذه المهمة التصنيفية.

الجدول (3): أداء النموذج البايزي في تصنيف سرطان الثدي

المقياس	Accuracy	Precision	Recall	F1-score	AUC
القيمة	0.965	0.976	0.93	0.95	0.997

تحليل مصفوفة الالتباس Confusion Matrix) (Analysis

في هذا القسم يتم تقييم أداء نموذجي التعلم الآلي **XGBoost** و **Naive Bayes** باستخدام مصفوفة الالتباس (Confusion Matrix)، والتي تُعد من أهم الأدوات المستخدمة في تقييم نماذج التصنيف، حيث توضح عدد العينات التي تم تصنيفها بشكل صحيح أو خاطئ لكل فئة (حميد وخبيث).

نموذجXGBoost

مصفوفة الالتباس الخاصة بنموذج XGBoost هي كما يلي:

$$\begin{bmatrix} 69 & 2 \\ 3 & 40 \end{bmatrix}$$

تُظهر النتائج أن نموذج XGBoost تمكن من تصنيف معظم العينات بشكل صحيح، حيث تم تسجيل 69 حالة صحيحة للفئة السلبية (True Negative) و40 حالة صحيحة للفئة الإيجابية (True Positive). في المقابل، تم تسجيل عدد قليل

من الأخطاء، حيث توجد حالتان من الإيجابيات الكاذبة (False Positive) و3 حالات من السلبيات الكاذبة (False Negative).
تشير هذه النتائج إلى أن النموذج يتمتع بأداء مستقر وقدرة جيدة على التمييز بين الحالات الحميدة والخبيثة، مع معدل خطأ منخفض نسبيًا.

النموذج البايزي(Naive Bayes)

أما مصفوفة الالتباس الخاصة بالنموذج البايزي فهي كما يلي:

$$\begin{bmatrix} 70 & 1 \\ 3 & 40 \end{bmatrix}$$

تُظهر النتائج أن النموذج البايزي حقق أداءً قويًا جدًا، حيث تمكن من تصنيف 70 حالة بشكل صحيح ضمن الفئة السلبية، و40 حالة ضمن الفئة الإيجابية. كما سجل النموذج حالة واحدة فقط كإيجابي كاذب (False Positive)، وثلاث حالات كسلبيات كاذبة (False Negative) . تُشير هذه النتائج إلى أن النموذج البايزي يتميز بدقة عالية في تقليل الأخطاء الإيجابية الكاذبة، مما يجعله مناسبًا للتطبيقات الطبية التي تتطلب تقليل احتمالية التشخيص الخاطئ للحالات السليمة.

بشكل عام، يوضح تحليل مصفوفة الالتباس أن كلا النموذجين قد حققا أداءً جيدًا في مهمة تصنيف سرطان الثدي. ومع ذلك، يمكن ملاحظة تفوق طفيف للنموذج البايزي من حيث تقليل الإيجابيات الكاذبة، في حين أظهر نموذج XGBoost أداءً متوازنًا وقويًا مع دقة عالية في كلا الفئتين. وبناءً على هذه النتائج، يمكن القول إن كلا النموذجين مناسبان لهذه المهمة التصنيفية، مع اختلافات طفيفة في نمط الأخطاء لكل نموذج.

مقارنة أداء النماذج(Model Comparison)

تم في هذا القسم إجراء مقارنة شاملة بين نموذجي التعلم الآلي **XGBoost** و **Naive Bayes** باستخدام مجموعة من المقاييس المتقدمة التي تعكس الأداء الحقيقي للنماذج، خصوصًا في التطبيقات الطبية التي تتطلب دقة عالية وتقليل الأخطاء الحرجة.

الجدول (4): مقارنة أداء النماذج

النموذج	Specificity	Balanced Accuracy	F. P. Rate	F. N. Rate
XGBoost	0.972	0.951	0.028	0.07
Bayesian	0.99	0.967	0.01	0.07

تُظهر النتائج في الجدول (4) أن كلا النموذجين قد حققا أداءً قويًا في مهمة تصنيف سرطان الثدي، إلا أن هناك اختلافات واضحة في طبيعة الأداء بينهما. بالنسبة لمقياس Specificity النوعية، فقد تفوق نموذج Naive Bayes بشكل واضح، حيث سجل قيمة 0.990 مقارنة بـ 0.972 لنموذج XGBoost، مما يدل على قدرة أعلى للنموذج البايزي على تصنيف الحالات السليمة بشكل صحيح وتقليل عدد الإيجابيات الكاذبة (False Positives). أما Balanced Accuracy فقد أظهر النموذج البايزي أيضًا أداءً أفضل بقيمة 0.967 مقابل 0.951، مما يعكس توازنًا أفضل بين الحساسية والنوعية.

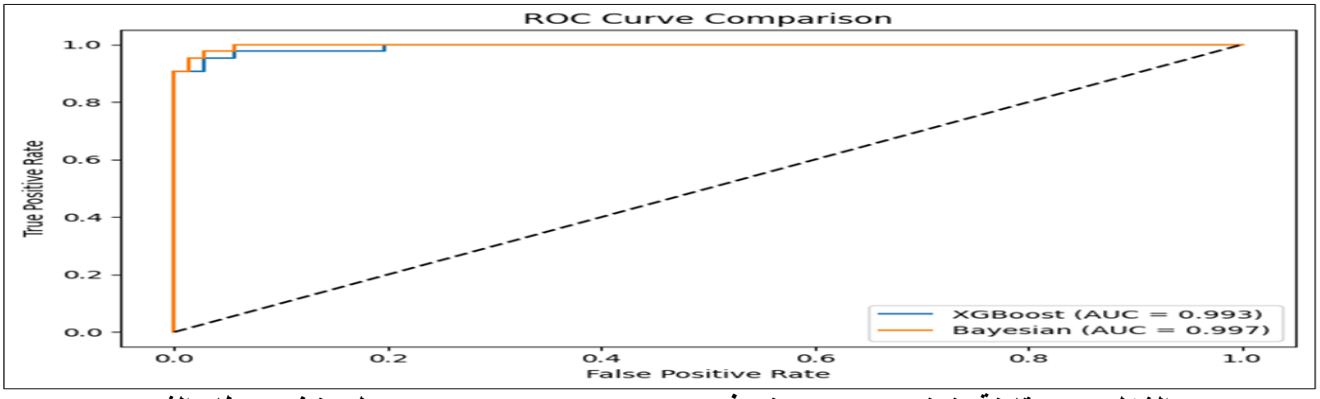
فيما يتعلق بـ False Positive Rate (FPR)، سجل نموذجNaive Bayes أقل قيمة (0.010)، وهو ما يُعد ميزة مهمة جدًا في التطبيقات الطبية، حيث أن تقليل التشخيص الخاطئ للحالات السليمة يقلل من القلق والإجراءات الطبية غير الضرورية. في المقابل، كان أداء XGBoost أقل قليلًا مع FPR يساوي 0.028. أما False Negative Rate (FNR) فقد كان متساويًا تقريبًا بين النموذجين، مما يشير إلى قدرة متقاربة في اكتشاف الحالات الإيجابية (الخبيثة).

بشكل عام، يمكن ملاحظة أن نموذج Naive Bayes يتفوق من حيث تقليل الأخطاء الإيجابية الكاذبة وتحقيق توازن أفضل في الأداء، مما يجعله أكثر ملاءمة في التطبيقات الطبية الحساسة. في المقابل، يُظهر نموذج XGBoost أداءً قويًا ومستقرًا مع نتائج متقاربة، مما يجعله خيارًا موثوقًا أيضًا في مهام التصنيف. وبناءً على هذه النتائج، يمكن القول إن اختيار النموذج يعتمد على طبيعة التطبيق: فإذا كان الهدف تقليل التشخيص الخاطئ للحالات السليمة، فإن Naive Bayes يكون الخيار الأفضل، بينما XGBoost يوفر أداءً متوازنًا وقويًا في مختلف المقاييس.

نتائج منحنى ROC

يوضح الشكل رقم (2) نتائج منحنى خاصية التشغيل للمستقبل (ROC Curve) المستخدم لتقييم الأداء التمييزي لنموذجي التعلم الآلي XGBoost : ونموذج بايز (Naive Bayes)في مهمة تصنيف سرطان الثدي. تُظهر النتائج أن كلا النموذجين حققا أداءً ممتازًا في التمييز بين الحالات الحميدة والخبيثة، حيث بلغت قيمة المساحة تحت المنحنى (AUC) لنموذج XGBoost حوالي 0.993، مما يعكس قدرة عالية جداً على التصنيف الصحيح وتقليل الأخطاء في التنبؤ. في المقابل، حقق نموذج بايز قيمة AUC أعلى قليلاً بلغت 0.997، مما يشير إلى أداء تمييزي أفضل بشكل طفيف مقارنة بنموذج XGBoost، وقدرته العالية على الفصل بين الفئات المختلفة بدقة كبيرة.

بشكل عام، تُظهر النتائج أن كلا النموذجين يتمتعان بكفاءة عالية جداً في التصنيف، مع تفوق بسيط للنموذج البايزي من حيث قيمةAUC ، مما يؤكد موثوقية كلا النموذجين في التطبيقات الطبية الحساسة مثل تشخيص سرطان الثدي.



الشكل (2): مقارنة منحنى ROC بين نموذجي XGBoost و Naive Bayesلتصنيف سرطان الثدي

تحليل أهمية الخصائص باستخدام نموذج (Feature Importance Analysis)

يوضح الجدول أهمية الخصائص (Feature Importance)المستخرجة من نموذج XGBoost ، حيث يبين مدى تأثير كل متغير في عملية اتخاذ القرار الخاصة بتصنيف حالات سرطان الثدي. تُظهر النتائج أن خاصية concave points meanكانت الأكثر تأثيرًا في النموذج، تليها perimeter worstو concave points worst، مما يشير إلى أن الخصائص المرتبطة بشكل حدود الخلايا وحجم الورم تلعب دورًا رئيسيًا في التمييز بين الحالات الحميدة والخبيثة. كما يتضح أن مجموعة من الخصائص الأخرى مثل radius worst و area worstو radius mean تساهم بشكل ملحوظ في تحسين دقة النموذج، بينما تظهر بعض الخصائص ذات تأثير أقل نسبيًا مثل fractal-dimension-worst وsymmetry mean.

الجدول 5 .درجات أهمية الخصائص المستخرجة من نموذجXGBoost

Feature	Importance	Feature	Importance
concave points mean	0.259	Radius se	0.010
perimeter worst	0.178	Concavity means	0.010
concave points worst	0.171	Smoothness worst	0.008
Radius worst	0.054	Compactness se	0.008
Area worst	0.045	Fractal dimension mean	0.007
Radius mean	0.040	Symmetry worst	0.007
Concavity worst	0.026	Texture se	0.005
Texture worst	0.025	Smoothness means	0.005
Area means	0.024	Compactness means	0.005
Concavity se	0.020	concave points se	0.004
Texture means	0.019	Fractal dimension worst	0.004
Perimeter se	0.019	Perimeter mean	0.004
Smoothness se	0.011	Symmetry se	0.004
Area se	0.011	Symmetry means	0.003
Fractal dimension se	0.011	Compactness worst	0.002

بشكل عام، تعكس هذه النتائج قدرة نموذج XGBoost على تحديد أهم المتغيرات المؤثرة في عملية التصنيف، مما يعزز قابلية تفسير النموذج ويدعم استخدامه في التطبيقات الطبية التشخيصية.

الاستنتاجات والتوصيات

أولا: الاستنتاجات:

تقدم هذه الدراسة نهجًا قائمًا على تعلم الآلة لتصنيف حالات سرطان الثدي باستخدام نموذجي XGBoostو Naive Bayes، اعتمادًا على مجموعة بيانات Breast Cancer Wisconsin Diagnostic. وقد تم تقييم أداء النماذج باستخدام مجموعة شاملة من المقاييس الإحصائية، بما في ذلك Accuracyو Precisionو Recall وF1-score، بالإضافة إلى مقاييس متقدمة مثل Specificity و ROC-AUC ومصفوفة الالتباس.

أظهرت النتائج أن كلا النموذجين حققا أداءً عاليًا في مهمة التصنيف، حيث سجل نموذج XGBoost دقة قوية واستقرارًا جيدًا في التنبؤ، بينما أظهر نموذج Naive Bayes أداءً متميزًا

بشكل طفيف من حيث قيمة AUC التي بلغت 0.997 مقارنة بـ 0.993 لنموذج XGBoost، مما يعكس قدرة أعلى على التمييز بين الفئات المختلفة. كما بين تحليل مصفوفة الالتباس انخفاض معدل الأخطاء في كلا النموذجين، مع قدرة جيدة على اكتشاف الحالات الإيجابية والسلبية. إضافة إلى ذلك، أوضح تحليل أهمية الخصائص أن المتغيرات المتعلقة بشكل وحجم الورم مثل concave points mean وperimeter worst و radius worst تلعب الدور الأهم في عملية التصنيف، مما يعزز قابلية تفسير النموذج من الناحية الطبية.

بشكل عام، تؤكد النتائج أن نماذج تعلم الآلة، وخاصة XGBoost وNaive Bayes، توفر أدوات فعالة وموثوقة لدعم التشخيص الطبي المبكر لسرطان الثدي. ويمكن اعتبار هذه النماذج مساعدة قوية للخبراء الطبيين في تحسين دقة التشخيص وتقليل الأخطاء البشرية.

ثانيا: التوصيات:

1. اعتماد تقنيات تعلم الآلة في أنظمة دعم القرار الطبي، لما أظهرته من كفاءة عالية في تحسين دقة تصنيف وتشخيص سرطان الثدي .
2. الاستفادة من نموذج Gaussian Naive Bayes في تطبيقات التشخيص المبكر لسرطان الثدي، نظرًا لتحقيقه أداءً متميزًا في معظم
3. مؤشرات التقييم مقارنةً بخوارزمية XGBoost في هذه الدراسة .توسيع نطاق الدراسات المستقبلية من خلال تطبيق نماذج تعلم الآلة على مجموعات بيانات طبية أكبر وأكثر تنوعًا، للتحقق من قدرة النماذج على التعميم في بيئات سريرية مختلفة .
4. إجراء مقارنات مع خوارزميات تعلم آلة أخرى مثل Random Forest و Support Vector Machine وLightGBM وCatBoost، بهدف تحديد النموذج الأكثر كفاءة في تشخيص سرطان الثدي .
5. دمج البيانات السريرية وبيانات التصوير الطبي والبيانات الجينية مع نماذج تعلم الآلة، لتطوير أنظمة تشخيص أكثر شمولًا ودقة .
6. تشجيع التعاون بين الباحثين في علوم الحاسوب والكوادر الطبية لتطوير أنظمة ذكية موثوقة تدعم التشخيص المبكر، وتسهم في تحسين جودة الرعاية الصحية وتقليل الأخطاء التشخيصية.

توافر البيانات:

تم تضمين البيانات المستخدمة لدعم نتائج هذه الدراسة في المقالة.

تضارب المصالح:

ليس لدينا تضارب في المصالح.

موارد التمويل:

لم نتلقى اي دعم مالي.

شكر وتقدير:

لا أحد.

References:

1. Wang, Y., Zhang, H., & Liu, J. (2024). Breast cancer classification using gradient boosting-based machine learning models. Journal of Medical Artificial Intelligence, 12(3), 145–158.
2. Li, X., Chen, M., & Zhao, L. (2024). Probabilistic graphical models for clinical decision-making: A Bayesian network approach in breast cancer analysis. Computational Biology and Medicine, 156, 106–118.
3. Khan, M. A., Ahmed, S., & Rehman, A. (2025). Predictive modeling of breast cancer using XGBoost and genomic data analysis. IEEE Access, 13, 22145–22160.
4. Alvarez, R., Gomez, P., & Martinez, J. (2024). Comparative analysis of machine learning algorithms for breast cancer diagnosis. Expert Systems with Applications, 238, 122–134.
5. Chen, Y., Wang, L., & Huang, X. (2025). Hybrid deep learning and Bayesian models for interpretable medical diagnosis systems. Artificial Intelligence in Medicine, 148, 102–115.
6. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Breast cancer Wisconsin (diagnostic) dataset. UCI Machine Learning Repository.
7. Patel, R., Singh, K., & Mehta, D. (2024). Explainable AI approaches for breast cancer classification using ensemble learning methods. Biomedical Signal Processing and Control, 89, 105612.
8. Zhou, Q., Li, H., & Sun, Y. (2024). Machine learning-based diagnostic systems for breast cancer prediction: A comparative study. Computers in Biology and Medicine, 170, 108214.
9. Rahman, F., & Kim, J. (2025). Integrating Bayesian inference with machine learning for robust medical diagnosis. IEEE Transactions on Computational Biology and Bioinformatics, 22(1), 88–101.

medical image-based breast cancer detection. IEEE Access, 13, 98765–98780.

14. El-Sayed, A., & Mahmoud, H. (2024). Bayesian approaches for uncertainty modeling in medical classification problems. Journal of Biomedical Informatics, 150, 104512.

15. Kim, D., Park, J., & Lee, S. (2025). Performance evaluation of machine learning models for cancer diagnosis using ROC-AUC analysis. Computational and Structural Biotechnology Journal, 23, 456–468.

16. Brown, T., Williams, J., & Davis, R. (2024). Machine learning interpretability in healthcare: A focus on feature importance methods. Health Informatics Journal, 30(2), 789–805.

10. Garcia, M., Lopez, A., & Hernandez, S. (2024). Feature selection and importance analysis in breast cancer datasets using tree-based models. Knowledge-Based Systems, 285, 111345.

11. Singh, P., & Verma, R. (2025). Comparative performance analysis of XGBoost and machine learning classifiers in healthcare applications. Expert Systems with Applications, 250, 123456.

12. Hassan, M., Ali, N., & Khan, S. (2024). Interpretable machine learning for cancer diagnosis using clinical and imaging data. Artificial Intelligence in Medicine, 145, 102789.

13. Wang, H., Liu, Z., & Zhang, T. (2025). Advanced ensemble learning techniques for