Breast Cancer Detection Using Machine Learning Techniques

Muhammed Hikmat Muhammed*

(Received 2 / 11 / 2024. Accepted 4 / 12 / 2024)

□ ABSTRACT □

Breast cancer affects approximately 10% of women worldwide at some point in their lives and has emerged as one of the most feared and prevalent cancers among women. The main dilemma arises when cancer cannot be properly detected in its early stages. Machine learning has proven to play a vital role in diagnosing diseases such as cancers. Effective methods for classification and data recognition are particularly essential in the medical field.

In this project, classification techniques were employed using the PCA-KNN algorithm on the Wisconsin Breast Cancer Dataset. The main objective was to evaluate the accuracy of data classification concerning the efficiency and effectiveness of the PCA-KNN algorithm in terms of precision, recall, specificity, and the F1 score. The experimental results demonstrated an accuracy of up to 99%.

Keywords: Machine learning , KNN, PCA, Breast cancer.

* **Master's - Industrial Automation Engineering – Tartus University – Tartus - Syria.**

الكشف عن سرطان الثدي باستخدام تقنيات تعلم الآلة

محمد حكمات محمد *

(تاريخ الإيداع 2 / 11 / 2024. قُبِل للنشر في 4/ 12 / 2024)

□ ملخّص □

سرطان الثدي هو أحد أكثر أنواع السرطانات شيوعًا بين النساء، حيث يصيب حوالي 10% منهن خلال حياتهن، مما يجعله مصدرًا رئيسيًا للقلق الصحي عالميًا. تتفاقم المشكلة عندما يتأخر التشخيص في المراحل المبكرة، مما يقلل من فرص العلاج. تسهم تقنيات تعلم الآلة بشكل كبير في تحسين دقة التشخيص وتقديم حلول مبتكرة في مجال الرعاية الصحية.

تم في هذا البحث استخدام خوارزمية PCA–KNN (تحليل المكونات الرئيسية – الجار الأقرب) لتصنيف بيانات سرطان الثدي من مجموعة بيانات جامعة .Wisconsin استُخدمت هذه الخوارزمية لتحسين دقة التصنيف عبر تقليل أبعاد البيانات وتحديد الأنماط ذات الصلة. يهدف البحث إلى تقييم أداء النموذج بناءً على مقاييس مثل الدقة، الحساسية، التحديد، ودرجة .F1 أظهرت النتائج التجريبية قدرة النموذج على تحقيق دقة تصل إلى 99%، مما يبرز كفاءة هذا النهج مقارنة بالأساليب التقليدية.

يوفر هذا البحث أداة فعّالة للتشخيص الطبي تسهم في تقليل الحاجة إلى الإجراءات الجراحية غير الضرورية، وتزيد من دقة التشخيص المبكر، مما يعزز فرص العلاج. كما يوصي البحث بتوسيع قاعدة البيانات المستخدمة واختبار نماذج تعلم عميق لتحليل الأداء بالمقارنة مع الخوارزميات الحالية. تشير النتائج إلى إمكانية تطبيق هذا النموذج كجزء من منظومة شاملة للكشف المبكر عن سرطان الثدي وتحسين خدمات الرعاية الصحية عالميًا.

الكلمات المفتاحية: سرطان الثدي، تعلم الآلة، PCA، KNN، تشخيص الأمراض.

* **ماجستير – قسم هندسة الأتمتة الصناعية – كلية الهندسة التقنية – جامعة طرطوس – طرطوس – سورية.**

مقدمة:

يعد سرطان الثدي من أخطر الأمراض التي تصيب النساء حول العالم وأكثرها فتكاً، تم تصنيف سرطان الثدي في المرتبة الأولى بين النساء الهنديات، ويكون لديها معدل إصابة متكيف حسب العمر يصل إلى 25.8 لكل 100.000 امرأة ومعدل وفيات يبلغ 12.7 لكل 100.000 امرأة، وتشير توقعات سرطان الثدي في الهند في عام 2020 إلى أن العدد سيرتفع إلى 1797900، وقد يؤدي تحسين الوعي الصحي وتوافر برامج الكشف عن سرطان الثدي ومرافق العلاج إلى ظهور صورة سريرية مواتية وإيجابية في البلد.

وفقاً لجمعية السرطان الأمريكية (ACS)، فإن سرطان الثدي هو أكثر أنواع السرطانات شيوعاً عند النساء الأمريكيات، بعد سرطانات الجلد، يبلغ متوسط خطر إصابة امرأة في الولايات المتحدة بسرطان الثدي في وقت ما من حياتها حوالي 12% أو احتمال 1 من كل 8، وتبلغ نسبة احتمال وفاة المرأة بسبب سرطان الثدي حوالي 2.6% أو فرصة واحدة من بين 38، ووفقاً لتقرير منظمة الصحة العالمية (WHO) لعام 2013، في عام 2011 توفيت حوالي نصف مليون امرأة في جميع أنحاء العالم نتيجة الإصابة بسرطان الثدي، حيث أنّ معدل الوفيات من سرطان الثدي مرتفع للغاية مقارنة بأنواع السرطان الأخرى.

إنّ التشخيص المبكر يزيد بشكل كبير من نجاح العلاج، على وجه التحديد أثناء إجراء التشخيص، حيث يقوم المتخصصون بتقييم كل من التنظيم العام والمحلي للأنسجة من خلال صور شريحة كاملة وصور مجهرية، ومع ذلك، فإن الكمية الكبيرة من البيانات وتعقيد الصور تجعل هذه المهمة تستغرق وقتاً طويلاً وغير اعتيادي، نتيجة لذلك يعد تطوير أدوات الكشف والتشخيص التلقائي أمراً صعباً وضرورياً في هذا المجال، يتكون تشخيص سرطان الثدي من اختبار فحص مثل التصوير الشعاعي للثدي أو الموجات فوق الصوتية كلما تم الكشف عن كتلة، يليها خزعة وفحص الأنسجة المرضية لعمل تشخيص واضح لأي نمو خبيث في أنسجة الثدي، يمكن أن تكون الأنسجة طبيعية أو حميدة أو خبيثة، ويمكن تصنيف الآفات الخبيثة على أنها إما "موضعية، حيث يتم تقييد الخلايا داخل نظام الأقنية – الفصيص الثديي، أو "غازية"، حيث تنتشر الخلايا خارج بنية الثدي [1]. لتحقيق الاستفادة القصوى من إمكانات تقنيات تعلم الآلة في تشخيص سرطان الثدي، يسعى هذا البحث إلى تقديم نموذج عملي يتميز بالمرونة وقابلية التطبيق في البيئات الطبية المختلفة. من خلال التركيز على خوارزميةPCA–KNN ، يهدف العمل إلى ليس فقط تحسين دقة التصنيف ولكن أيضًا تقديم نهج يمكن دمجه بسهولة مع الأدوات الطبية الحالية. إنّ تعزيز موثوقية التشخيص وتقليل الحاجة إلى تدخلات جراحية غير ضرورية يمكن أن يساهم بشكل كبير في تحسين جودة حياة المرضى وتقليل الأعباء المالية والطبية على الأنظمة الصحية. كما يفتح هذا البحث آفاقًا لتوسيع استخدام النماذج التنبؤية في مجالات أخرى من الرعاية الصحية، ما يعزز من أهمية البحث كجزء من التطور المستقبلي للتكنولوجيا الطبية.

مشكلة البحث:

وفقًا لإحصائيات منظمة الصحة العالمية(WHO) ، يُعد سرطان الثدي من أكثر الأمراض تهديدًا لحياة النساء بسبب التأخر في التشخيص المبكر. يؤدي هذا التأخير إلى انتشار المرض في الجسم، مما يقلل فرص العلاج ويزيد من معدلات الوفيات.

لقد أدى استخدام تقنيات تعلم الآلة وتنقيب البيانات إلى تغيير عملية تشخيص مرض سرطان الثدي بأكملها. تشخيص سرطان الثدي باستخدام تقنية تعلم الآلة يمكن من تمييز أورام الثدي الحميدة من الخبيثة. تعتبر معظم طرق تنقيب

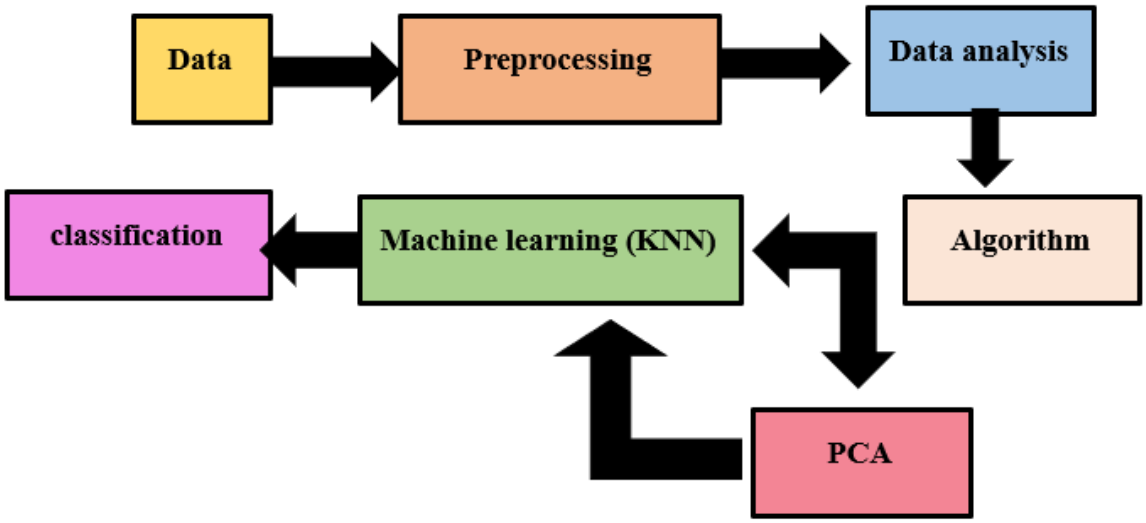
البيانات المستخدمة بشكل شائع في هذا المجال فئة تصنيف Classification التي تقوم بالتشخيص التطبيقي وذلك بتصنيف المرضى إما لمجموعة "حميدة" غير سرطانية أو مجموعة "خبيثة " سرطانية
هذا البحث يتناول تشخيص مرض سرطان الثدي باستخدام تقنيات تعلم الآلة بالتطبيق قاعدة بيانات من جامعة Wisconsin .

أهمية البحث وأهدافه:

إن استخدام الأنظمة الحاسوبية في هذا المجال أصبح أمراً في غاية الأهمية نظراً لأنه سيعطي نتائج أكثر دقة تساعد العنصر البشري في تحسين النتائج وتخفيض معدل الخطأ، والذي من الممكن أن يساعد في إنقاذ أرواح بشرية.
يهدف البحث إلى بناء خوارزمية للكشف عن سرطان الثدي باستخدام تقنيات تعلم الآلة والذي يمكن أن يحل محل طرق الفحص البصرية الحالية من قبل الطبيب ويساهم في تخفيض عدد الخزعات الجراحية غير الضرورية وزيادة الموثوقية وذلك بالمساعدة على التشخيص الصحيح.

طرائق البحث ومواده:

تم تنفيذ البحث باستخدام لغة البرمجة Python في بيئةVisual Studio Code ، على جهاز حاسوب بمواصفات: معالج Intel Core i5-8250U بسرعة 1.80 GHz وذاكرة عشوائية GB.8 .
يتضمن البحث عدة مراحل بدءاً من الحصول على البيانات وصولاً إلى مرحلة التشخيص من خلال بناء نموذج تعلم الآلة مع الاستعانة بمكتبات Keras و Tansorflow من أجل بناء النموذج ويبين الشكل (1) المخطط الصندوقي لأهم مراحل العمل خلال هذا البحث.



الشكل (1) المخطط الصندوقي لمراحل العمل

الدراسات المرجعية:

- ✓ في دراسة أجرتها Hiba Asri وزملاؤها عام 2016 [2]، تمت مقارنة أداء خوارزميات تعلم الآلة المختلفة (Decision Tree, SVM, Naive Bayes, KNN)في تشخيص سرطان الثدي باستخدام قاعدة بيانات Wisconsin.ركزت الدراسة على تقييم كفاءة كل خوارزمية من حيث الدقة، الحساسية، والخصوصية، وذلك لتحليل

مدى موثوقية هذه التقنيات في التنبؤ بحالة المريض. أظهرت النتائج التجريبية تفوق خوارزميةSVM ، التي حققت أعلى دقة بلغت 97.13% مع أقل معدل خطأ مقارنة بالخوارزميات الأخرى. تم تنفيذ جميع التجارب باستخدام أداة التنقيب عن البياناتWEKA ، مما عزز من موثوقية التحليل واستقرار النتائج.

✓ في عام 2018، استعرض Puneet Yadav وآخرون [3] فعالية خوارزميات تعلم الآلة مثل شجرة القرار (Decision Tree)وخوارزمية SVM لتشخيص سرطان الثدي. أوضحت نتائج الدراسة أن شجرة القرار حققت دقة تراوحت بين 90% و94%، بينما تفوقت خوارزمية SVM بدقة تراوحت بين 94.5% و97%. أكدت الدراسة أهمية تحسين خوارزميات التصنيف لتقليل الأخطاء وزيادة الثقة في أنظمة التشخيص.

✓ في عام 2019، قام Akshya Yadav وزملاؤه [4] بتطوير نموذج يعتمد على خوارزمية SVM لتحليل احتمالية إصابة الأفراد بسرطان الثدي. ركزت الدراسة على تحسين أداء الخوارزمية باستخدام طريقة PCA لتحليل المكونات وتقليل أبعاد البيانات. ساعدت هذه الطريقة في تقليل أثر فرط التعلم والقيم المتطرفة، مما أدى إلى رفع دقة الخوارزمية من 95.1% إلى 96.5%. استُخدمت مجموعة بيانات من جامعةWisconsin ، مما يعزز من موثوقية النتائج وإمكانية تطبيقها على نطاق واسع.

قاعدة البيانات المستخدمة:

تم تطوير مجموعة البيانات المستخدمة في هذا المشروع بواسطة الدكتور H. Wolberg من مستشفى جامعة Wisconsin، الولايات المتحدة الأمريكية، بهدف تسهيل تشخيص سرطان الثدي باستخدام تقنيات تعلم الآلة. اشتملت البيانات على عينات مأخوذة من سوائل المرضى الذين يعانون من كتل ثدي صلبة، حيث تم تحليلها باستخدام برنامج رسومي متخصص يسمى **.Xcyt**يتميز البرنامج بقدرته على إجراء تحليل الخصائص الخلوية رقميًا باستخدام خوارزميات ملائمة للمنحنى، ما يتيح استخراج عشر ميزات أساسية من كل خلية في العينة. تشمل هذه الميزات المتوسط والقيمة القصوى والخطأ القياسي لكل خاصية، مما يولد 30 متجهًا عدديًا لكل حالة.

تحتوي قاعدة البيانات على 569 عينة، منها 357 حالة حميدة و212 حالة خبيثة، مع 32 عمودًا: العمود الأول هو رقم الحالة، والثاني يمثل نتيجة التشخيص (حميدة أو خبيثة)، بينما تتضمن الأعمدة الأخرى متوسطات وخصائص إحصائية للعشرة ميزات. لا تحتوي مجموعة البيانات على قيم مفقودة، مما يعزز موثوقيتها في التحليل واستخدامها لتدريب النماذج.[5]

خوارزمية الجار الأقرب:(KNN)

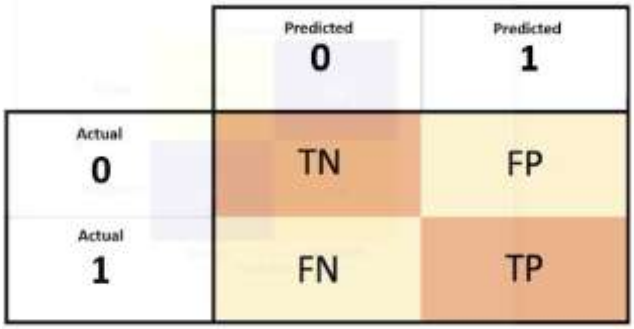
تعد خوارزمية الجار الأقرب (KNN) من أبسط خوارزميات التعلم تحت الإشراف، حيث تعتمد على افتراض أن العناصر المتشابهة تكون قريبة في الخصائص. تتميز الخوارزمية بسهولة تنفيذها وقلة متطلباتها الزمنية مقارنة بغيرها. تستخدم خوارزمية KNN البيانات الكاملة كمجموعة تدريب دون تقسيمها، حيث تعمل على تصنيف العناصر الجديدة بناءً على قربها من **k**أقرب الجيران المحددين مسبقًا. يتم تحديد قيمة **k**من قبل المستخدم، ويتم قياس التشابه باستخدام مقاييس مثل المسافة الإقليدية أو مسافةHamming ، مما يجعلها فعالة في التعامل مع التصنيفات الأولية.[6]

مصفوفة الارتباك:

مصفوفة الارتباك، التي تعرف أيضًا باسم مصفوفة الخطأ، هي أداة تُستخدم لتقييم أداء النماذج التصنيفية. تعرض المصفوفة العلاقة بين النتائج الحقيقية والنتائج التي يتنبأ بها النموذج، مما يجعلها أداة فعالة لتحديد الأخطاء وتصنيفها. تحتوي المصفوفة على المؤشرات التالية:[7]

1. **الفئة الإيجابية الحقيقية (True Positive – TP):** الحالات المصنفة بشكل صحيح كإيجابية.
2. **الفئة الإيجابية الكاذبة (False Positive – FP):** الحالات المصنفة بشكل خاطئ كإيجابية.
3. **الفئة السلبية الحقيقية (True Negative – TN):** الحالات المصنفة بشكل صحيح كسلبية.
4. **الفئة السلبية الكاذبة (False Negative – FN):** الحالات المصنفة بشكل خاطئ كسلبية.

تعتبر مصفوفة الارتباك أداة أساسية لفهم الأخطاء النموذجية وتحسين الأداء العام للنموذج، خاصة في التطبيقات الطبية الحساسة مثل تشخيص سرطان الثدي.[7]



الشكل (2): الشكل العام لمصفوفة الالتباس

معايير التقييم:

1. الصحة(accuracy): وهي عدد العينات التي تم توقعها بشكل صحيح على العدد الإجمالي للعينات وتعطى بالعلاقة [8]:

$$Acc = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (1)$$

2. الحساسية (recall or sensitivity): وهي عدد العينات الصحيحة التي تم كشفها على عدد العينات الصحيحة الإجمالي، وتعطى بالعلاقة[8]:

$$SN = \frac{Tp}{Tp + Fn} \quad (2)$$

3. الخصوصية (Specificity): عدد العينات الخاطئة التي تم كشفها على العدد الإجمالي للعينات الخاطئة، وتعطى بالعلاقة[9]:

$$SP = \frac{Tn}{Tn + Fp} \quad (3)$$

4. المقياس (F1–score): ويعبر عن العلاقة التوافقية بين المقياسيين (precision) و(recall)، ويعطى بالعلاقة[9]:

$$F1 - score = \frac{2 * Tp}{2 * Tp + Fn + Fp} \quad (4)$$

المعالجة المسبقة للبيانات:

قبل أن نتعرف على محتوى البيانات، نحتاج أولاً إلى النظر في الهيكل العام للبيانات، أي عدد الصفوف (نقاط البيانات) وعدد الأعمدة (الميزات) فيها والجدول (1) يوضح عدد الصفوف وعدد الأعمدة.

الجدول (1): عينة من قاعدة البيانات.

	target	radius_mean	texture_mean	...	concave points_worst	symmetry_worst	fractal_dimension_worst
0	M	17.99	10.38	...	0.2654	0.4601	0.11890
1	M	20.57	17.77	...	0.1860	0.2750	0.08902
2	M	19.69	21.25	...	0.2430	0.3613	0.08758
3	M	11.42	20.38	...	0.2575	0.6638	0.17300
4	M	20.29	14.34	...	0.1625	0.2364	0.07678

[5 rows x 31 columns]

يحتوي العمود (target) على قيمتين: خبيثة وحميدة. يمكن بناء نماذج التعلم الآلي على بيانات تتكون من أرقام فقط، لذلك سنستبدل Malignant بالرقم 1 و Benign بالرقم 0. ويمكن استخدام أي رقمين ولكن 0 و 1 هما الأكثر استخدامًا لأغراض التصنيف. بمجرد استبداله، فإن الكود سيكون بمثابة فحص لنحصل على عمود ناتج من الأرقام 1 و 0 والجدول (2) يظهر عينة من قاعدة البيانات بعد القيام بعملية المعالجة لعمود (target).

الجدول (2): عينة من قاعدة البيانات بعد عملية المعالجة.

	target	radius_mean	texture_mean	...	concave points_worst	symmetry_worst	fractal_dimension_worst
0	1	17.99	10.38	...	0.2654	0.4601	0.11890
1	1	20.57	17.77	...	0.1860	0.2750	0.08902
2	1	19.69	21.25	...	0.2430	0.3613	0.08758
3	1	11.42	20.38	...	0.2575	0.6638	0.17300
4	1	20.29	14.34	...	0.1625	0.2364	0.07678

[5 rows x 31 columns]

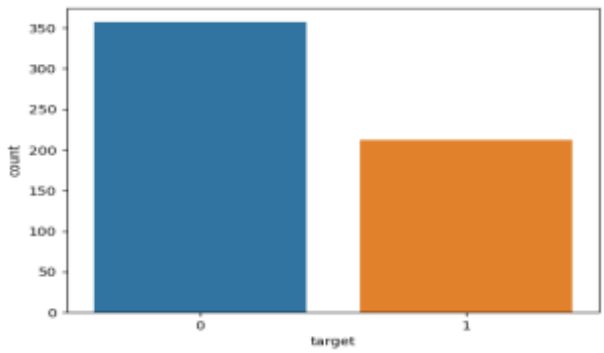
كما قمنا بعملية فحص لجميع البيانات الموجودة ضمن قاعدة البيانات وتأكدنا من عدم وجود عينة فارغة والجدول (3) يبين ذلك.

الجدول (3): عدد العينات في كل ميزة.

0	target	569 non-null	int64
1	radius_mean	569 non-null	float64
2	texture_mean	569 non-null	float64
3	perimeter_mean	569 non-null	float64
4	area_mean	569 non-null	float64
5	smoothness_mean	569 non-null	float64
6	compactness_mean	569 non-null	float64
7	concavity_mean	569 non-null	float64
8	concave points_mean	569 non-null	float64
9	symmetry_mean	569 non-null	float64
10	fractal_dimension_mean	569 non-null	float64
11	radius_se	569 non-null	float64
12	texture_se	569 non-null	float64
13	perimeter_se	569 non-null	float64
14	area_se	569 non-null	float64
15	smoothness_se	569 non-null	float64
16	compactness_se	569 non-null	float64
17	concavity_se	569 non-null	float64
18	concave points_se	569 non-null	float64
19	symmetry_se	569 non-null	float64
20	fractal_dimension_se	569 non-null	float64
21	radius_worst	569 non-null	float64
22	texture_worst	569 non-null	float64
23	perimeter_worst	569 non-null	float64
24	area_worst	569 non-null	float64
25	smoothness_worst	569 non-null	float64
26	compactness_worst	569 non-null	float64
27	concavity_worst	569 non-null	float64
28	concave points_worst	569 non-null	float64
29	symmetry_worst	569 non-null	float64
30	fractal_dimension_worst	569 non-null	float64

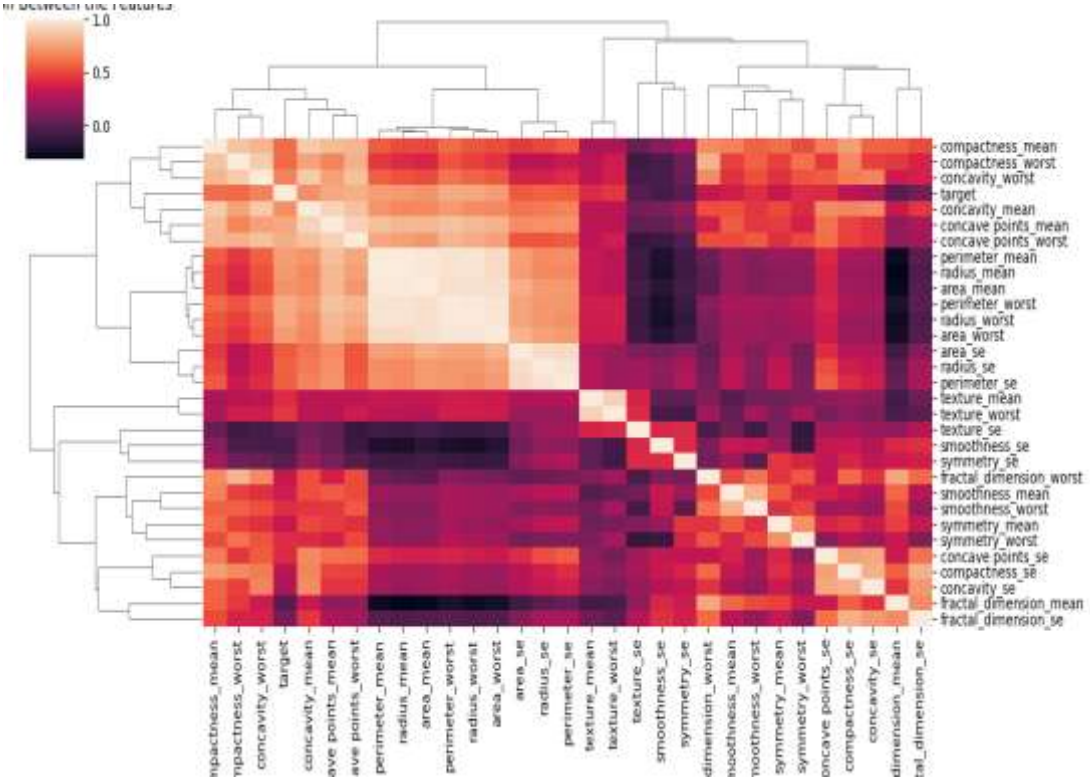
dtypes: float64(30), int64(1)

قمنا بعملية رسم الهيستوغرام للبيانات الموجودة ضمن عمود الـ (target) لمعرفة عدد النساء المصابة بسرطان الثدي الحميد وعدد النساء المصابة بسرطان الثدي الخبيث ومن الشكل (3) نلاحظ وجود 357 حالة مصابة بسرطان الثدي الحميد و 212 حالة مصابة بسرطان الثدي الخبيث.



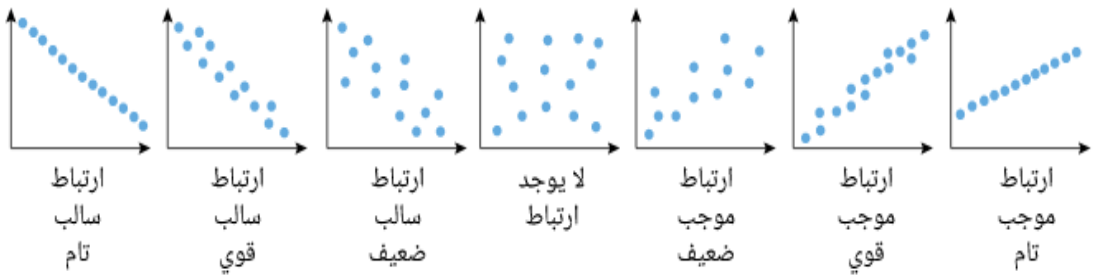
الشكل (3): عدد النساء المصابة بسرطان الثدي الحميد والخبيث.

مصفوفة الارتباط (Correlation Matrix) هي مجرد جدول يحتوي على معاملات الارتباط لميزات مختلفة .توضح المصفوفة كيف ترتبط جميع أزواج القيم الممكنة في الجدول ببعضها البعض. وبالتالي تمكننا من تلخيص مجموعة بيانات كبيرة وإيجاد وإظهار الأنماط في البيانات ويتراوح معامل الارتباط من −1 إلى +1، حيث (−1) تعني ارتباطا سلبيا مثاليا، (+1) تعني ارتباطا إيجابيا مثاليا، و (0) تعني عدم وجود ارتباط بين المتغيرات. والشكل (4) يوضح درجة الارتباط بين السمات الموجودة ضمن قاعدة البيانات لدينا.

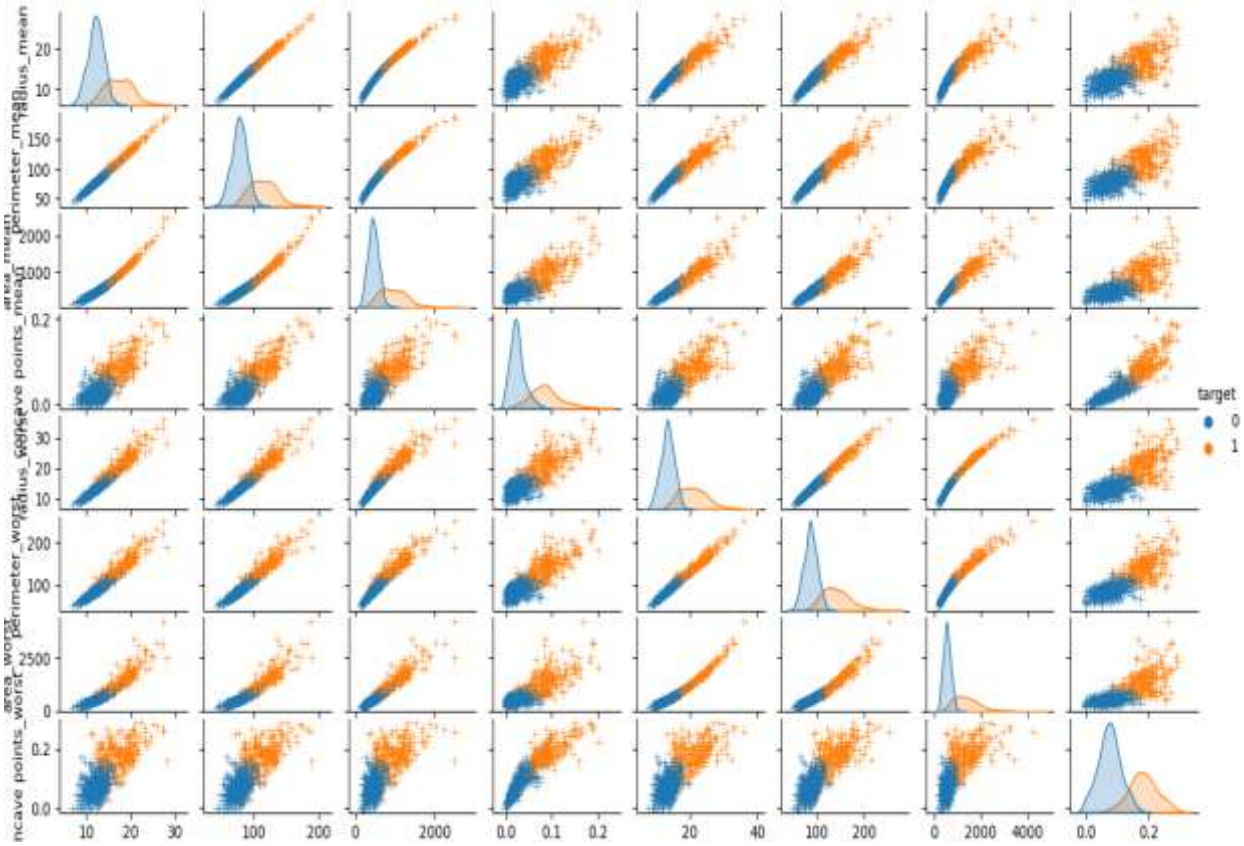


الشكل (4): درجة الارتباط بين السمات

لتوضيح الشكل (4) قمنا برسم مخطط Paris Plot، الذي يوفر أداةً ممتازة لاستكشاف البيانات، إذ يُمكّننا من معاينة العلاقات بين عدة أزواج من السمات، حيث يمكننا تحديد نوع العلاقة بين السمات من خلال معامل بيرسون والشكل (5) يوضح بعض قيم معامل بيرسون، والشكل (6) يبين العلاقات بين السمات لقاعدة البيانات المستخدمة.



الشكل (5): بعض قيم معامل بيرسون

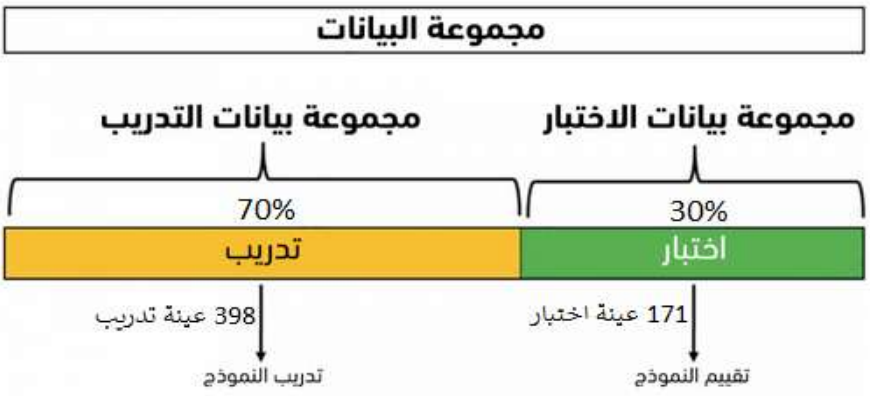


الشكل (6): العلاقة بين السمات

من الشكل (6) نلاحظ وجود ارتباط موجب قوي بين السمتين (radius mean و area mean) في حين أنه يوجد ارتباط موجب ضعيف بين السمتين (radius mean و fractal_dimension_worst).

تقسيم البيانات:

قسمنا مجموعة البيانات التي لدينا إلى قسمين (70% تدريب – 30% اختبار) ليمنحنا فكرة عن كيفية عمل الخوارزمية في مرحلة الاختبار وبهذه الطريقة يتم اختبار الخوارزمية المطبقة على بيانات لم تتدرب عليها مسبقاً والشكل (7) يوضح عملية التقسيم.



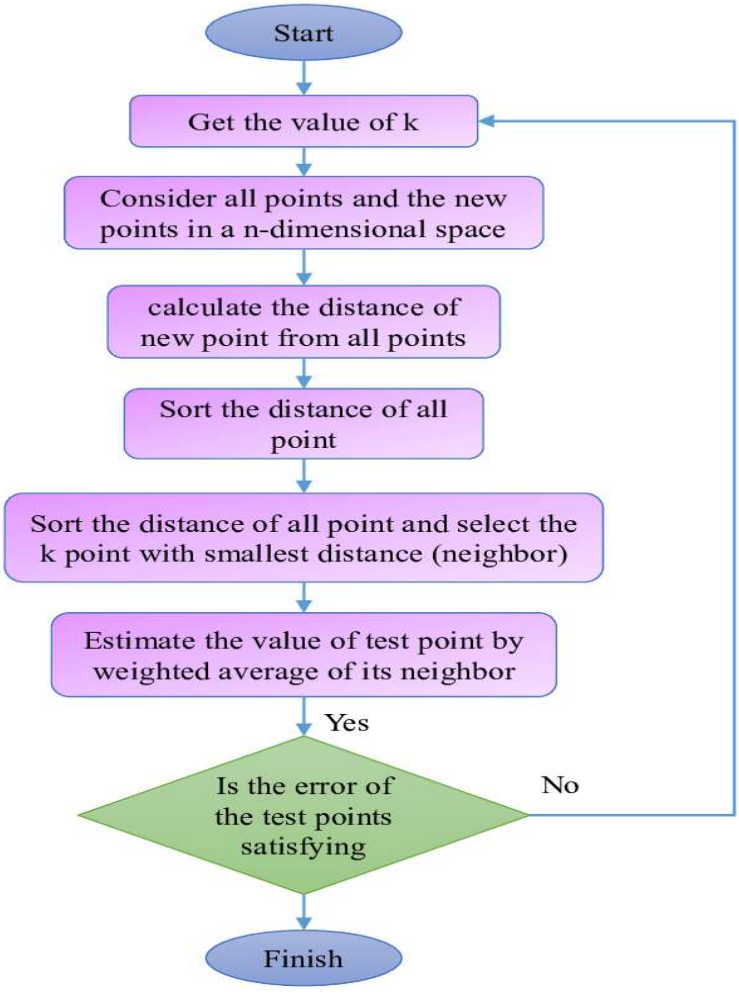
الشكل (7): تقسيم البيانات.

تدريب واختبار نموذج تعلم الآلة:

تم إجراء عملية تدريب لنموذج تعلم الألة بعد خطوة المعالجة المسبقة وإعادة ترتيب مجموعة البيانات والخطوات الأخرى التي تم توضيحها، حيث تم تقسيم التجارب جزئين، كل جزء يحتوي على تجربة واحدة، تم تقييم نتائج النموذج بعد كل تجربة باستخدام جميع مقاييس التقييم، ابتداء من التجربة الأولى والتي تتضمن تدريب النموذج على البيانات دون إجراء أي خطوات تحسينية، وانتهاء بالتجربة الثانية والتي تحتوي على النتائج النهائية التي تم اعتمادها بعد إجراء مجموعة من الخطوات التحسينية.

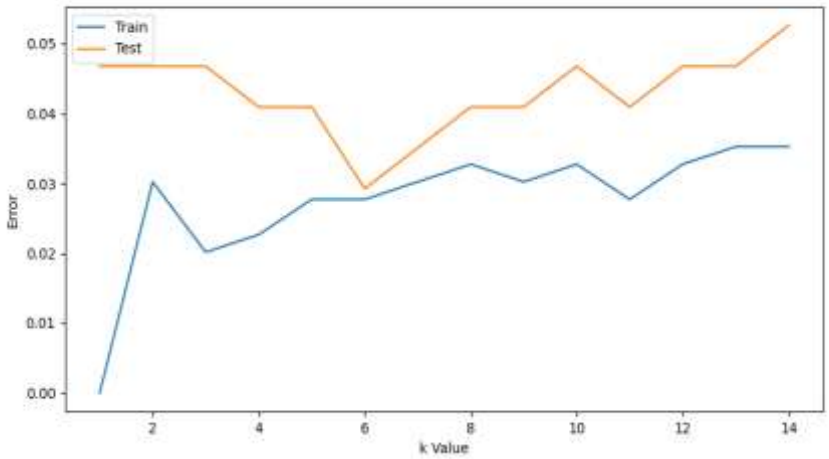
التجربة الأولى:

تم استخدام خوارزمية الجار الأقرب (Knn) في عملية التصنيف كونه واحداً من أبسط خوارزميات التعلم الخاضع للإشراف. تعتمد هذه الخوارزمية على افتراض أن الأشياء المتشابهة قريبة من بعضها البعض. بالمقارنة مع خوارزميات التصنيف الأخرى، والشكل (8) يوضح آلية عمل الخوارزمية.



الشكل (8): آلية عمل الخوارزمية.

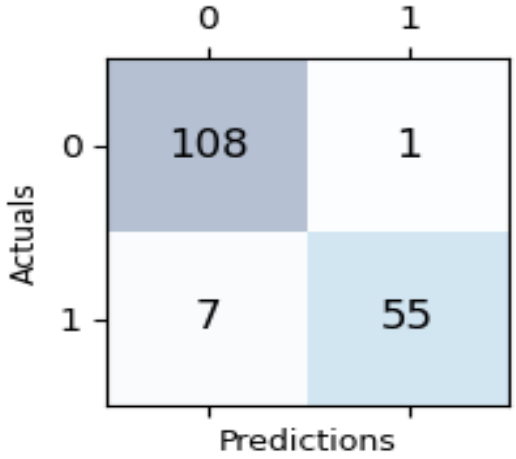
قبل البدء بعملية التدريب قمنا برسم قيمة لـ K ومنحني معدل الخطأ لتحديد أفضل قيمة لها والشكل (9) يوضح ذلك.



الشكل (9): منحني الخطأ لقيم مختلفة لـ K لبيانات التدريب والاختبار

من الشكل (9) نلاحظ عند قيم K المنخفضة، هناك فرط في البيانات/ تباين كبير. لذلك فإن خطأ الاختبار مرتفع وخطأ التدريب منخفض. عند k=1 في بيانات التدريب، يكون الخطأ دائمًا صفراً، لأن أقرب الجيران لتلك النقطة، تلك النقطة نفسها. لذلك، على الرغم من أن خطأ التدريب منخفض، فإن خطأ الاختبار مرتفع عند قيم K المنخفضة. وهذا ما يسمى overfitting كلما قمنا بزيادة قيمة K، يتم تقليل خطأ الاختبار ولكن بعد قيمة K معينة، يتم إدخال التحيز/ نقص الملاءمة ويرتفع خطأ الاختبار. لذلك يمكننا القول في البداية أن خطأ بيانات الاختبار مرتفع بسبب التباين ثم ينخفض ويستقر ومع زيادة أخرى في قيمة K، فإنه يزداد مرة أخرى بسبب التحيز. من منحنى الخطأ يمكننا اختيار k=6 لتنفيذ خوارزمية KNN الخاصة ببياناتنا.

بعد تدريب النموذج وبالاعتماد على مقاييس التقييم المستخدمة قمنا برسم مصفوفة الارتباك لكل مجموعة بيانات والشكل (10) يبين مصفوفة الارتباك.



الشكل (10): يبين مصفوفة الارتباك عند تطبيق مصنف Knn.

تشير مصفوفة الارتباك إلى أن النموذج تمكن من تصنيف 108 عينات حميدة بشكل صحيح من بين 109 عينات، مع خطأ واحد تم تصنيفه كعينة خبيثة. كما أنه من بين 62 عينة خبيثة استطاع النموذج التنبؤ بـ 55 عينة بشكل صحيح وأخطأ بـ7 عينات تنبأ بهم على أنهم حالة حميدة.

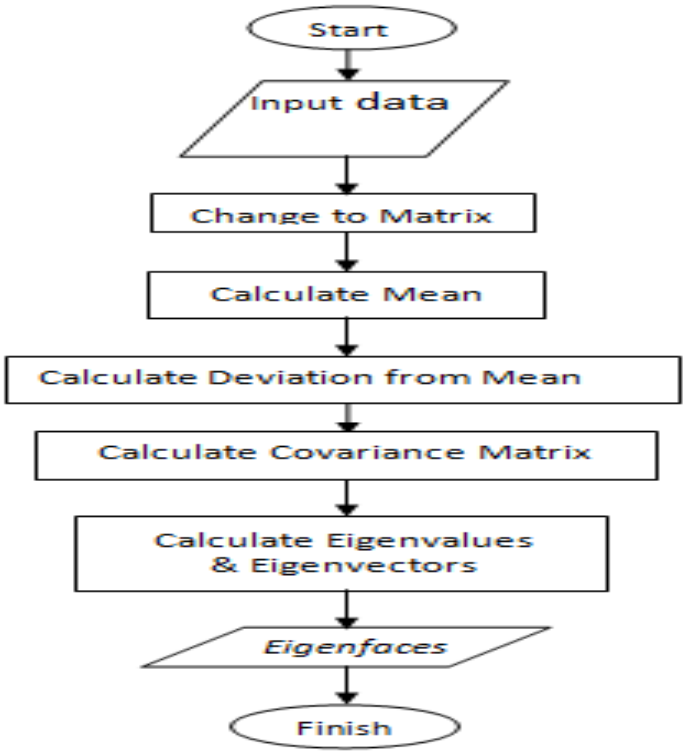
ويوضح الجدول (4) النتائج التي يتم من خلالها قياس أداء النموذج بواسطة المقاييس الموضحة والتي تم حساب قيمها عن طريق القيم الخاصة بمصفوفة الارتباك توضح النتائج أن نسبة دقة النموذج 95%.

الجدول (4): نتائج أداء النموذج بعد التدريب

	precision	recall	f1-score	support
0	0.94	0.99	0.96	109
1	0.98	0.89	0.93	62
accuracy			0.95	171
macro avg	0.96	0.94	0.95	171
weighted avg	0.95	0.95	0.95	171

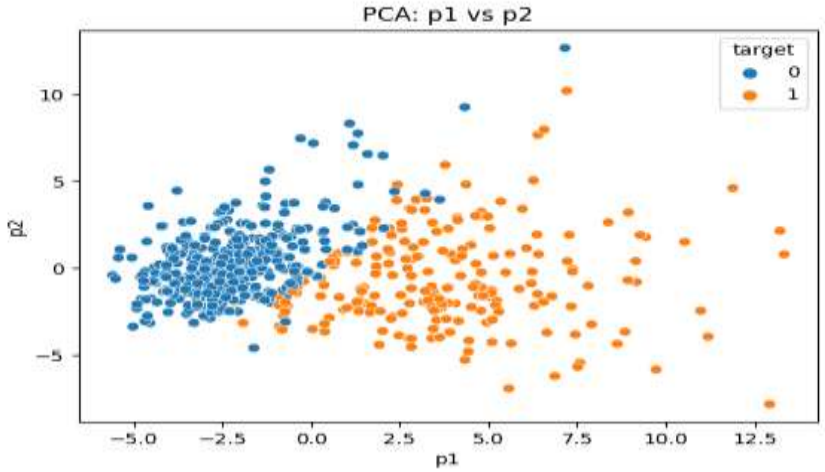
التجربة الثانية:

قمنا باستخدام pca حيث تقوم الخوارزمية بجمع البيانات ومحاولة إيجاد انماط فيما بينها، ثم جمع كل واحده في Clusterخاصة بها بجانب ما يشابهها والشكل (11) يبين آلية عمل الخوارزمية.



الشكل (11): آلية عمل خوارزمية PCA

لدينا في قاعدة البيانات 30 سمة سنعمل على استخراج سمتين جدد من السمات الموجودة في قاعدة البيانات ومن ثم سنقوم بتطبيق خوارزمية Knn على السمتين الجديدتين والشكل (12) يوضح أشعة السمات الجديدة (P1,P2) والجدول (5) بعض القيم من السمات الجديدة.

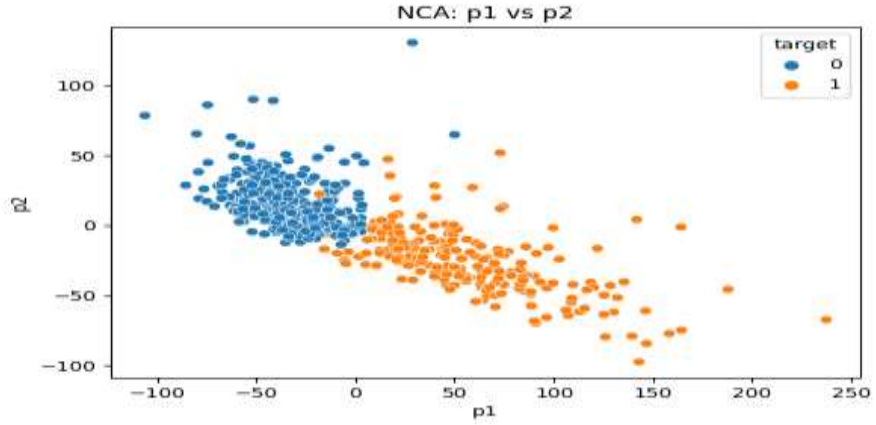


الشكل (12): السمات الجديدة.

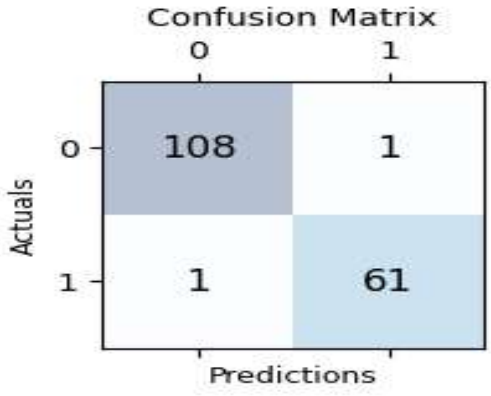
الجدول (5): قيم السمات الجديدة.

	p1	p2	target
0	9.433854	1.789756	1
1	2.482282	-3.868116	1
2	5.865425	-1.184200	1
3	7.205733	10.197309	1
4	4.062830	-2.024530	1

بعد تدريب النموذج وبالاعتماد على مقاييس التقييم المستخدمة والشكل (13) يبين توزع البيانات بعد تطبيق خوارزمية Knn على الأشعة الجديدة ومن ثم قمنا برسم مصفوفة الارتباك لكل مجموعة بيانات والشكل (14) يبين مصفوفة الارتباك.



الشكل (13): توزع البيانات بعد تطبيق خوارزمية Knn.



الشكل (14): مصفوفة الارتباك.

نلاحظ من مصفوفة الارتباك أنه من بين 109 عينة حميدة استطاع النموذج أت يتنبأ لـ 108 حالات بشكل صحيح وأخطأ في حالة واحدة تنبأ بها على أنها عينة خبيثة. كما أنه من بين 62 عينة خبيثة استطاع النموذج التنبؤ بـ 61 عينة بشكل صحيح وأخطأ بـ1 عينات تنبأ بهم على أنهم حالة حميدة.

ويوضح الجدول (6) النتائج التي يتم من خلالها قياس أداء النموذج بواسطة المقاييس الموضحة والتي تم حساب قيمها عن طريق القيم الخاصة بمصفوفة الارتباك توضح النتائج أن نسبة دقة النموذج 99%.

الجدول (6): نتائج أداء النموذج بعد التدريب على البيانات الجديدة

	precision	recall	f1-score	support
0	0.99	0.99	0.99	109
1	0.98	0.98	0.98	62
accuracy			0.99	171
macro avg	0.99	0.99	0.99	171
weighted avg	0.99	0.99	0.99	171

ومن خلال المقارنة بين النتائج التي تم الحصول عليها من التجربة الأولى والمتمثلة في الجدول (4) ونتائج الجدول (6) التي تحصلنا عليها من التجربة الثانية، نلاحظ أن قيم Precision التي توضح عدد العينات التي صنفها النموذج بطريقة صحيحة بالنسبة لكل فئة، أن القيم في الفئة الإيجابية أصبحت أعلى من القيم في الفئة السلبية بعد تطبيق PCA على عكس النتائج السابقة التي كانت فيها القيم السلبية أعلى، و توضح قيم Recall التي تشير إلى عدد العينات في الفئة التي صنفها النموذج بطريقة صحيحة نسبة إلى عدد العينات التي صنفها النموذج أنها تنتمي لهذه الفئة أنه عند تطبيق PCA أصبحت القيم الفئة الايجابية أقل من الفئة السلبية على عكس النتائج السابقة التي كانت فيها قيم الفئة الإيجابية أعلى، ونستنتج من ذلك أنه عند تطبيق PCA مع خوارزمية الـ Knn أصبح النموذج يعطي نتائج أفضل بالنسبة للفئة الإيجابية نسبة لجميع مجموعات البيانات.

مقارنة النتائج مع الدراسات المرجعية:

1. الدراسة الحالية:

- **الخوارزمية المستخدمة**.PCA–KNN :
- **الدقة المحققة**.99% :
- **المنهجية** :استخدام خوارزمية PCA لتقليل الأبعاد وخوارزمية KNN لتحسين التصنيف، مما أدى إلى دقة عالية في تمييز الأورام الحميدة والخبيثة.
- **مجموعة البيانات** :بيانات من جامعة.Wisconsin
- **التفسير** :تحقيق الدقة العالية يعزى إلى فعالية تقليل الأبعاد (PCA) في تقليل الضوضاء وتعزيز أهمية السمات المؤثرة، بجانب أداء KNN الجيد في التصنيف.

2. دراسة:Hiba Asri (2016)

- **الخوارزميات المستخدمة**.Decision Tree, SVM, Naive Bayes, KNN :
- **الدقة المحققة:**
  - أعلى دقة: 97.13% باستخدام.SVM
- **المنهجية** :مقارنة أداء خوارزميات متعددة على قاعدة بيانات.Wisconsin
- **التفسير** :خوارزمية SVM أظهرت دقة جيدة بسبب قدرتها على الفصل بين الفئات باستخدام الدعم الشعاعي، لكنها لم تستفد من تقنيات تقليل الأبعاد مثل.PCA

3. دراسة:Puneet Yadav (2018)

- **الخوارزميات المستخدمة** Decision Tree :و.SVM
- **الدقة المحققة:**
  - Decision Tree: 90%–94%.
  - SVM: 94.5%–97%.
- **المنهجية** :استخدام شجرة القرار وخوارزمية SVM لتصنيف بيانات سرطان الثدي.
- **التفسير** :على الرغم من الأداء الجيد، إلا أن شجرة القرار أقل كفاءة مقارنة بـ SVM بسبب حساسيتها للتداخلات البسيطة في البيانات. لم تستخدم الدراسة أي تقنيات لتحسين السمات.(Feature Selection)

4. دراسة:Akshya Yadav (2019)

- **الخوارزميات المستخدمة** SVM :مع.PCA
- **الدقة المحققة**.96.5% :
- **المنهجية** :دمج خوارزمية SVM مع PCA لتحسين التصنيف.
- **التفسير** :استخدام PCA ساهم في تحسين دقةSVM ، لكنه لم يصل إلى أداء الدراسة الحالية التي استفادت من الجمع بين PCA وخوارزمية.KNN

المقارنة النهائية: الجدول (7): مقارنة مع الدراسات المرجعية

التفسير الرئيسي وراء الأداء	الدقة المحققة	الخوارزميات المستخدمة	الدراسة
الجمع بين PCA و KNNعزز دقة التصنيف بفضل تقليل الأبعاد والتصنيف الفعال.	**99%**	PCA–KNN	**الدراسة الحالية**
SVMفعال في التصنيف لكنه لم يستخدم تقنيات تحسين السمات.	97.13%	SVM	**Hiba Asri (2016)**
الأداء أقل بسبب الاعتماد على خوارزميات غير معززة بتحليل السمات.	94.5%–97%	Decision Tree, SVM	**Puneet Yadav (2018)**
استخدام PCA ساعد على تحسين الأداء لكنه بقي أقل من.PCA–KNN	96.5%	SVMمعPCA	**Akshya Yadav (2019)**

نلاحظ من الجدول (7) أن النموذج المستخدم أعطى أفضل دقة من النماذج المستخدمة في الدراسات السابقة وذلك باستخدام أداة مختلفة وحديثة التي لم يتم استخدامها في الدراسات السابقة.

الاستنتاجات والتوصيات:

تمحورت الدراسة حول استخدام خوارزمية PCA–KNN لتصنيف بيانات سرطان الثدي، مستندة إلى مجموعة بيانات موثوقة من جامعة .Wisconsin أظهرت النتائج التجريبية قدرة النموذج المقترح على تحقيق دقة تصنيف بلغت 99%، متفوقًا على النماذج الأخرى المستعرضة في الدراسات المرجعية. تشير النتائج إلى أن تطبيق تحليل المكونات الرئيسية (PCA)قد ساهم بشكل فعال في تحسين الأداء من خلال تقليل عدد السمات وتعزيز الدقة، خاصة في تصنيف الأورام الخبيثة. يؤكد هذا التفوق أن استخدام خوارزمية PCA–KNN يمثل خطوة واعدة نحو تحسين أنظمة التشخيص الطبي وتقليل الأخطاء البشرية. كما يعكس البحث إمكانية تقليل الإجراءات الجراحية غير الضرورية وزيادة موثوقية التشخيص، مما يعزز جودة الرعاية الصحية للمرضى.

الاستنتاجات:

- يشير هذا البحث إلى أهمية توظيف تقنيات تعلم الآلة، لا سيما خوارزمية **PCA–KNN**، في مجال التشخيص الطبي. من خلال تطبيق الخوارزمية على بيانات سرطان الثدي من جامعةWisconsin ، تم تحقيق دقة تصنيف بلغت 99%، مما يعزز جدوى هذه المنهجية في تحسين التشخيص المبكر وتقليل الأخطاء البشرية.
- تؤكد النتائج أن استخدام تحليل المكونات الرئيسية (PCA) يساعد في تحسين الأداء من خلال تقليل أبعاد البيانات والتركيز على السمات الأكثر تأثيرًا. يعكس هذا البحث تفوق خوارزمية PCA–KNN على النماذج السابقة من حيث الدقة والفعالية، ما يجعلها خيارًا واعدًا للتطبيقات الطبية المستقبلية.

- يبرز البحث أيضًا الدور الحاسم لتقنيات تعلم الآلة في تطوير أنظمة تشخيص أكثر ذكاءً واستدامة. ومن خلال استخدام هذه النماذج، يمكن تقليل الحاجة إلى التدخلات الجراحية غير الضرورية وتحسين موثوقية التشخيص، مما يؤدي إلى تحسين جودة الرعاية الصحية للمرضى وتقليل الأعباء المالية على أنظمة الرعاية الصحية.
- يشير البحث كذلك إلى إمكانيات واسعة لتطوير هذه التقنيات، مثل دمجها مع خوارزميات تعلم عميق واختبارها في بيئات سريرية حقيقية، مما يفتح آفاقًا جديدة في مجال التشخيص الطبي الذكي. وبذلك، يُساهم البحث في تحقيق تقدم مهم نحو جعل الذكاء الاصطناعي جزءًا لا يتجزأ من مستقبل الطب الحديث.

التوصيات:

- **دمج تقنيات تعلم الآلة:**
  - يوصى بدمج خوارزمية PCA–KNN مع تقنيات تعلم الآلة الأخرى، مثل الشبكات العصبية أو خوارزميات التعلم العميق، لاستكشاف إمكانيات تحسين الأداء التشخيصي.
- **اختبار قاعدة بيانات أكبر وأكثر تنوعًا:**
  - توسيع نطاق التجارب باستخدام مجموعات بيانات أكبر ومتنوعة من بيئات ومناطق جغرافية مختلفة، للتأكد من شمولية النتائج وقابليتها للتعميم.
- **التطبيق السريري:**
  - اختبار النموذج في بيئة سريرية واقعية لتقييم أدائه على بيانات حقيقية واستخدامه كأداة مساعدة للأطباء في تشخيص سرطان الثدي.
- **أتمتة المنظومات الطبية:**
  - توسيع تطبيق النموذج ليكون جزءًا من منظومة شاملة للرعاية الصحية تشمل الكشف المبكر عن الأمراض الأخرى، مما يعزز كفاءة التشخيص الطبي.
- **استخدام خوارزميات متقدمة:**
  - استكشاف إمكانيات خوارزميات التعلم العميق مثل الشبكات العصبية التلافيفية (CNN) لمعالجة صور الأنسجة مباشرة، مما قد يؤدي إلى تحسين النتائج مقارنة بالبيانات الرقمية المستخلصة.
- **التعاون بين التخصصات:**
  - تعزيز التعاون بين علماء البيانات والمختصين في المجال الطبي لتطوير نماذج مخصصة تأخذ في الاعتبار الجوانب الطبية والمعايير السريرية.

References:

1. Bagchi A, Pramanik P, Sarkar R. A Multi-Stage Approach to Breast Cancer Classification Using Histopathology Images. Diagnostics. 2023;13(126):1–23. doi:10.3390/diagnostics13010126.
2. Asri H, Mousannif H, Al Moatassime H, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science. 2016;83:1064–1069.
3. Yadav P, Yadav D, Yadav R. Diagnosis of Breast Cancer using Decision Tree Models and SVM. International Research Journal of Engineering and Technology (IRJET). 2018;5(3):March-2018.

4. Yadav A, Jain A, Jain R. Breast Cancer Prediction using SVM with PCA Feature Selection Method. International Journal of Science Research in Computer Science Engineering and Information Technology. 2019;5(2). ISSN: 2456-3307.
5. Wolberg WH, Street WN, Mangasarian OL. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository. 1992.
6. Brudfors M. Generative Models for Preprocessing of Hospital Brain Scans [Doctoral dissertation]. University College London; 2020.
7. Charte F, Rivera AJ, Del Jesus MJ. Multilabel classification: problem analysis, metrics and techniques. Springer International Publishing; 2016.
8. Aous M, Ghada S. Designing a Multiclassification Convolutional Neural Networks Model for the Diagnosis of Lung Cancer and Covid-19. Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series. 2022;44(6):185–200.
9. Dina MI, Nada ME, Amany MS. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. Elsevier. 2021;15(3):1–13.4o