تحسين التنبؤ بمرض الكلى المزمن باستخدام أقل عدد من السمات

رشا كاظم مسعود[1]

الملخص

أدت التكاليف الباهظة للديلزة في المراحل الأخيرة من مرض الكلى المزمن (chronic kidney disease (CKD)) إلى ضرورة الكشف المبكر عن المرض قبل تقدمه إلى مراحل أخرى، إلا أن التحدي الأكبر هو أن معظم الناس لا يعانون من أي أعراض أو علامات في المراحل المبكرة من المرض، ولا يتم كشف المرض لديهم إلا في المراحل المتقدمة، وعندها تكون السيطرة على المرض صعبة. جذبت فكرة الكشف المبكر عن مرض الكلى المزمن العديد من الأطباء والباحثين لتخفيض معدل الوفيات وإيقاف تقدم المرض في مراحله المبكرة والتقليل من عدد المرضى الخاضعين للديلزة وتكاليف الرعاية الصحية المرافقة، يهدف هذا البحث إلى التنبؤ بمرض الكلى المزمن من خلال أقل عدد ممكن من البارامترات والفحوصات المخبرية. تم في هذا البحث بناء نظام ذكي يعتمد على الشبكات العصبونية الصنعية ((intelligent artificial neural network (ANN)) للتنبؤ بمرض الكلى المزمن واستخدمت قاعدة بيانات لـ 400 عينة، و كل عينة لها 23 سمة تصف حالتها (بعضها وصف سريري ،والآخر قيم فحوصات مخبرية)، أعطت الشبكة العصبونية دقة عالية (99.5%) وتوافق هذا مع الأبحاث السابقة، ثم تم تخفيض عدد السمات باستخدام الخوارزمية الجينية مع الشبكة العصبونية، وذلك بهدف الحصول على السمات الأكثر ارتباطاً بالمرض، فأصبح عدد السمات 3 سمات والتي حافظت على أداء الشبكة العصبونية، للتأكد من أهمية هذه السمات الناتجة و مدى فعاليتها فقد استخدمت في التنبؤ بالمرض باستخدام خوارزمية عنقدة k-means وهي خوارزمية ذات تعلم غير موجه (unsupervised learning) وكانت دقة النتائج (99.5%) مما أكد صحة السمات المستنتجة.

الكلمات المفتاحية: مرض الكلى المزمن (CKD)، الخوارزمية الجينية، الشبكة العصبونية الصنعية (ANN).

[1]أستاذ مساعد في قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق، دمشق، سورية، ص.ب. 86
البريد الإلكتروني: rqies@yahoo.com.

Improving the Prediction of Chronic Kidney Disease Using the Least Number of Features

Rasha Kazem Massoud

Abstract

High costs of dialysis during late stages of chronic kidney disease highlight the importance of CKD early detection. However, most people do not show any major signs or symptoms at early stages, so the disease is generally detected at later stages. CKD's early detection can reduce mortality rate of the disease, control its progress during early stages, and lower the number of dialysis or transplantation patients. This paper aims to predict CKD using the least number of clinical and physiological tests. An intelligent artificial neural network (ANN) system was constructed to predict CKD, a dataset with 400 observations and 23 features was used, the ANN system accuracy is 99.5%, which agrees with the literature studies. Then the number of features was reduced in order to find the most related features from the dataset by applying genetic algorithm to the ANN, the algorithm reduced the number of features to three, while maintaining the ANN performance. In order to validate the importance of the deducted features, a k-means clustering algorithm was used, which is an unsupervised learning algorithm; the three features were able to detect the disease even without supervision, with an accuracy of 99.5%.

Keywords: chronic kidney disease (CKD), genetic algorithm (GA), artificial neural network (ANN).

1- مقدمة

يعدُّ مرض الكلى المزمن (CKD) أحد أهم مشاكل الصحة العامة في العالم، وتقدر نسبة انتشار المرض في دول العالم المتقدم والدول النامية بين 10% و16%، إلا أن التوعية لهذا المرض قليلة جداً (أقل من 20% من النسبة السكانية في معظم البلدان)، وهذا المرض يزيد من النسبة المرضية و نسبة الوفيات. فقد وُجِد في عام 2010م ان نسبة الوفيات المتعلقة بمرض الـCKD ازدادت بنسبة 82.3% في العقدين الأخيرين، فهي تأتي مباشرة بعد مرض الإيدز والسكري، وقد أفضت دراسة أُجريت في الولايات المتحدة الأمريكية بين عامي 2002م و م2016إلى أن تأثير هذا المرض وانتشاره قد خطا بعيداً متجاوزاً الأمراض الأخرى غير المُعْدِيَة، وأصبح عبئاً ثقيلاً على المعنيين بالصحة العامة، وهذا شكل تحدياً للكشف المبكر عن هذا المرض والتدخل الفعال في بدايته قبل تفاقمه، حيث تم صرف 98 مليار دولار في الولايات المتحدة في عام 2015م على الرعاية الصحية لمرضى الفشل الكلوي في الحالة الأخيرة، والحالة ليست أفضل في الصين بسبب العبء المادي الناتج عن انتشار هذا المرض [1].

بينما أظهرت دراسة أجريت في الصين ان نسبة مرض الكلى المزمن الناتج عن مرض السكري أصبحت منذ عام 2010م أكبر بكثير من مرض الكلى المزمن الناتج عن التهاب كبيبات الكلى، وهذا يعني ان مرض السكري يؤدي دوراً مهما في الإصابة بهذا المرض وانتشاره [1].

وقد سعت عدة دول في العالم مثل الصين والولايات المتحدة و كندا وبريطانيا واليابان وأستراليا إلى بناء انظمة لرصد هذا المرض للحصول على أكبر قدر من الإحصاءات و المعلومات عنه لتستطيع السيطرة عليه وإدارته بشكل فعال [2]، وقد بنيت معظم هذه الدراسات بالاعتماد على سجلات المرضى الرقمية، وذلك من أجل تشخيص المرض ،وإدارته، تقييم عوامل الخطورة المرافقة له. وهدفت معظم هذه الدراسات إلى بناء نظام دعم اتخاذ قرار طبي ذكي (Computer Decision Support System (CDSS)) بناء على معطيات سجلات المرضى والإرشادات الطبية [3].

أدت التكاليف الباهظة للديلزة في المراحل الأخيرة من مرض الكلى المزمن إلى ضرورة الكشف المبكر عن المرض قبل تقدمه إلى مراحل أخرى، إلا ان التحدي الأكبر هو أنَّ معظم الناس لا يعانون من أيِّ أعراض أو علامات في المراحل المبكرة من المرض ولا يتم كشف المرض لديهم إلا في المراحل المتقدمة، وعندها تكون السيطرة على المرض صعبة. جذبت فكرة الكشف المبكر عن مرض الـ CKD العديد من الأطباء والباحثين لتخفيض معدل الوفيات و إيقاف تقدم المرض في مراحله المبكرة و التقليل من عدد المرضى الخاضعين للديلزة وتكاليف الرعاية الصحية المرافقة لذلك [4].

جمع الباحثون الفحوصات المخبرية والمعطيات التي تتعلق بمرضى الكلى المزمن وقاموا ببناء نماذج لتقدم المرض إما بالاعتماد على التحليل الإحصائي ،أو على تعلم الآلة-5] [8. بينما قام آخرون بالتنبؤ بهذا المرض في مراحله الأولى حيث استخدم Sarada J. وزملاؤه عام 2018 خوارزميتي C4.5 و J48 المبنيتين على شجرة القرار في التنبؤ بالمرض وتصنيفه، وقد أعطت دقة % 99.5 [9]، أما Elhoseny M. وزملاؤه فقد قدموا في بحثهم عام 2019م نظاماً ذكياً للتنبؤ وتصنيف مرض الـ CKD، حيث اختاروا السمات عن طريق الكثافة الاحتمالية (Density Feature Selection_DFS)، ومن ثم بنيت قواعد شرطية اعتماداً

على خبرة الطبيب في تصنيف المرض، تمت أمثلة هذه القواعد باستخدام خوارزمية مستعمرة النمل (Ant Colony based Optimization (D-ACO) algorithm) وكانت دقة هذا النظام 95%, بينما كانت الحساسية 96% والخصوصية 93.33% [10]. استخدمت Gharibdousti M.S. وزملاؤها عام 2017 م عدة تقنيات تعتمد على تعلم الآلة لتصنيف مرض الـ CKD، والتنبؤ المبكر به كشجرة القرار وآلة متجه الدعم والشبكات العصبونية والارتداد الخطي، وقد أعطت التقنيات الثلاث الأخيرة دقة عالية جداً [11].

اعتمد الباحثون المذكورون آنفاً قاعدة البيانات نفسها والتي تعتمد على 24 سمة يتم الحصول عليها من نتائج الفحوصات المخبرية أو من معطيات المريض، يهدف هذا البحث إلى التقليل من عدد السمات مع المحافظة على أداء المصنف و معرفة أهم السمات التي تساعد في كشف المرض، مما يقلل من تكاليف الفحوصات المخبرية، وذلك عن طريق استخدام الخوارزمية الجينية بشكل هجين مع الشبكات العصبونية في استخلاص السمات.

2-الطرائق المستخدمة:

2-1 قاعدة البيانات و تهيئتها:

تم الحصول على قاعدة البيانات من مخزن UCI لتعلم الآلة [12]، وهذه البيانات لمرضى الكلى المزمن، وهي تتألف من 400 عينة من جنوب الهند تتراوح أعمارهم بين 90-2 عاماً، منهم 250 مريضاً و150 سليماً. تحوي هذه البيانات على 24 سمة (feature) معظمها سريري والقسم الآخر فيزيولوجي، ولذلك بعضها له قيم اسمية والبعض الآخر قيمه رقمية. وجزء من هذه المعطيات مفقود سواء كان رقمياً أم اسمياً. اقترحت الدراسات للتخلص من هذه المشكلة حذف جميع العينات التي تكون فيها المعطيات مفقودة [13]، و هذا يؤدي إلى انخفاض عدد العينات إلى 181 عينة، كما حسب [11] القيمة المتوسطة لكل مزية بالنسبة لجميع العينات ووضعها مكان المزية المفقودة، وهذه الطريقة ليست مثالية في تعويض السمات الناقصة. أما هذا البحث فقد استخدم طريقة مختلفة لمعالجة البيانات تعتمد على مسافة الارتباط بين العينة ذات السمة المفقودة و بقية العينات ذات السمات المكتملة وتم استبدال السمات المفقودة بسمات العينة المكتملة الأقرب (ذات المسافة الأصغر)، حيث تم حساب مسافة الارتباط من المعادلة (1):

$$d_{st} = 1 - \frac{(X_s - \bar{X}_s)(X_t - \bar{X}_t)'}{\sqrt{(X_s - \bar{X}_s)(X_s - \bar{X}_s)'}\sqrt{(X_t - \bar{X}_t)(X_t - \bar{X}_t)'}}$$

$$\bar{X}_s = \frac{1}{n}\sum_j X_{sj} \quad \text{و} \quad \bar{X} = \frac{1}{n}\sum_j X_{tj} \quad \text{حيث}$$

و n هو عدد السمات الموجودة في العينة غير المكتملة و j هو رقم السمة بينما X_{sj} هي السمة رقم j في العينة غير المكتملة و $_{tj}$ هي السمة رقم j في العينة المكتملة.

واختيرت مسافة الارتباط لأنها تقوم بتقييس البيانات فيتساوى تأثير السمات في حساب المسافة مهما اختلفت القيم الرقمية لهذه السمات. بعد ملء الفراغات الموجودة في قاعدة البيانات تم تحويل السمات ذات القيم الاسمية إلى قيم رقمية، فمثلاً المزية ذات القيمة "normal" تم استبدالها بالرقم 1 بينما القيمة "abnormal" تم استبدالها بالقيمة 0 ويبين الجدول1 شرح سمات قاعدة البيانات ورمز كل مزية، تم حذف السمة الرابعة (Rbc) من الجدول لأنها تحوي على

عدد كبير من الفراغات، ومن ثمَّ أصبح عدد السمات الكلي 23 سمة.

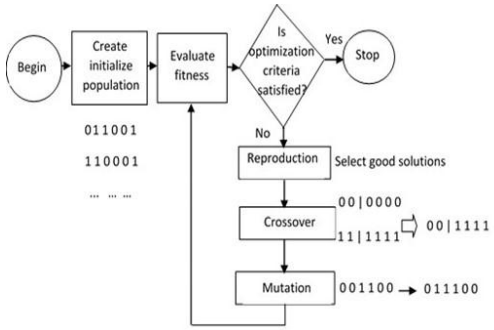
الجدول (1) السمات الموجودة في قاعدة البيانات و رموزها.

Specific Gravity	1	Sg	Pus Cell clumps	13	Pcc
Albumin	2	Al	Age	14	Age
Sugar	3	Su	Blood pressure	15	Bp
Red Blood Cells	4	Rbc	Blood Glucose Random	16	Bgr
Pus Cell	5	Pc	Blood Urea	17	Bu
Bacteria	6	Ba	Serum Creatinine	18	Sc
Hypertension	7	Htn	Sodium	19	Sod
Diabetes Mellitus	8	Dm	Potassium	20	Pot
Coronary Artery Disease	9	Cad	Hemoglobin	21	Hemo
Appetite	10	Appet	Packed Cell Volume	22	Pcv
Pedal Edema	11	Pe	White Blood Cell Count	23	Wc
Anemia	12	Ane	Red Blood Cell Count	24	Rc

2-2 الخوارزمية الجينية:

تعدُّ الخوارزمية الجينية إجرائية أمثلة عودية، حيث في غياب أي معرفة عن نطاق المشكلة تبدأ الخوارزمية الجينية بحثها في تجمع عشوائي من الحلول (population)، لتطبق بعدها مجموعة من العمليات المستوحاة من علوم الأحياء المسماة بالاستنساخ (Reproduction) والتهجين (Crossover) والطفرة (Mutation)، لتعمل على تحديث الحلول وتتكرر هذه العملية من أجل عدد معرف مسبقاً من المرات أو الوصول إلى هدف محدد[14].

بدايةً يُنشأ تجمعٌ أولي من الحلول الممكنة، ويُحول كل عنصر في هذا التجمع إلى مجموعة من السلاسل (الصبغيchromosome) التي تؤثر فيها عوامل الخوارزمية الجينية، وبعدها يُقيّم كل فرد من التجمع باستخدام التابع الهدف الذي تفترضه المشكلة، مما يسبب قدح عملية اختيار أزواج الأفراد الذين سيتم تزاوجهم معاً خلال الاستنساخ، وتُعين قيمة اللياقة المشتقة من قياس الأداء الخام للفرد والتي يعطيها التابع الهدف، تستخدم قيمة اللياقة لتحديد اتجاه الانزياح نحو الأفراد الأكثر لياقة، فيمتلك الأفراد الأكثر لياقة احتمالاً أعلى لاختياره للتزاوج بينما يقل احتمال التزاوج من أجل الأفراد الأقل لياقة ويبين الشكل 1 آلية عمل الخوارزمية الجينية [14].



الشكل 1 آلية عمل الخوارزمية الجينية[14].

تمثيل التجمع والتهيئة:

يمثل الصبغي تركيب البيانات الحاملة لسلسلة من البارامترات والتي تسمى بالمورثة (gene)، وقد تخزن هذه السلسلة على شكل سلسلة بت ثنائي (التمثيل الثنائي Binary Representation) أو على شكل مصفوفة من الأعداد الصحيحة (تمثيل النقطة العائمة أو تمثيل الترميز الحقيقي) الذي يمثل الأعداد ذات الفاصلة العشرية، وتعدُّ المورثة قسم جزئي من الصبغي يُرمز عادة قيمة بارامتر واحد.

تابع الهدف وقيمة اللياقة:

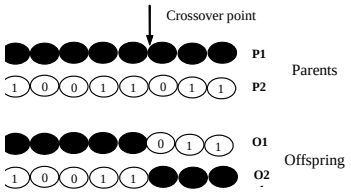
يتصل تابع الهدف مباشرة مع المشكلة أو حيث يراد تطبيق الخوارزمية الجينية، وفي معظم الحالات يكون الهدف من اختيار الخوارزمية الجينية الأمثل أو عملية البحث هو إما تصغير أو تكبير قيمة التابع الهدف، حيث يمتلك الأفراد الأكثر لياقة القيمة العددية الأدنى للتابع الهدف المرافق لهم، ونظراً لاختلاف قيم الهدف باختلاف المشكلة فإن الحصول على توحيد لنطاقات مشكلة مختلفة يتطلب إعادة تقييس قيمة الهدف بالنسبة لقيمة اللياقة [15].

الاختيار:

تبدأ عملية الاختيار حالما يتم تقييم الأفراد وتعيين قيم اللياقة لاختيار الأفراد (يفترض الأصلح، والأكثر لياقة) ليكونوا آباء وأمهات الجيل القادم، ويفرز التجمع من الأفضل إلى الأسوأ، ثم ينسخ كل فرد على عدد المرات الممكنة تبعاً للياقته.

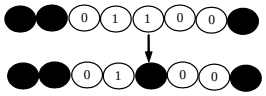
التهجين والطفرة:

يعدُّ التهجين والطفرة عوامل أساسية، فالتهجين هو العملية المسؤولة عن إنتاج صبغيات في الخوارزمية الجينية، فهي تقوم بالتبديل بين جزأين من المعلومات المورثية في صبغيين لتولد صبغيات جديدة، وتُختار نقاط التبديل بشكل عشوائي، فيظهر الشكل 2 أبسط أشكال التهجين، وهو تهجين النقطة الوحيدة، حيث تتبادل عند نقطة التهجين أجزاءُ الوالدين P1, P2 لتنتج ذرية جديدة O1, O2.



الشكل 2 عملية التهجين [14].

أما الطفرة فهي عامل مورثي أساسي يقدم مورثات جديدة في التجمع ليساعد الخوارزمية الجينية على الخروج من مصيدة الحلول الموضعية (Local Minima)، حيث تطبق الطفرة أحياناً لتغير بت واحد عشوائي في الصبغي، ويظهر الشكل 3 عملية الطفرة المطبقة على البت الخامس في السلسلة، ونظراً لعدم ارتباط التعديل الذي تحدثة الطفرة بأي بنى تركيبية سابقة للصبغي في التجمع تشكل التراكيب الجديدة لزيادة فضاء البحث.



الشكل 3 عامل الطفرة[14] .

إعادة الإدخال

تم تطوير عدة استراتيجيات لتحديد كيفية وأي الأفراد من الجيل الجديد سيحل محل الجيل القديم، عند تحديد أي أعضاء التجمع القديم يجب تبديله فإن الاستراتيجية الأكثر وضوحاً هي تبديل الأعضاء الأقل لياقة، ليؤكد أن تنفيذ استبدال الأعضاء الأقل لياقة يتم بفعالية استراتيجية نخبوية تَعُدُّ أنَّ الأكثر لياقة هو الأصلح للبقاء احتمالياً خلال الأجيال المتعاقبة، فيضمن استخدام النخبوية عدم تناقص اللياقة الأعظمية للتجمع من جيل لآخر، وتنتج عادة تقارباً أسرع للتجمع.

3-2 عنقدة k-means:

تستخدم عنقدة k-means كطريقة لتقسيم المعطيات إلى عدد k من المجموعات (العناقيد)، حيث تعامل هذه الطريقة كل عينة من المعطيات ككائن أو غرض (object)، وكل غرض له مكانه في الفضاء فتقوم بإيجاد الأغراض الأكثر قرباً من بعضها وتضعها في عنقود واحد بحيث تكون أبعد

ما يمكن عن العنقود الآخر. ويُعْرَفُ كل عنقود بمركزه والأغراض المنتمية له، ومركز العنقود هو النقطة التي يكون مجموع المسافات بينها وبين بقية الأغراض أقل ما يمكن. تكون خطوات الخوارزمية كما يأتي [16]:

1-اختر عشوائياً عدداً k من العينات لتكون مراكز للعناقيد (centroid).

2- احسب المسافات بين مركز العنقود وجميع العينات من أجل كل مركز.

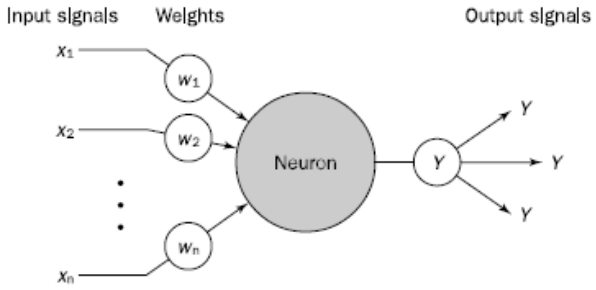
3- اربط بين كل عينة والعنقود ذي المركز الأقرب.

4- احسب متوسط العينات في كل عنقود للحصول على مراكز جديدة.

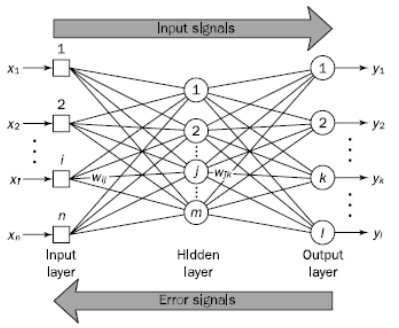
5- أعدْ الخطوة من 2 إلى 4 حتى تثبت المراكز، أو حتى تصل إلى العدد الأعظمي المحدد من الدورات.

2-4 الشبكة العصبونية:

هي تقنية حسابية مصممة لمحاكاة الطريقة التي يؤدي بها الدماغ البشري مهمة معينة، ذلك عن طريق معالجة ضخمة موزعة على التوازي ومكونة من وحدات معالجة بسيطة تدعى العصبونات (neurons). يقوم العصبون بجمع إشارات الدخل الموزونة، ومن ثم يحسب مستوى التنشيط (خرج العصبون) من خلال تطبيق أحد توابع التنشيط كما هو مبين في الشكل 4، ترتبط العصبونات مع بعضها عن طريق وصلات، ولكل وصلة وزن عددي خاص بها فتتعلم الشبكة عن طريق التعديلات المتكررة للأوزان، حيث تعدُّ هذه الأوزان الذاكرة طويلة الأمد للشبكات العصبونية. ويبين الشكل 5 بنية الشبكة العصبونية متعددة الطبقات ذات الانتشار الخلفي



الشكل 4 بنية العصبون[17] .



الشكل 5 شبكة عصبونية ذات انتشار خلفي بثلاث طبقات [17].

تتكون الشبكة العصبونية متعددة الطبقات من طبقة دخل و طبقة مخفية واحدة أو أكثر تحوي على مجموعة من العصبونات الحسابية و من طبقة خرج، وتنتشر إشارات الدخل في اتجاه أمامي بتتابع من طبقة إلى أخرى. يكون لخوارزمية التعلم في الشبكات العصبونية ذات الانتشار الخلفي طوران: في الطور الأول يتم إدخال بيانات التدريب إلى الشبكة عبر طبقة الدخل، وتقوم الشبكة بنقل هذه البيانات من طبقة إلى أخرى حتى تصل إلى طبقة الخرج، فإذا كان خرج الشبكة مختلفاً عن الخرج المطلوب يتم حساب الخطأ، ويبدأ الطور الثاني بنقل الخطأ بتغذية خلفية خلال الشبكة من طبقة الخرج عبر الطبقات المخفية إلى طبقة

الدخل مرة أخرى مع تعديل الأوزان خلال مرحلة الانتقال [17].

3- منهجية البحث:

تمت تجزئة العمل إلى قسمين: التصنيف عن طريق التعليم الموجه باستخدام الشبكات العصبونية، والتصنيف عن طريق التعليم غير الموجه باستخدام العنقدة. تكمن أهمية التعليم غير الموجه في تحديد قدرة السمات (المداخل) على تصنيف البيانات دون أي توجيه أو معرفة تامة بالتصنيف الحقيقي للبيانات، فالسمات وحدها تكون قادرة على تمييز العينة وتصنيفها. أما في التعليم الموجه فتلقن خوارزمية التعليم تصنيف كل عينة مما يتيح لها تدريب الأوزان في كل طبقة وضبطها. استخدمت الشبكة العصبونية ذات التغذية الأمامية والانتشار العكسي في هذا البحث لتصنيف العينات الموجودة في قاعدة المعطيات، حيث أعطت أداءً عالياً ولذلك اتجه البحث نحو تقليل عدد السمات الداخلة إلى الخوارزمية مع الاحتفاظ بالأداء العالي للشبكة العصبونية، واستُخدمتْ الخوارزمية الجينية لهذا الغرض حيث كان تابع الهدف (objective function) هو زيادة أداء الشبكة العصبونية و تقليل الخطأ، تم الحصول على السمات الأكثر تأثيراً في أداء الشبكة، وتم حساب العدد الأمثلي لها، وتم اختبار السمات المُثلى على التعليم غير الموجه لتحديد قدرتها على التصنيف باستخدام هذا النوع من التعليم.

4-مناقشة النتائج:

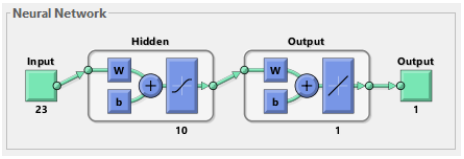
4-1 الشبكة العصبونية ذات التغذية الأمامية:

استُخدمتْ في هذا البحث الشبكة العصبونية لتصنيف العينات المريضة والسليمة في قاعدة المعطيات السابقة، وهي شبكة عصبونية ذات تغذية أمامية وانتشار عكسي ذات طبقة مخفية واحدة و10 عصبونات كما هو موضح في الشكل6، حيث تم حساب عدد العصبونات s من المعادلة (2) حسب [13]:

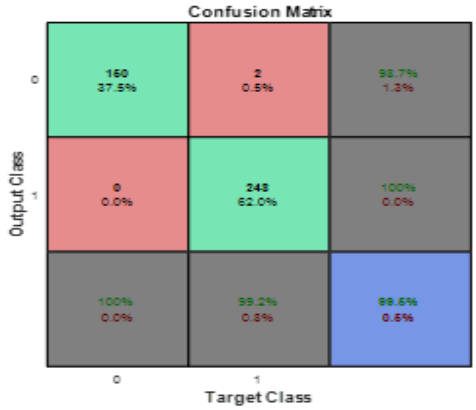
$$s = \sqrt{0.43n \times m + 0.12n^2 + 2.54m + 0.77n + 0.35} + 0.51 \quad .. (2)$$

حيث m هو عدد المداخل ،و n هو عدد المخارج.

أعطت الشبكة أداء ممتازاً حيث كان مجموع مربعات الأخطاء مساوياً 0.0049 و دقة الشبكة 99.5% كما تبين مصفوفة الارتباك في الشكل7، وهذا يتوافق مع المراجع [13] و [11].



الشكل 6 بنية الشبكة العصبونية المستخدمة



الشكل (7) مصفوفة الارتباك الناتجة عن أداء الشبكة

4-2 استخدام الخوارزمية الجينية في استخلاص السمات:

بما أنَّ استخدام الشبكة العصبونية أعطى دقة عالية فهل يمكن استخدام عدد أقل من السمات (مداخل الشبكة

العصبونية) والحصول على الأداء نفسه؟ للإجابة عن هذا السؤال صُمِّمتْ في هذا البحث خوارزمية جينية بحيث يكون تابع الهدف هو أداء الشبكة العصبونية، في البداية تم اختيار عدد السمات مساوياً لـ 8 سمات، وكانت مهمة الخوارزمية الجينية هي البحث عن السمات الثماني التي تعطي الأداء الأفضل للشبكة العصبونية. تم تكرار الخوارزمية الجينية 7 مرات، بعد ذلك تم تقليل عدد السمات إلى 4 سمات وكُرِّرتْ الخوارزمية الجينية 8 مرات، حافظت الشبكة العصبونية على أدائها لذلك تم تقليل عدد السمات إلى 3 فانخفض أداء الشبكة العصبونية، ويوضح الجدول 2 السمات المتكررة التي أنتجتها الخوارزمية الجينية في كل مرة.

الجدول 2 السمات المتكررة الناتجة عن الخوارزمية الجينية.

	السمات الأكثر تكراراً	sg	al	bgr	hemo	rc	htn	dm
GA مع 8 سمات	التكرار من أصل 7	7	7	4	7	5	2	6
GA مع 4 سمات	التكرار من أصل 8	8	6	4	8	1	2	3
GA مع 3 سمات	التكرار من أصل 5	5	1	0	5	0	4	0

ويمكن ملاحظة السمات المتكررة الأكثر أهمية في الجدول 2 والتي تكررت في كل مرة تم تطبيق الخوارزمية الجينية فيها وهي sg و hemo، حيث اختارت الخوارمية الجينية هاتين السمتين من بين السمات المختارة في كل مرة. أما السمة الثالثة الأكثر تكراراً فكانت al عندما كان عدد السمات 8 و4. وكانت هذه السمة هي htn عندما كان عدد السمات 3 فقط، وهذا يعطي أهمية لهذه السمة عندما تكون مداخل الشبكة 3 فقط، ومن ثمَّ تكون السمات الأربع الأكثر أهمية هي:sg و hemo و htn و al.

4-3 نتائج التعليم غير الموجه باستخدام k-means clustering:

لمعرفة مدى قدرة السمات على تصنيف البيانات دون أي توجيه، أي دون معرفة الخرج الحقيقي للبيانات، استخدم في هذا البحث خوارزمية k-means clutering. تم استخدام السمات الـ23 لتصنيف العينات، فكانت دقة تصنيف العينات 96.8% وعندما استخدمت 4 سمات كانت الدقة 99%، أما عندما استخدمت 3 سمات فكانت الدقة 99.5%، أي أن السمات الثلاث sg و al و htn قادرة على توصيف البيانات وتصنيفها دون وجود معلّم أو موجه. ويبين الشكل 8 مصفوفات الارتباك الناتجة عن استخدام k-means clustering.

4-4 تحسين أداء الشبكة العصبونية:

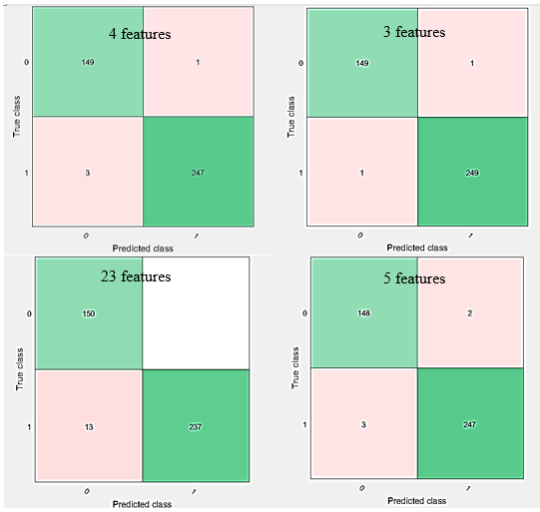
لتحسين أداء الشبكة العصبونية تمت زيادة عدد الطبقات المخفية إلى 3 طبقات مخفية بحيث يكون عدد العصبونات في الطبقتين الأولى والثانية 15 عصبوناً، وفي الطبقة الثالثة 20 عصبوناً كما يوضح الشكل 9. ومن ثمَّ تمت مقارنة الأداء مع الشبكة السابقة المؤلفة من طبقة مخفية واحدة مؤلفة من 10 عصبونات. تحسن أداء الشبكة ودقتها في الحالات الثلاث (3 سمات و4 سمات و23 سمة)، وتقارب منحني أداء الشبكة أكثر مع هبوط مجموع مربعات الأخطاء إلى القيمة الدنيا لمجموعات التدريب والتقييم والاختبار حيث كانت دقة الشبكة 100% من أجل 23 سمة. ويبين الشكل 10 مصفوفات الارتباك للشبكة العصبونية من أجل 3 سمات و4 سمات، ونلحظ أنه كلما قل عدد السمات ينخفض أداء الشبكة العصبونية، فمن الشكل 10 a وكما هو ملاحظ من

مصفوفة الارتباك نجد أن الدقة هي 99.8% من أجل 3 سمات بينما تكون 100% من أجل 4 سمات، وأنَّ الحساسية هي 99.3% من أجل 3 سمات و100% من أجل 4 سمات، أما النوعية فهي 100% من أجل كل من 3 سمات و4 سمات.

5-الاستنتاجات و الخلاصة:

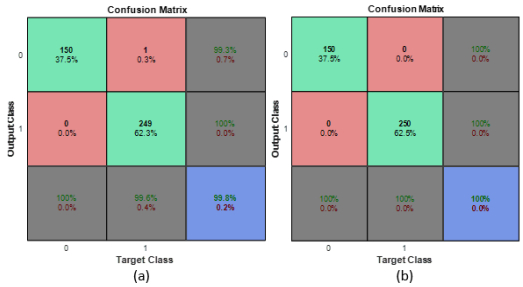
تمكَّن البحث من استخلاص السمات الأكثر صلة بالكشف عن مرض الكلى المزمن من خلال قاعدة بيانات مرض الكلى المزمن المؤلفة من 400 عينة والمتاحة على شبكة الإنترنت، وذلك بالاعتماد على خوارزميات الذكاء الصنعي، والمقارنة بين التعليم الموجه والتعليم غير الموجه للتأكد من النتائج. ووجد البحث أنَّ السمات الأكثر أهمية هي: الوزن النوعي (sg)، وكمية الهيموغلوبين في الدم (hemo)، وقياس ضغط الدم (htn) والألبومين (al)، وتبين النتائج وجود ارتباط كبير بين السمات في قاعدة البيانات، أي أن قيمة أحد السمات تعدُّ مؤشراً يدل على بعض السمات الأخرى، ومن ثمَّ يمكن تخفيض عدد السمات للتخلص من السمات المترابطة. وهذه النتائج تتوافق مع [10] و[13] اللذين حصلا على 14 سمة و8 سمات بالترتيب من بينهم السمات المستنتجة من بحثنا، كما أنَّها تتوافق و[11] الذي أشار أنَّ أداء خوارزميات التنبؤ بمرض الكلى المزمن يبقى نفسه عند استخدام ثماني سمات أو أكثر.

من السمات المهمة الأكثر صلة والتي استنتجها البحث قياس ضغط الدم، وقد أكدت الدراسات ارتباط ارتفاع ضغط الدم وفقر الدم بمرض الكلى المزمن[2,3].



الشكل 8 مصفوفة الارتباك الناتجة عن خوارزمية k-means clutering باستخدام عدد مختلف من السمات.

الشكل 9 بنية الشبكة العصبونية المعدلة.



الشكل 10 مصفوفة الارتباك الناتجة عن الشبكة العصبونية: (a) من أجل 3 سمات و(b) من أجل 4 سمات

References المراجع

1. **Zeng XX, Liu J, Ma L, Fu P. (2018) Big Data Research in Chronic Kidney Disease. Chin Med J, 131, pp. 2647-50.**
2. **Bello AK, Ronksley PE, Tangri N, et al. (2017) A national surveillance project on chronic kidney disease management in Canadian primary care: a study protocol. BMJ Open, 7:e016267.**
3. **Wang J, Bao B, Shen P, et al. (2019) Using electronic health record data to establish a chronic kidney disease surveillance system in China: protocol for the China Kidney Disease Network (CK-NET)- Yinzhou Study. BMJ Open, 9:e030102.**
4. **Jing Zhao, Shaopeng Gu , Adam McDermaid (2019) Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression, Mathematical Bio- sciences, 310,pp.24-30.**
5. **Obrador, G.T., et al.(2011) Establishing the global kidney disease prevention network (KDPN): a position statement from the National Kidney Foundation. American Journal of Kidney Diseases, 57(3): p. 361-370.**
6. **Rucci, P., et al., (2013) A clinical stratification tool for chronic kidney disease progression rate based on classification tree analysis. Nephrology Dialysis Transplantation, 29(3): p. 603-610.**
7. **Stevens, L.A., et al. (2006) Assessing kidney function—measured and estimated glomerular filtration rate. New England Journal of Medicine, 354(23): p. 2473-2483.**
8. **Gaspari, F., et al. (2004) Performance of different prediction equations for estimating renal function in kidney transplantation. American Journal of Transplantation, 4(11): p. 1826-1835.**
9. **Sarada, J. and Lakshmi, Neelam Venugopal Muthu, (2018) Data Analytics on Chronic Kidney Disease Data. IADS International Conference on Computing, Communications & Data Engineering (CCODE).**
10. **Elhoseny M., Shankar K., Uthayakumar J.,(2019) Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease. Scientific reports, 9:9583.**
11. **Gharibdousti M. S., Azimi K., Hathikal S., Won D. H. (2017) Prediction of Chronic Kidney Disease Using Data Mining Techniques, Proceedings of the 2017 Industrial and Systems Engineering Conference, May, pp:20135-2140.**
12. **"UCI Machine Learning Repository: Kidney failure Data Set [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_ Kidney_Disease [Last accessed on 21/12/2019].**
13. **Misir R, Mitra M, Samanta RK. (2017) A reduced set of features for chronic kidney disease prediction. J Pathol Inform;8:24.**
14. **Chipperfield, A., Fleming, P., Pohlheim, H. and Fonseca, C. (1994). Genetic algorithms toolbox user's guide, Res. Rep. 512, Dept. Automatic Control and Systems Engineering, Univ. Sheffield, Sheffield, U.K.**
15. **Goldberg, D., E. (1989). Genetic algorithms in search, optimisation and machine learning. Addison Wesley Longman, Publishing Co. Inc., New York.**
16. **Lloyd, Stuart P. "Least Squares Quantization in PCM." IEEE Transactions on Information Theory. Vol. 28, 1982, pp. 129–137.**
17. **Negnevitsky, M. (2008) Artificial Intelligence: A Guide to Intelligent Systems. 2nd Edition, Pearson Education, Harlow.**

Received	**2020/2/5**	**إيداع البحث**
Accepted for Publ.	**2020/6/24**	**قبول البحث للنشر**