جامعة حلب



كلية الهندسة المعلوماتية

للعام الدراسي 2024/2025

Classification of Breast Cancer   
using Backpropagation

بإشراف: المهندسة هديل عبد الرحمن

الاسم: محمود بنّان



* **ملخص المشروع:**

يهدف هذا المشروع إلى **تطوير نموذج للتعلم الآلي قادر على التنبؤ بحالة بقاء مرضى سرطان الثدي** (سواء كانوا "أحياء" أو "متوفين")، وذلك باستخدام بيانات سريرية شاملة.

قمت بتصميم وتدريب شبكة عصبونية اصطناعية بهدف التنبؤ **بحالة بقاء مرضى سرطان الثدي على قيد الحياة**.   
أظهر النموذج أداءً ممتازاً على بيانات التدريب، حيث حقق **دقة وصلت إلى 95% مع نسبة خطأ منخفضة جداً بلغت 4%.**   
عند تقييم النموذج على بيانات اختبار جديدة، استطاع **تحقيق دقة جيدة بنسبة 90%، مع نسبة خطأ بلغت 0%1.**

* **حول المشروع:**

لا يزال سرطان الثدي أحد أكثر أنواع السرطانات انتشارًا على مستوى العالم، ويعد الكشف المبكر والتشخيص الدقيق أمرًا بالغ الأهمية للعلاج الفعال وتحسين نتائج المرضى. تلعب مجموعات البيانات مثل مجموعة بيانات "سرطان الثدي" دوراً حيويًا في تمكين الباحثين وعلماء البيانات من تطوير واختبار النماذج التنبؤية للتشخيص والبقاء على قيد الحياة.

* **حول البيانات:**

تم الحصول على هذه المجموعة من بيانات مرضى سرطان الثدي من تحديث نوفمبر 2017 لبرنامج المراقبة والوبائيات والنتائج النهائية (SEER) التابع للمعهد الوطني للسرطان (NCI). يوفر برنامج SEER إحصائيات سرطان قائمة على السكان.

تتضمن مجموعة البيانات هذه المريضات الإناث اللاتي تم تشخيصهن بسرطان الثدي الغازي القنوي والفصيصي (والذي يندرج تحت رموز الأنسجة NOS 8522/3 في سجلات SEER الأولية) خلال الفترة من 2006 إلى 2010.

لضمان جودة البيانات ودقتها، تم استبعاد المرضى الذين لديهم:

* حجم ورم غير معروف.
* عقد ليمفاوية إقليمية غير مفحوصة (LNs).
* عقد ليمفاوية إقليمية إيجابية غير معروفة.
* أشهر بقاء على قيد الحياة أقل من شهر واحد.

بعد هذه المعايير، بلغ العدد النهائي للمرضى المدرجين في مجموعة البيانات 4024 مريضًا.

**معلومات عن السمات:**

1. العمر: عمر المريض عند تشخيص المرض
2. العرق: العرق الذي ينتمي إليه المريض.
3. الحالة الاجتماعية: الحالة الاجتماعية للمريض (أعزب، متزوج، إلخ)
4. مرحلة T: تشير إلى حجم الورم الأولي وامتداده. جزء من نظام TNM لتصنيف السرطان.
5. مرحلة N: تشير إلى ما إذا كان السرطان قد انتشر إلى العقد الليمفاوية القريبة. جزء من نظام TNM.
6. المرحلة السادسة: غالبًا ما يكون تصنيفًا إجماليًا لمرحلة السرطان يجمع بين معلومات T و N و M .
7. التمييز: يشير إلى مدى شبه الخلايا السرطانية بالخلايا الطبيعية، مما يعكس عدوانية السرطان.
8. الدرجة: درجة الورم النسيجية، تشير إلى مدى عدوانية الخلايا السرطانية وسرعة نموها.
9. مرحلة A: قد يشير إلى تصنيف فرعي للمرحلة أو نظام تدريج بديل.

10- حجم الورم: حجم الورم الأولي.

11- حالة الإستروجين: تشير إلى ما إذا كانت الخلايا السرطانية تحتوي على مستقبلات الإستروجين (موجب/سالب).

12- حالة البروجسترون: تشير إلى ما إذا كانت الخلايا السرطانية تحتوي على مستقبلات البروجسترون

(موجب/سالب).

13- العقد الليمفاوية الإقليمية المفحوصة: عدد العقد الليمفاوية القريبة من الورم التي تم فحصها.

14- العقد الليمفاوية الإقليمية الإيجابية: عدد العقد الليمفاوية القريبة التي تحتوي على خلايا سرطانية.

15- أشهر البقاء على قيد الحياة: عدد الأشهر التي عاشها المريض بعد التشخيص.

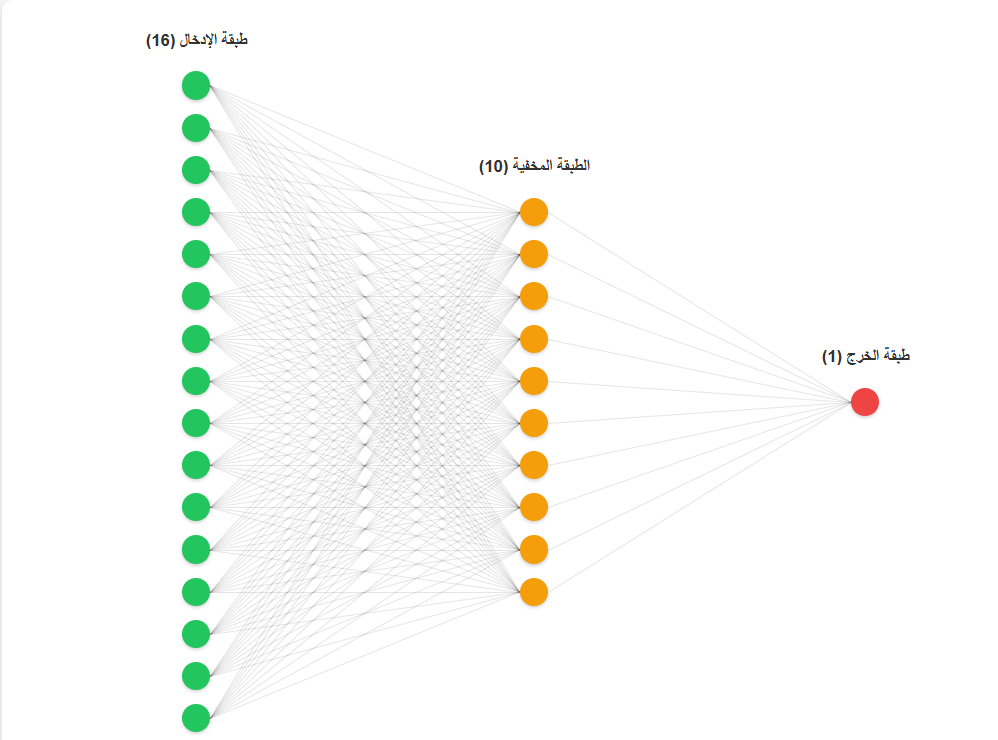
16- الحالة: المتغير المستهدف، يشير إلى حالة بقاء المريض على قيد الحياة (عادةً "حي" أو "متوفى").

* في مرحلة التحضير، تم إجراء عملية معالجة أولية لمجموعة البيانات بهدف ضمان جودتها، حيث لوحظ وجود بعض القيم غير الصحيحة التي استدعت التصحيح. بعد التأكد من سلامة البيانات، تم تقسيمها إلى مجموعتين رئيسيتين:   
  خُصصت نسبة 80% من البيانات لعملية تدريب الشبكة العصبونية،   
  بينما تم حجز نسبة الـ 20% المتبقية كمجموعة بيانات اختبار لتقييم أداء النموذج على بيانات لم يرها مسبقًا.
* **بنية الشبكة:**

تم بناء النموذج باستخدام شبكة عصبونية اصطناعية تعتمد على خوارزمية الانتشار الخلفي (Backpropagation)، وتتألف بنيتها من الطبقات التالية:

* **طبقة الإدخال: (Input Layer)** مكونة من 16 عصبوناً، حيث يستقبل كل عصبون إحدى الميزات (features) الـ 16 من مجموعة البيانات.
* **الطبقة المخفية: (Hidden Layer)** طبقة واحدة تحتوي على 01 عصبوناً، مسؤولة عن استخلاص الأنماط من بيانات الإدخال.
* **طبقة الإخراج: (Output Layer)** تحتوي على عصبون واحد فقط، وظيفته تقديم التصنيف النهائي الثنائي (إما ان المريض متوفى او على قيد الحياة).

وقد تم استخدام دالة التفعيل **Sigmoid** لكل العصبونات الموجودة في الطبقة المخفية وطبقة الإخراج .



* **منهجية العمل والنتائج:**

اتبعت منهجية عمل منظمة بدأت بمرحلة **معالجة البيانات الأولية**، حيث تم تنظيف مجموعة البيانات لضمان جودتها، وتضمنت هذه المرحلة خطوتين أساسيتين: أولا، تصحيح الأخطاء في تنسيق الأرقام، مثل تحويل الفاصلة العشرية من "," إلى "." (على سبيل المثال: 15,5 أصبحت 15.5) ثانيا، تم حذف الصفوف (السجلات) التي تحتوي على حقول فارغة. بعد ذلك، تم **تقسيم البيانات**، حيث خُصصت نسبة **80%** من السجلات لمجموعة **التدريب** ونسبة  **20%** المتبقية لمجموعة **الاختبار**.

للوصول إلى البنية المثلى للشبكة، تم إجراء سلسلة من التجارب المنهجية مع تغيير عدد العصبونات في الطبقة المخفية ومعدل التعلم وكانت أبرز النتائج كالتالي:

* **عند استخدام 5 عصبونات:**

في البداية كان معدل التعلم 0.1 وفي منتصف التدريب تم تعديله الى 0.01

بعد 10000 دورة تدريب تم الوصول الى معدل خطأ 6% ومعدل نجاح 92 % للبيانات التي تدرب عليها وبمعدل

خطأ 8% ومعدل نجاح 89% للبيانات الجديدة.

* **عند استخدام 10 عصبونات:**

في البداية كان معدل التعلم 0.1 وفي منتصف التدريب تم تعديله الى 0.01

بعد 10000 دورة تدريب تم الوصول الى معدل خطأ 5% ومعدل نجاح 93 % للبيانات التي تدرب عليها وبمعدل خطأ 7% ومعدل نجاح 90% للبيانات الجديدة.

* **عند استخدام 20 عصبوناً:**

في البداية كان معدل التعلم 0.1 وفي منتصف التدريب تم تعديله الى 0.01

بعد 10000 دورة تدريب تم الوصول الى معدل خطأ 4% ومعدل نجاح 95 % للبيانات التي تدرب عليها وبمعدل خطأ 11% ومعدل نجاح 87% للبيانات الجديدة.

* **عند استخدام 30 عصبوناً:**

في البداية كان معدل التعلم 0.1 وفي منتصف التدريب تم تعديله الى 0.01

بعد 10000 دورة تدريب تم الوصول الى معدل خطأ 4% ومعدل نجاح 96 % للبيانات التي تدرب عليها وبمعدل خطأ 10% ومعدل نجاح 88% للبيانات الجديدة.

تُظهر هذه المقارنات أثر تغيير هذه البارامترات بشكل مباشر على أداء النموذج، وقد ساعدت هذه العملية في تحديد التوليفة الأفضل من البارامترات التي قادت في النهاية إلى تحقيق أفضل نتيجة ممكنة على بيانات الاختبار، والتي هي:

1. 10 عصبونا في الطبقة المخفية (hidden layer) وعصبون واحدة لطبقة الخرج
2. معدل تعلم0.1 في بداية التعليم 0.01 في منتصف التعلم
3. تابع تفعيل لكلا من الطبقتين hidden, output هو تابع Sigmoid
4. 0,0001دورة تدريب لكامل البيانات (entire dataset)

* رسم بياني للدقة والخطأ اثناء التدريب:

