Breast Cancer Classification Based on Analysis of Patients' DNA Arrays

Dr. Sobhi Al Shikha*

(Received 8 / 9 / 2024. Accepted 16 / 10 / 2024)

□ ABSTRACT □

Cancer, especially breast cancer, is one of the most common causes of death worldwide according to the World Health Organization. For this reason, extensive research efforts have been made in the field of accurate and early cancer diagnosis in order to increase the odds of cure. Microarray technology has proven effective among the tools available for cancer diagnosis. Microarray technology analyzes the expression level of thousands of genes simultaneously. Although the sheer number of features or genes in microarray data may appear useful, many of these features are irrelevant or redundant leading to deterioration in classification accuracy. To overcome this challenge, feature selection techniques are a mandatory preprocessing step before the classification process.

Methods and Materials: Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest algorithms are used along with the most common statistical method (logistic regression). Building prediction models using a large data set the nature of the categories, performance evaluation criteria are defined in advance.

Results: After applying different algorithms to the selected data, the results showed that the decision tree (DT) achieved an accuracy of 89.1% thanks to its ability to deal with multi-dimensional data and reduce the overlap between features. The performance of the neural network (ANN) was improved to 90.2% when applying reinforcement learning techniques (Boosting), while deep neural networks (DNN) achieved the highest accuracy of 92.5%.

Conclusion: A comparative study of multiple breast cancer predictive models using a large data set and three classification methods gave us insight into the relative predictive power of different data mining. After analyzing the data, we came to this conclusion: This model can be easily integrated into electronic health records (EHR) systems within hospitals and clinics To improve early diagnosis of breast cancer. Medical teams can be trained to use the model effectively through customized training programs, reducing human error and speeding up the diagnostic process.

Keywords: Breast tumor classification, DNA arrays, neural networks (ANN), support vector machines (SVM), decision trees (DT).

* **Assistant Professor, Faculty Member, College of Information Engineering, Al-Shahbaa Private University, Aleppo, Syria. sobhialshikha@gmail.com**

تصنيف سرطان الثدي اعتماداً على تحليل مصفوفات الحمض النووي (DNA) للمرضى

د. صبحي الشيخة*

(تاريخ الإيداع 8 / 9 / 2024. قُبِل للنشر في 16/ 10 / 2024)

□ ملخّص □

تعتبر أمراض السرطان، وخاصة سرطان الثدي، أحد أكثر أسباب الوفاة شيوعًا في جميع أنحاء العالم وفقًا لمنظمة الصحة العالمية. ولهذا السبب، تم بذل جهود بحثية واسعة النطاق في مجال التشخيص الدقيق والمبكر للسرطان من أجل زيادة احتمالات الشفاء. وقد أثبتت تقنية المصفوفات الدقيقة فعاليتها من بين الأدوات المتاحة لتشخيص السرطان. حيث تقوم تقنية المصفوفات الدقيقة بتحليل مستوى التعبير لآلاف الجينات في وقت واحد. على الرغم من أن العدد الهائل من الميزات أو الجينات في بيانات المصفوفة الدقيقة قد يبدو مفيدًا، إلا أن العديد من هذه الميزات غير ذات صلة أو زائدة عن الحاجة مما يؤدي إلى تدهور دقة التصنيف. للتغلب على هذا التحدي، تم تطبيق تقنيات **تحليل المكونات الرئيسية (PCA) والتصفية الإحصائية** لاختيار الميزات الأكثر أهمية فقط، مما ساهم في تقليل تعقيد النماذج وتحسين دقة التصنيف بنسبة تصل إلى 5%. أظهرت النتائج أن تقليل عدد الميزات ساعد بشكل كبير على تحسين أداء **شجرة القرار (DT)** مقارنة بالخوارزميات الأخرى.

الطرق والمواد: الشبكة العصبية الاصطناعية (ANN)، آلة ناقلات الدعم (SVM)، أشجار القرار (DT)، تُستخدم خوارزميات الغابة العشوائية جنبًا إلى جنب مع الطريقة الإحصائية الأكثر شيوعًا (الانحدار اللوجستي). بناء نماذج التنبؤ باستخدام مجموعة بيانات كبيرة طبيعة الفئات، يتم تحديد معايير تقييم الأداء مسبقًا

النتائج:

بعد تطبيق خوارزميات مختلفة على البيانات المختارة، أظهرت النتائج أن شجرة القرار (DT) حققت دقة 89.1٪ بفضل قدرتها على التعامل مع البيانات متعددة الأبعاد وتقليل التداخل بين الميزات، تم تحسين أداء الشبكة العصبية (ANN) إلى 90.2% عند تطبيق تقنيات **التعلم المعزز (Boosting)**، بينما حققت **الشبكات العصبية العميقة (DNN)** أعلى دقة بنسبة 92.5%.

الاستنتاج: دراسة مقارنة لنماذج تنبؤية متعددة لسرطان الثدي باستخدام مجموعة كبيرة من البيانات وثلاث طرائق للتصنيف أعطتنا نظرة ثاقبة على القدرة النسبية على التنبؤ باستخراج البيانات المختلفة. وبعد تحليل البيانات، توصلنا إلى هذا الاستنتاج: يمكن دمج هذا النموذج بسهولة في أنظمة **السجلات الصحية الالكترونية (EHR)** داخل المستشفيات والعيادات لتحسين التشخيص المبكر لسرطان الثدي. يمكن تدريب الفرق الطبية على استخدام النموذج بشكل فعال من خلال برامج تدريبية مخصصة، مما يقلل من نسبة الأخطاء البشرية ويسرع عملية التشخيص.

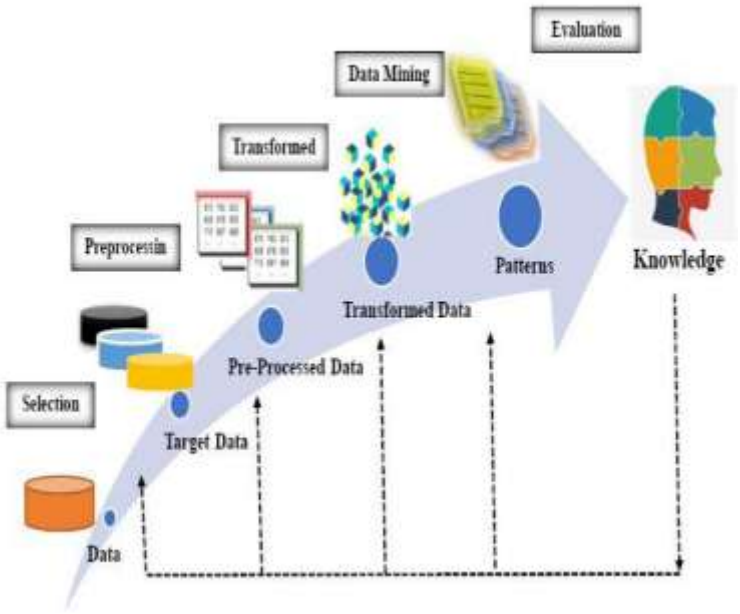
الكلمات المفتاحية: تصنيف أورام الثدي، مصفوفات (سلاسل) الحمض النووي (DNA)، الشبكات العصبونية (ANN)، آلات الدعم الموجه (SVM)، شجرة القرار (DT).

* **مدرس، عضو هيئة تدريسية، كلية الهندسة المعلوماتية، جامعة الشهباء الخاصة، حلب، سورية.sobhialshikha@gmail.com**

مقدمة:

يعتبر تصنيف سرطان الثدي باستخدام بيانات مصفوفات الحمض النووي (DNA microarrays) أحد التحسينات الرئيسية في مجال التشخيص الطبي الحديث. تتضمن هذه البيانات عدداً هائلاً من الميزات الجينية التي قد تكون غير ضرورية أو زائدة، مما يؤدي الى انخفاض دقة التصنيف وهو ما يعرف ب "لعنة الأبعاد". للتغلب على هذه المشكلة، تستخدم تقنيات اختيار الميزات لتحسين جودة البيانات وتحقيق نتائج أكثر دقة عند تطبيق نماذج تعلم الالة.

يعد التنقيب عن البيانات ((Knowledge Discovery in Databases (KDD) خطوة حيوية في اكتشاف المعرفة في عملية قواعد البيانات والتي هي عبارة عن عملية الغربلة ضمن كميات كبيرة جدًا من البيانات للحصول على معلومات مفيدة وله دور مهم في التنبؤ بالأمراض [2]، [4]. بعض من تقنيات استخراج البيانات واسعة النطاق هي الارتباط والتصنيف والتجميع والتنبؤ وتسلسل الأنماط [4]. في هذا البحث، سوف نقوم بتطبيق استخراج البيانات والتعلم الآلي لإنشاء أداة آلية تتنبأ بسرطان الثدي تعتمد على عوامل تتعلق بالمريض ونموذج التنبؤ. تم بناء النموذج من مجموعتين لبيانات أشخاص لا يعانون من السرطان والمرضى الذين يعانون من السرطان. يهدف الطب الدقيق إلى علاج كل مريض بشكل مختلف على الرغم من أنهم يشتركون في أعراض سريرية مماثلة. المراحل الخمس الرئيسية لعملية KDD يمكن تسلسلها كما يلي: الاختيار، المعالجة المسبقة، التحويل، استخراج البيانات، التفسير/التقييم، كما هو موضح في الشكل (1).



الشكل 1: مراحل عمل خوارزمية KDD

في مرحلة اختيار البيانات: البيانات ذات الصلة بالمشكلة هي التي تعتمد ويتم استرجاعها من مجموعات البيانات.

المعالجة المسبقة: تتضمن تحويل البيانات الأولية إلى بيانات مفهومة الشكل. ومن ثم تنظيف البيانات ومعالجتها مسبقًا عن طريق تحديد استراتيجيات للتعامل مع الحقول المفقودة وتعديل البيانات حسب المتطلبات.

تحويل البيانات: هي المرحلة التي يتم فيها تحويل البيانات المعينة إلى النماذج المناسبة لإجرائية استخراج البيانات.

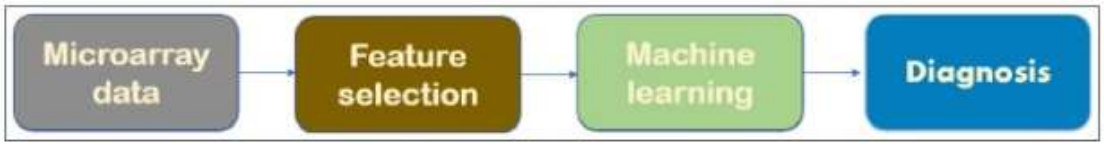
استخراج البيانات: هو الخطوة الأساسية التي يتم فيها تطبيق التقنيات الذكية للحصول على معلومات أو أنماط مفيدة من البيانات. مع تقييم الأنماط، هناك أنماط مثيرة للاهتمام تمثل المعرفة يتم تحديدها بدقة على أساس تدابير معينة.

تمثيل المعرفة: يتم تحديدها بدقة على أساس قياسات معينة. تمثيل المعرفة هو المرحلة الأخيرة التي يتم تنفيذها حيث يتم تقديم المعرفة بصريا للمستخدم.

في هذه الخطوة، تُستخدم تقنيات التصور لمساعدة المستخدمين على فهم وتفسير النتائج المقدمة من النماذج التي تم إنشاؤها في الخطوة السابقة

منهجية تصنيف السرطان:

يتم تحليل بيانات مصفوفة الحمض النووي الدقيقة من خلال الخطوات التالية. أولاً، تتم معالجة البيانات مسبقًا باستخدام تقنيات اختيار الميزات لإزالة الميزات المزعجة والمتكررة والحصول على الميزات المفيدة فقط. ثم يتم استخدام المجموعة الفرعية الناتجة لتدريب نموذج التعلم لتشخيص أنواع السرطان الفرعية كما هو موضح في الشكل 2



الشكل 2: مراحل تصنيف السرطان

طرائق البحث ومواده:

جمع البيانات يمكن أن تكون مجموعات البيانات متاحة للعامة، مثل مجموعة البيانات عبر الإنترنت أو تعاون المنظمة مع فريق بحث. البيانات متاحة للعامة تم تحميلها من المركز الوطني لمعلومات التكنولوجيا الحيوية (NCBI) التي تعمل على تطوير العلوم والصحة من خلال توفير الوصول للمعلومات الطبية الحيوية والجينومية.

نقوم بجمع عينات من تسلسل الجين BRCA2 الحميد والخبيثة. العدد الإجمالي لمجموعة البيانات هو 154,801 عينة [27]، [28].

معالجة البيانات هي طريقة لاستخراج البيانات تقوم بتحويل البيانات الخام إلى نموذج يمكن فهمه. بيانات العالم الحقيقي غالبًا ما تكون غير مكتملة أو متعارضة، وتفتقر إلى بعض السلوكيات أو الاتجاهات وبما في ذلك العديد من الأخطاء. المعالجة المسبقة للبيانات هي طريقة مجربة وصحيحة لحل مثل هذه المشاكل. حيث تتضمن خطوات المعالجة المسبقة:

1: إضافة المكتبات

2: إضافة مجموعة البيانات

3: التحقق من القيم المفقودة

4: تحديد قيمة الصف

5: تقسيم مجموعة البيانات إلى مجموعتين تدريب واختبار.

هذه هي الطريقة التي نستورد بها المكتبات في بايثون باستخدام ملف استيراد الكلمات الرئيسية، المكتبة الأكثر شعبية التي يستخدمها أي عالم البيانات.

بناء النموذج الخطوة الأولى من المعالجة المسبقة هي تحويل ملف تنسيق SAR إلى تنسيق ملف Fasta، ثم تنسيق ملف Fastq الخطوة الثانية هي وضع العلامات على البيانات. يستخدم التعلم الخاضع للإشراف لخوارزميات التصنيف تصنيف البيانات يدويًا. للمساعدة في هذا، تم استيراد عمود جديد "الفئة" إلى البيانات التي تم جمعها للإشارة إلى تصنيف التسلسل، إيجابي أو سلبي. يعتبر التسلسل الجيني الإيجابي (الفئة 1) ورمًا خبيثًا لسرطان الثدي. التسلسل الجيني السلبي (الفئة 0) تسلسل حميد لسرطان الثدي. الخطوة الثالثة: جمع مجموعة البيانات في إطار البيانات (التسلسل، الفئة).

المعالجة المسبقة للبيانات:

في خطوة المعالجة المسبقة، تم التعامل مع القيم المفقودة باستخدام تقنيات تعويض البيانات مثل **إدخال المتوسط أو الحد الأقصى للقيم المشابهة** بالإضافة الى ذلك، تم تطبيق **تحليل المكونات الرئيسية (PCA)** لاختيار الميزات الأكثر تأثيراً وإزالة الميزات غير الضرورية، مما ساهم في تحسين دقة التصنيف. ثم تحميل البيانات من **NCBI** وتوزيع العينات بالتساوي باستخدام تقنية **K – fold cross – validation** لضمان التوازن بين مجموعات التدريب والاختبار.

الخطوة الرابعة حقيبة الكلمات (شكل 4)

الخطوة الخامسة تقسيم البيانات: تم توزيع العينات باستخدام تقنية **K – fold cross – validation** لضمان توازن تام بين العينات المختلفة وتقليل الانحياز في التوزيع. تم تقسيم البيانات الى خمس مجموعات (fold)، مما أتاح تحسين دقة النماذج بشكل أكبر وتحقيق نتائج أكثر موثوقية عند الاختبار.

المصنف: مع زيادة استخراج البيانات، تلعب أشجار القرار دورًا جديًا في استخراج البيانات وتحليلها. تعلم شجرة القرار يحتاج إلى استخدام مجموعة من بيانات التدريب لعمل شجرة القرار الذي يجمع بيانات التدريب بدقة. إذا كانت عملية التعلم قيد التشغيل، فإن شجرة القرار ستجمع العناصر الجديدة بشكل مثالي كبيانات الإدخال. نموذج شجرة القرار مع طريقة اختيار شجرة القرار

precision = 0.890 , accuracy = 0.891

recall = 0.891, and F1 = 0.890

النموذج الأفضل أداءً تم بناؤه لأنه يتمتع بأعلى دقة.

نماذج التنبؤ:

استخدمنا ثلاثة أنواع مختلفة من نماذج تصنيف الشبكة العصبية، آلة ناقلات الدعم، أشجار القرار، تم اختيار هذه النماذج لإدراجها في بحثنا نظرًا لشعبيتها الشديدة مؤخرًا بالإضافة إلى دقتها الأفضل

آلات دعم المتجهات (SVM):

آلة ناقل الدعم (SVM) هي خوارزمية تدريب لتعلم قواعد التصنيف والانحدار للبيانات.

المعادلة لSVM $w*x + b = 0$: حيث w هو وزن الناقل ، x وهو

متجه الإدخال، وb هو مصطلح التحيز.

الشبكات العصبية الاصطناعية:

الشبكة العصبية الاصطناعية (ANN)، بشكل عام، هي تقنية تعتمد على شبكة من الخلايا العصبية الاصطناعية المصممة للقيام بمهام معينة مثل التصنيف والتجميع والتعرف على الانماط، يتم إعطاء الأنماط للشبكة عن طريق "الإدخال". "الطبقة" المرتبطة هي واحدة أو أكثر من "الطبقات المخفية" التي يتم فيها الانتهاء من معالجة الأنماط. الطبقات الخفية، ثم الارتباط بـ "طبقة الإخراج" حيث يتم تقديم الإجابة كإخراج. يمكن أن يكون لـ ANN استخدامات مختلفة، والتي تشمل التصنيف: حيث يمكن تدريب الشبكة العصبية على تصنيف شيء معين نمط أو مجموعة بيانات في فئة محددة مسبقًا. يستخدم التنبؤ شبكات التغذية الأمامية يمكن تدريب الشبكة العصبية على إنتاج المخرجات المتوقعة من مدخلات معينة [32].

شجرة القرار:

مع نمو تقنيات استخراج البيانات، أصبحت شجرة القرار خوارزمية التصنيف القوية والتي أصبحت أكثر شعبية مع ازدياد استخراج البيانات في المنطقة وتلعب دورا رئيسيا في استخراج البيانات والتحليل. احتياجات تعلم شجرة القرار باستخدام مجموعة من بيانات التدريب لتصميم شجرة القرار. إذا كانت عملية التعلم نشطة، ستقوم الشجرة بتصنيف بيانات الإدخال الجديدة بدقة. تختلف أشجار القرار عبر العديد من الأبعاد، مثل قواعد التوقف والفصل والمعايير وحالة الفرع ونوع الشجرة النهائي وتشغيل الفرع [33].

معايير تقييم الأداء

في هذه الدراسة، استخدمنا مقاييس الأداء والتي هي: الدقة، الاستدعاء، الضبط، و F1، مصفوفات الارتباك والمعروفة باسم مصفوفات الخطأ في مجال التعلم الآلي والتي غالبا ما تستخدم لتقييم وتصور أداء النماذج والخوارزميات.

جدول المعلومات والتفاصيل من التصنيف الفعلي (مصنف من قبل الإنسان) و التصنيفات المتوقعة (المصنفة حسب النموذج). حجم المصفوفة يتم تحديده من خلال عدد الاصناف، حيث يمثل سطرها التصنيف الحقيقي، ويمثل عمودها التصنيف المتوقع.

الدقة، الاستدعاء، قياس F1، تم قياسها لتحديد جودة المصنف المختبر، كما هو مبين في الجدول 1، والذي يتضمن المزيد من التفاصيل لتوضيح القياسات التي تم تطبيقها من خلال مصفوفة الارتباك [33]، [35].

- TP (إيجابي حقيقي): عدد الحالات التي تدل على أن النموذج تنبأ بشكل صحيح للصنف p وفعليا هو كان الصنف p
- FP (إيجابي كاذب): عدد الحالات التي تم فحصها بالنموذج, والتي تنبأ عنها بشكل غير صحيح للفئة p وفي الواقع هي تنتمي بالفعل إلى الصنف n

جدول 1: مصفوفة الارتياب

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative(FN)
	Negative	False Positive (FP)	True Negative (TN)

TN (سلبي حقيقي): عدد الحالات التي تم توقع النموذج لها بشكل صحيح للصنف n وفي الواقع ينتمي إلى الفئة n.

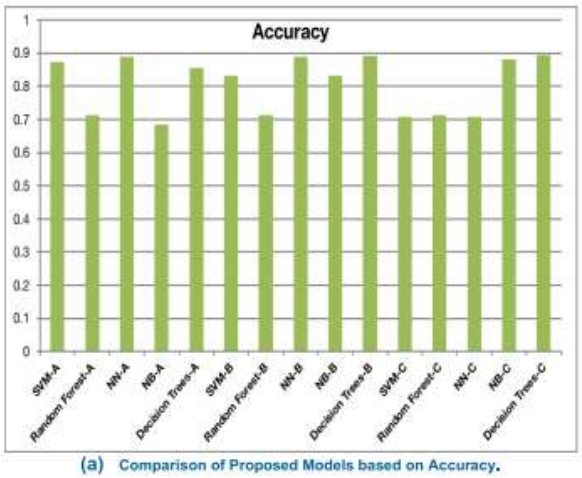
• FN (سلبي كاذب): عدد حالات النموذج تتبأ بشكل غير صحيح للفئة n وينتمي إلى الفئة p.

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN). \quad (2)$$

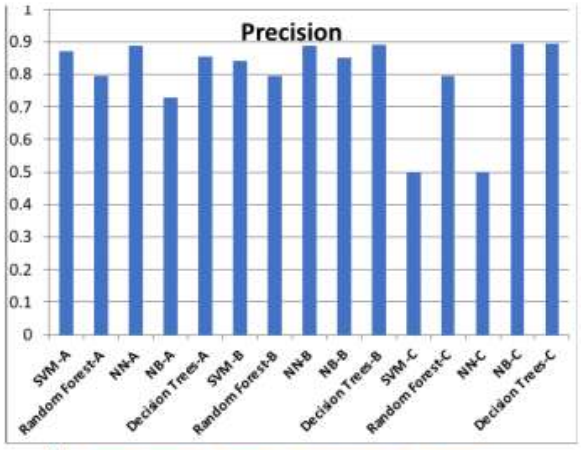
$$\text{Precision} = TP/(TP + FP). \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

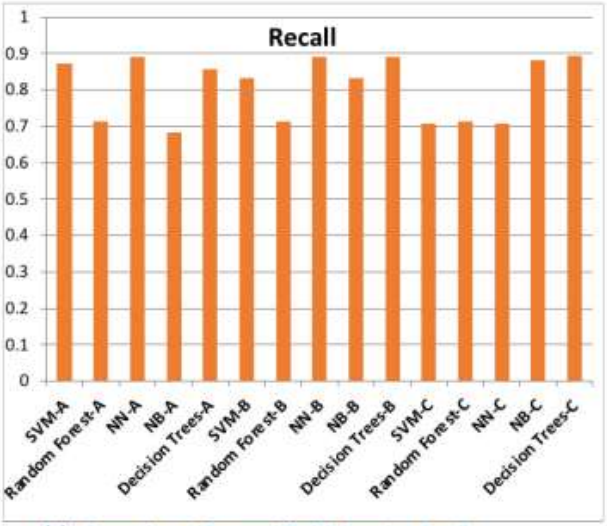
$$F1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (5)$$



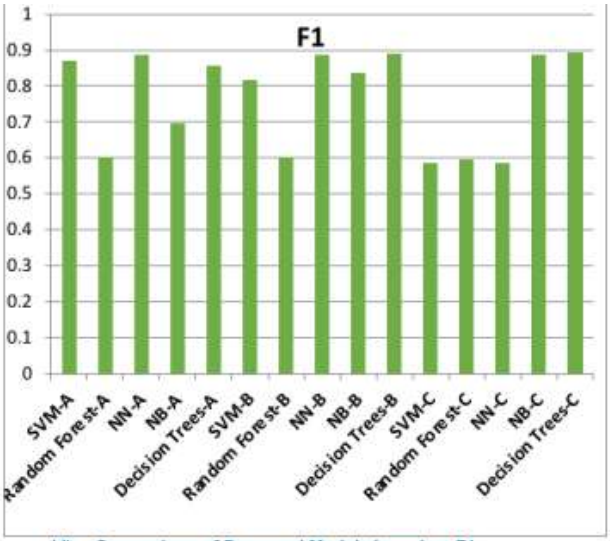
(a) Comparison of Proposed Models based on Accuracy.



(b) Comparison of Proposed Models based on Precision.



(c) Comparison of Proposed Models based on Recall.



(d) Comparison of Proposed Models based on F1.

الشكل 4: مقارنة تقنيات التصنيف المستخدمة من حيث الدقة والاستعادة والضبط

الجدول 2: المقارنة بين المصنفات المدروسة

	Algorithm	Accuracy	Precisio	Recall	F1
Experiment A	SVM	0.873	0.871	0.873	0.871
	Random Forest	0.712	0.796	0.712	0.602
	NN	0.889	0.888	0.889	0.888
	NB	0.684	0.730	0.684	0.696
	Decision Trees	0.856	0.856	0.856	0.856
Experiment B	SVM	0.832	0.841	0.832	0.816
	Random Forest	0.712	0.796	0.712	0.602
	NN	0.889	0.888	0.889	0.888
	NB	0.833	0.850	0.833	0.837
	Decision Trees	0.891	0.890	0.891	0.890
Experiment C	SVM	0.707	0.500	0.707	0.586
	Random Forest	0.712	0.795	0.712	0.596
	NN	0.707	0.500	0.707	0.586
	NB	0.882	0.893	0.882	0.885
	Decision Trees	0.895	0.895	0.895	0.895

الاستنتاجات والتوصيات:

تم تقييم الخوارزميات المستخدمة في هذا البحث باستخدام مقاييس الدقة المذكورة سابقا.

لقد قمنا بالكثير من التجارب باستخدام أساليب وتقنيات مختلفة، كما هو موضح في مصنف الجدول 2. حيث تم استخدام ما مجموعه 154801 تسلسل اختبار في التجارب، وكان مطلوبًا تقسيمها إلى زوج من المجموعات.

الجدول 2 والشكل 4 تمثل الدقة والاستدعاء والضبط ونتائج قيم F1 التي تم تحقيقها أثناء تطبيق طرق مختارة لتصنيف المجموعة من خلال ثلاث تجارب للاختبار (التجربة A، التجربة B، والتجربة C).

في التجربة A: حقق مصنف NN أعلى دقة بنسبة 88.9%.

في التجربة B: أعلى دقة 89.1%، حيث تم تحقيقها باستخدام نموذج شجرة القرار

في التجربة C: بلغت نسبة أداء نموذج شجرة القرار 89.5%. لكننا وجدنا أن النماذج الخوارزمية تتفاعل أحيانًا بشكل مختلف مع التقنيات التي استخدمناها. على سبيل المثال، زاد أداء نموذج شجرة القرار بنسبة 89.5% عندما استخدمنا التجربة C

References:

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. • Jemal, and F. Bray, ‘‘Global cancer statistics 2023: Advances in early detection and AI applications in healthcare’’, CA Cancer J. Clin., vol. 74, no. 1, pp. 110–145, March 2023. doi: 10.3322/caac.23010.

[2] S. Gupta, D. Kumar, and A. Sharma. (2011). Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. Accessed: Aug. 31, 2023. [Online]. Available: https://www.semanticscholar.org/ paper/DATA-MINING-CLASSIFICATION-TECHNIQUES-APPLIEDFOR-Gupta Kumar/2bdad1956bb8f0e5c0583210f8f6a8359ce36219

[3] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao, ‘‘Big data and analytics in healthcare: Introduction to the special section,’’ Inf. Syst. Frontiers, vol. 18, no. 2, pp. 233–235, Apr. 2016, doi: 10.1007/s10796-016-9641-2.

[4] W. Raghupathi and V. Raghupathi, ‘‘Big data analytics in healthcare: Promise and potential,’’ Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1–10, Dec. 2014, doi: 10.1186/2047-2501-2-3.

[5] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, ‘‘–Omic and electronic health record big data analytics for precision medicine,’’ IEEE Trans. Biomed. Eng., vol. 64, no. 2, pp. 263–273, Feb. 2017, doi: 10.1109/TBME.2016.2573285.

[6] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, ‘‘Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012,’’ Int. J. Cancer, vol. 136, no. 5, pp. 359–386, Mar. 2015, doi: 10.1002/ijc.29210.

[7] Y. Wang, L. Kung, W. Y. C. Wang, and C. G. Cegielski, ‘‘An integrated big data analytics-enabled transformation model: Application to health care,’’ Inf. Manag., vol. 55, no. 1, pp. 64–79, Jan. 2018, doi: 10.1016/j.im.2017.04.001.

[8] A. Sahu, K. K. Ajeeshkumar, M. N. Peerzada, M. K. Yadav, and S. Verma, ‘‘Nature-inspired computing in breast cancer research: Overview, perspective, and challenges of the state-of-the-art techniques,’’ in NatureInspired Intelligent Computing Techniques in Bioinformatics, K. Raza, Ed. Singapore: Springer, 2023, pp. 45–62, doi: 10.1007/978-981-19-6379-7_3.

[9] tern recognition model to distinguish cancerous DNA sequences via signal processing methods,’’ Soft Comput., vol. 24, no. 21, pp. 16315–16334, Nov. 2020, doi: 10.1007/s00500-020-04942-4.

[10] M. H. Dunham, Data Mining Introductory and Advanced Topics. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[11] S. Marsland, Machine Learning: An Algorithmic Perspective, 1st ed. Boca Raton, FL, USA: Chapman, 2009.

[12] S. Chumbar. KDD Process in Data Science: A Beginner’s Guide. Accessed: Nov. 16, 2023. [Online]. Available: https://medium. com/@shawn.chumbar/kdd-process-in-data-science-a-beginners-guide426d1f0fc062

[13] V. K. Tiwari, A. Heesacker, O. Riera-Lizarazu, H. Gunn, S. Wang, Y. Wang, Y. Q. Gu, E. Paux, D. Koo, A. Kumar, M. Luo, G. Lazo, R. Zemetra, E. Akhunov, B. Friebe, J. Poland, B. S. Gill, S. Kianian, and J. M. Leonard, ‘‘A whole-genome, radiation hybrid mapping resource of hexaploid wheat,’’ Plant J., vol. 86, no. 2, pp. 195–207, Apr. 2016, doi: 10.1111/tpj.13153.

[14] S. Song, D. Tian, C. Li, B. Tang, L. Dong, J. Xiao, Y. Bao, W. Zhao, H. He, and Z. Zhang, ‘‘Genome variation map: A data repository of genome variations in BIG data center,’’ Nucleic Acids Res., vol. 46, no. 1, pp. 944–949, Oct. 2017, doi: 10.1093/nar/gkx986.
[15] Genetic Map. Accessed: Nov. 16, 2023. [Online]. Available: https://www.genome.gov/genetics-glossary/Genetic-Map
[16] K.-C. Wong, Big Data Analytics in Genomics, 1st ed. New York, NY, USA: Springer, 2016.
[17] S.-B. Cho and H.-H. Won, ‘‘Machine learning in DNA microarray analysis for cancer classification,’’ in Proc. 1st Asia–Pacific Bioinf. Conf., Jan. 2003, pp. 189–198.
[18] J. O. Afolayan, M. O. Adebiyi, M. O. Arowolo, C. Chakraborty, and A. A. Adebiyi, ‘‘Breast cancer detection using particle swarm optimization and decision tree machine learning technique,’’ in Intelligent Healthcare: Infrastructure, Algorithms and Management, C. Chakraborty and M. R. Khosravi, Eds. Singapore: Springer, 2022, pp. 61–83, doi: 10.1007/978-981-16-8150-9_4.
[19] M. O. Adebiyi, M. O. Arowolo, M. D. Mshelia, and O. O. Olugbara, ‘‘A linear discriminant analysis and classification model for breast cancer diagnosis,’’ Appl. Sci., vol. 12, no. 22, p. 11455, Nov. 2022, doi: 10.3390/app122211455.
[20] M. O. Adebiyi, J. O. Afolayan, M. O. Arowolo, A. K. Tyagi, and A. A. Adebiyi, ‘‘Breast cancer detection using a PSO-ANN machine learning technique,’’ in Using Multimedia Systems, Tools, and Technologies for Smart Healthcare Services. Hershey, PA, USA: IGI Global, 2023, pp. 96–116, doi: 10.4018/978-1-6684-5741-2.ch007.
[21] Y.-D. Zhang, S. C. Satapathy, D. S. Guttery, J. M. Górriz, and S.-H. Wang, ‘‘Improved breast cancer classification through combining graph convolutional network and convolutional neural network,’’ Inf. Process. Manag., vol. 58, no. 2, Mar. 2021, Art. no. 102439, doi: 10.1016/j.ipm.2020.102439.
[22] Y.-D. Zhang, C. Pan, X. Chen, and F. Wang, ‘‘Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling,’’ J. Comput. Sci., vol. 27, pp. 57–68, Jul. 2018, doi: 10.1016/j.jocs.2018.05.005.
[23] D. Delen, G. Walker, and A. Kadam, ‘‘Predicting breast cancer survivability: A comparison of three data mining methods,’’ Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, Jun. 2005, doi: 10.1016/j.artmed.2004. 07.002.
[24] S.-B. Cho and J. Ryu, ‘‘Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features,’’ Proc. IEEE, vol. 90, no. 11, pp. 1744–1753, Nov. 2002, doi: 10.1109/JPROC.2002. 804682.
[25] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, ‘‘Predicting the clinical status of human breast cancer by using gene expression profiles,’’ Proc. Nat. Acad. Sci. USA, vol. 98, no. 20, pp. 11462–11467, Sep. 2001, doi: 10.1073/pnas.201162998.
[26] K. P. Soh, E. Szczurek, T. Sakoparnig, and N. Beerenwinkel, ‘‘Predicting cancer type from tumour DNA signatures,’’ Genome Med., vol. 9, no. 1, p. 104, Nov. 2017, doi: 10.1186/s13073-017-0493-2. [27] Y.
[27] Mingquan, G. Lingyun, and W. Chunyuan, ‘‘Gene expression data classification based on artificial bee colony and SVM,’’ J. Shandong Univ., vol. 48, no. 3, pp. 10–16, 2018, doi: 10.6040/j.issn.1672-3961.0.2017.405.
[28] H. Hijazi and C. Chan, ‘‘A classification framework applied to cancer gene expression profiles,’’ J. Healthcare Eng., vol. 4, no. 2, pp. 255–284, Jun. 2013, doi: 10.1260/2040-2295.4.2.255.

[29] R. A. Musheer, C. K. Verma, and N. Srivastava, ‘‘Novel machine learning approach for classification of high-dimensional microarray data,’’ Soft Comput., vol. 23, no. 24, pp. 13409–13421, Dec. 2019, doi: 10.1007/s00500-019-03879-7.
[30] N. Mohd Ali, R. Besar, and N. A. A. Aziz, ‘‘A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation,’’ Bull. Electr. Eng. Informat., vol. 12, no. 2, pp. 1047–1054, Apr. 2023, doi: 10.11591/eei.v12i2.4838.
[31] S. Afreen, A. K. Bhurjee, and R. M. Aziz, ‘‘Gene selection with game Shapley Harris hawks optimizer for cancer classification,’’ Chemometric Intell. Lab. Syst., vol. 242, Nov. 2023, Art. no. 104989, doi: 10.1016/j.chemolab.2023.104989.
[32] C. Boniface, C. Deig, C. Halsey, T. Kelley, M. B. Heskett, C. R. Thomas, P. T. Spellman, and N. Nabavizadeh, ‘‘The feasibility of patient-specific circulating tumor DNA monitoring throughout multi-modality therapy for locally advanced esophageal and rectal cancer: A potential biomarker for early detection of subclinical disease,’’ Diagnostics, vol. 11, no. 1, p. 73, Jan. 2021, doi: 10.3390/diagnostics11010073.
[33] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[34] S. Mandt, M. D. Hoffman, and D. M. Blei, ‘‘Stochastic gradient descent as approximate Bayesian inference,’’ 2017, arXiv:1704.04289.
[35] S. A. El Rahman, A. Al-Montasheri, B. Al-Hazmi, H. Al-Dkaan, and M. Al-Shehri, ‘‘Machine learning model for breast cancer prediction,’’ in Proc. Int. Conf. 4th Ind. Revolution (ICFIR), Feb. 2019, pp. 1–8, doi: 10.1109/ICFIR.2019.8894777.
[36] A. K. Mishra and B. Ratha. (2016). Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis. Accessed: Sep. 1, 2023. [Online]. Available: https://www.semanticscholar. org/paper/Study-of-Random-Tree-and-Random-Forest-Data-MiningMishra-Ratha/173c9872946718df7a3e6fe29c3822f2f8631149
[37] S. O. Abdulsalam, M. O. Arowolo, and O. Ruth, ‘‘Stroke disease prediction model using ANOVA with classification algorithms,’’ in Artificial Intelligence in Medical Virology (Medical Virology: From Pathogenesis to Disease Control), J. M. Chatterjee and S. K. Saxena, Eds. Singapore: Springer, 2023, pp. 117–134, doi: 10.1007/978-981-99-0369-6_8.
[38] M. O. Arowolo, R. O. Adesina, M. O. Adebiyi, and A. A. Ad ‘‘A comparative of classification models for predicting of heart diseases,’’ in Proc. Int. Conf. Sci., Eng. Bus. Sustain. Develop. Goals (SEB-SDG), vol. 1, Apr. 2023, pp. 1–8, doi: 10.1109/SEB-SDG57117.2023.10124478.
[39] M. O. Arowolo, H. E. Aigbogun, P. E. Michael, M. O. Adebiyi, and A. K. Tyagi, ‘‘Chapter 12—A predictive model for classifying colorectal cancer using principal component analysis,’’ in Data Science for Genomics, A. K. Tyagi and A. Abraham, Eds. Cambridge, MA. USA: Academic Press, 2023, pp. 205–216, doi: 10.1016/B978-0-323-98352- 5.00004-5.
[40] M. Arowolo, V. Akubor, S. Misra, L. Garg, M. Adebiyi, and J. Awotunde, ‘‘Development of a chi-square approach for classifying ischemic stroke prediction,’’ in Proc. Int. Conf. Inf. Syst. Manag. Sci., 2022, pp. 268–279, doi: 10.1007/978-3-031-13150-9_23.