تحسين التنبؤ بدرجة خطورة حالة المرضى المصابين بالفيروس التاجي كورنا

د.م. فطمة الطراب[1] ، أ.د.عصام الأمين[2] ، م. دعاء حمرة[3]

الملخص

خلفية البحث وهدفه: رغم مرور أكثر من عامين على بداية انتشار فيروس كورونا لا تبدو المؤشرات على أن الجائحة على وشك الرحيل بالرغم من حملات اللقاحات الضخمة والإجراءات ضد هذا الوباء، وكذلك توافر اختبارات اكتشاف العدوى مبكرا.

وقدّم خلال السنتان الماضيتان نماذج متعددة للتعلم العميق باستخدام الشبكات العصبونية للوصول الى نماذج آلية سريعة تساعد في الكشف عن الحالات المصابة بمرض فيروس كورونا كوفيد 19، واعتمد الباحثون فيها على مجموعة كبيرة من السمات.

يهدف البحث الى تخفيض عدد هذه السمات مع المحافظة على أداء المصنف ومعرفة أهم السمات التي تساعد في كشف درجة خطورة الإصابة بهذا المرض وفق حالات الشفاء والوفاة من اجل إيقاف انتقال المرض وانتشاره وخاصة في أماكن العمل الشديدة الخطورة مثل بيئات العمل في الرعاية الصحية.

مواد البحث وطرائقه: استخدمت الخوارزميات الاتية لتخفيض السمات: خوارزميةBoruta، الخوارزمية الوراثية، خوارزمية الجار الاقرب KNN، معامل Ridge مع الشبكات العصبونية، ونوقشت كل منها وفقاً لعدد السمات المستخلصة في كل طريقة.

النتائج: أظهرت النتائج أن كل خوارزميات الاستخلاص تتفق بسمة العمر كأهم سمة لتزايد خطر الاصابة بفيروس كورونا. وأن ارتفاع درجة الحرارة سمة مشتركة بدرجات أهمية مختلفة بين خوارزميات الاستخلاص. كما أن السعال وارتشاح السوائل وعدد العدلات تلعب دورا مهما في ازياد احتمال الاصابة بفيروس كورونا.

الاستنتاج: إن أقل عدد من السمات التي ساعدت في كشف خطورة حالة المرضى للإصابة بفيروس كورونا مع البقاء على دقة أداء الشبكة في كشف الاصابة هو ثلاثة سمات حسب كل من الخوارزميات خوارزمية Boruta ومعامل Ridge، بينما اكبر عدد من السمات هو خمسة وذلك حسب كل من الخوارزمية الوراثية وخوارزمية الجار الأقرب KNN.

كلمات مفتاحية: كوفيد19، الخوارزمية الجينية، الشبكات العصبونية، خوارزمية الجار الأقرب KNN.

[1] مدرسة - قسم أمراض وجراحة الأذن والأنف والحنجرة والوجه والعنق – كلية الطب البشري – جامعة دمشق .
eng.fatmah.tarab@gmail.com

[2] أستاذ مساعد - قسم أمراض وجراحة الأذن والأنف والحنجرة والوجه والعنق – كلية الطب البشري – جامعة دمشق .

[3] باحثة - قسم الهندسة الطبية- كلية الهمك - جامعة دمشق .

Improving the prediction of the severity of the condition of patients infected with the Corona virus

Dr.Fatmah Tarrab [4], Dr.Isam Alamin [5], Eng. Doaa Hmra [6]

Abstract

Background and aim: Although more than two years have passed since the beginning of the spread of the Corona virus, there are no indications that the pandemic is about to leave, despite the massive vaccination campaigns and measures against this epidemic, as well as the availability of tests to detect infection early.

During the past two years,were multiple models of deep learning using neural networks presented to reach rapid automated models that help detect cases of COVID-19, and researchers relied on a wide range of features.

This research aims to reduce the number of these features while keeping the performance of the classifier intact and knowing the most important features that help in detecting the degree of risk of infection according to the cases of recovery and death in order to stop the transmission and spread of the disease, especially in high-risk workplaces such as work environments in health care.

Materials and methods: The following algorithms were used for feature reduction: Boruta algorithm, genetic algorithm, KNN algorithm and Ridge coefficient with neural networks, and each of them was discussed according to the number of features extracted in each method.

Results: The results showed that all the extraction algorithms agreed about the age as the most important for the increased risk of infection with COVID-19. The high temperature is a common feature with various degrees of importance among the extraction algorithms. In addition, coughing, fluid leaching and the number of neutrophils play an important role in increasing the possibility of infection with the Corona virus.

Conclusion: The least number of features that helped in detecting the severity of the patients' case of infection is three according to (the Boruta algorithm and the Ridge coefficient), while the largest number of features is five, according to both the genetic algorithm and the KNN algorithm.

Keywords: COVID-19, genetic algorithm, neural networks.

4 Lecturer in ENT Department -Faculty of Medicine-Damascus University.

5 Associated Professor in ENT Department -Faculty of Medicine-Damascus University.

6 Researcher – Department of Biomedical Engineering- Faculty of Mechanical &Electrical Engineering - Damascus University.

1-المقدمة:

رغم مرور أكثر من عامين على بداية انتشار فيروس كورونا لا تبدو المؤشرات على أن الجائحة على وشك الرحيل بالرغم من حملات اللقاحات الضخمة والإجراءات ضد هذا الوباء، وكذلك توافر اختبارات اكتشاف العدوى مبكرا.

إن العقبة الرئيسية في السيطرة على انتشار هذا المرض هو عدم الكفاءة في الكشف المبكر عن هذا المرض وخاصة في الأشخاص الذين لا يعانون من أي اعراض [1,2].

وقد أظهرت الدراسات الحديثة بمجرد أن يبدأ تفشي مرض فيروس كورونا كوفيد 19 سيستغرق أقل من أربعة أسابيع لتمتلئ المشافي بالحالات المصابة وتطغى على نظام الرعاية الصحية ويزداد عدد الوفيات[4]. مما يستدعي البحث عن آلية فعالة في الكشف المبكر عنه من اجل إيقاف انتقال المرض وانتشاره وخاصة في أماكن العمل الشديدة الخطورة مثل بيئات العمل في الرعاية الصحية.

وقدّم خلال السنتان الماضيتان نماذج متعددة للتعلم العميق باستخدام الشبكات العصبونية للوصول الى نماذج آلية سريعة تساعد في الكشف عن الحالات المصابة بمرض فيروس كورونا كوفيد 19 في صور الأشعة المقطعية المحورية للصدر (CT) وفي صور الأشعة السينية البسيطة للصدر [3-5] .

كما قامت أبحاث أخرى في تصنيف حالات الإصابة بمرض فيروس كورونا حسب درجة العدوى بالاعتماد على قواعد البيانات المتاحة عن المرضى المصابين باستخدام صور الاشعة المقطعية المحورية للصدر [6,7]، وفي دراسة حديثة تم تقديم نموذج تنبؤ تصنيفي باستخدام الشبكات العصبونية للتنبؤ بدرجة خطورة الحالة المصابة بمرض فيروس كورونا كوفيد 19، حيث يتوقع النموذج المقترح مخاطر الوفيات على أساس المعلومات الديموغرافية والأعراض الفيزيولوجية للمرضى والتقرير الطبي لموجودات الصور الشعاعية للصدر السينية والطبقي المحوري بالإضافة الى نتائج اختبارات تحليل الدم ، وأظهرت النتائج كفاءة النموذج التصنيفي المقترح في التنبؤ بحالات الشفاء والوفاة بأعلى أداء تصنيف بدقة 95.9٪ [8] .

وحققت النماذج السابقة نجاحاً، واعتمد الباحثون فيها على مجموعة من السمات يتم الحصول عليها من صور الأشعة السينية والمقطعية المحورية للصدر (CT) أو من السمات الديموغرافية وبعض الأعراض السريرية والفيزيولوجية للمرضى بالإضافة الى نتائج اختبارات تحليل الدم.

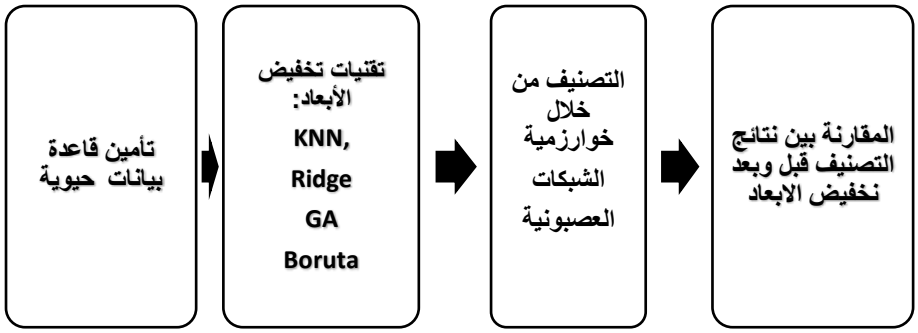
يهدف هذا البحث الى تخفيض عدد هذه السمات مع المحافظة على أداء المصنف ومعرفة أهم السمات التي تساعد في كشف درجة خطورة الإصابة بهذا المرض وفق حالات الشفاء والوفاة من اجل إيقاف انتقال المرض وانتشاره وخاصة في أماكن العمل الشديدة الخطورة مثل بيئات العمل في الرعاية الصحية وذلك عن طريق استخدام خوارزميات تخفيض الابعاد مع الشبكات العصبونية في استخلاص السمات.

2- الطرائق المستخدمة في البحث:

فيما يلي مخطط صندوقي لاهم مراحل العمل خلال هذا البحث، حيث يبدأ بتأمين قاعدة البيانات ومعالجتها ثم مرحلة تخفيض الأبعاد من خلال أربعة خوارزميات وبعدها مرحلة الحصول على النتائج ومناقشتها. فيما يلي مخطط صندوقي لاهم مراحل العمل خلال هذا البحث المبين في الشكل (1).

2-1 قاعدة البيانات وتهيئتها:

استخدمنا في هذه الدراسة مجموعة بيانات لـ (100) مريض من كلا الجنسين، وتم تأكيد تشخيص إصابتهم بمرض فيروس كورونا كوفيد 19 من خلال إيجابية



الشكل (1) مخطط صندوقي لاهم مراحل العمل خلال هذا البحث

تفاعل سلسلة البوليميراز العكسي (Reverse transcription polymerase chain reaction) . وهذه البيانات قسم منها مأخوذ من قاعدة بيانات خاصة بصور الأشعة السينية للصدر وكذلك الطبقي المحوري لـ(586) مريض من مستودع GitHub [15-16]. يتألف هذا المستودع من صور بالأشعة السينية / التصوير المقطعي المحوسب للمرضى الذين يعانون بشكل رئيسي من عدة متلازمات للأمراض التنفسية وهي: متلازمة الضائقة التنفسية الحادة ARDS، كوفيد 19 COVID-19، متلازمة الشرق الأوسط التنفسية (MERS) ، الالتهاب الرئوي، المتلازمة التنفسية الحادة الوخيمة (SARS). والقسم الثاني من البيانات تم الحصول عليه

من تقارير المرضى بمشفى المواساة الجامعي. وتتضمن قاعدتي البيانات العديد من المتغيرات بما في ذلك المعلومات الديموغرافية والأعراض الفيزيولوجية بالإضافة الى الاختبارات الدموية والتقارير الطبية لموجودات الصور الشعاعية والملاحظات السريرية والتي تتضمن الأمراض المزمنة والقصة السريرية للمرضى واستخلصنا منها (15) متغير (سمة) كما هو موضح في الجدول (1).

الجدول (1) السمات المدروسة وعددها (15) في قاعدتي البيانات

	السمات
المعلومات الديموغرافية	العمر
	الجنس
الأعراض الفيزيولوجية	درجة الحرارة
	نسبة الاشباع بالأكسجين
	الحمى
	السعال
	ضيق التنفس
الاختبارات الدموية	عدد الكريات البيضWBCs
	عدد العدلات
	عدد اللمفاويات
موجودات الصور الشعاعية للصدر	ارتشاح السوائل
	العتامة الخلالية
	الاندماج السنخي
	سماكة البنية النسيجية
وجود أحد هذه الأمراض المزمنة المصاحبة	السكري، أمراض القلب والأوعية الدموية، ارتفاع ضغط الدم، أمراض الكلى

وفي مرحلة تنظيف البيانات، أزلنا عناصر البيانات غير المفيدة والمتكررة، بالإضافة الى البيانات غير الكاملة والمفقودة. وتم اختيار (49) مريض من مستودع GitHub بمتوسط عمر (57) من كلا الجنسين (31) ذكور و(18) إناث بنسبة (40.9%) حالة وفاة و(59.1%) حالة شفاء، و(51) مريض

من مشفى المواساة الجامعي بمتوسط عمر (48) وبنسبة (53%) ذكور و(47%) إناث بنسبة (7.8%) حالة وفاة و(92.2%) حالة شفاء، وتأكدت إصابتهم بكوفيد 19 ولديهم جميع المعلومات المطلوبة في هذه الدراسة.

2-2 تطبيق تقنيات تخفيض الأبعاد:

استخدمت أربعة تقنيات لتخفيض الابعاد للوصول الى عدد مناسب للميزات. وهي:

2-2-1-خوارزميةBoruta:

تعتمد على إجراء العديد من خوارزميات الغابات العشوائية، نحسب أهمية جميع السمات لكل تشغيل. ومن ثم إجراء اختبار إحصائي لجميع السمات Two-Sided Equality Test. لرفض وقبول السمات[10].

2-2-2- معامل انحدار (Ridge): تعد Ridge من التقنيات القوية المستخدمة عمومًا لإنشاء نماذج بارزة في وجود عدد كبير من الميزات[11]. يضيف معامل تصحيح(plenty) يعادل مربع مطال المعاملات(الأوزان)، من خلال تابع (التكلفة) المعطى في المعادلة (1):

$$\mathbf{cost(w)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \left\{ \mathbf{yi} - \sum_{\mathbf{j=0}}^{\mathbf{M}} \mathbf{wj\ xij} \right\}^{\mathbf{2}} + \boldsymbol{\lambda} \sum_{\mathbf{j=0}}^{\mathbf{M}} \mathbf{w_j^2}$$

$\boldsymbol{\lambda}$ هي المعلمة التي تحقق التوازن بين مجموع المربعات المتبقية RSS (residual sum of squares) و مجموع مربعات الأوزان. وكانت أفضل قيمة لها $\boldsymbol{\lambda = 1}$.

2-2-3-الخوارزمية الوراثية (Genetic Algorithm):

تمر الخوارزميَّة الجينيَّة بثلاث مراحل مهمة[12]:

- انتقاء الوالدين (Parent Selection): وهي المرحلة الأولى حيث يتم اختيار الجيل الجديد بوضع جميع أفراد الجيل القديم في مجموعة، واختيار المناسب منها حسب قيمه الملاءمة.
- التَّزاوج أو العبور (Crossover): ويتم فيها المزاوجة بين كروموسومين (Parents) وتوريث الجينات لتشكل كروموسومين جديدين (Offsprings).
- الطَّفرة (Mutation): وهي المرحلة الأخيرة وتسهم بشكل جيد في الوصول إلى الحل الأمثل، حيث أن حدوث تغير مفاجئ غير متوقع (عشوائي) في الجيل يكون له تأثير إيجابي في الاقتراب من الحل الأمثل.

وتحدث هذه العمليات بشكل عشوائي وبنسبة معيَّنة. يتم تمثيل مجموعة فرعيَّة من المُميّزات بكروموسوم ويتكون منn، حيث n هو عدد المُميّزات في هذا الكروموسوم، وكل جين أو عنصر في الكروموسوم تشير إلى مَيّزة من المُميّزات. لاستخدام الخوارزميَّة الجينيَّة نقوم في البداية بإنشاء مجتمع عشوائي من الكروموسومات أو الأفراد، وتقوم الخوارزميَّة بتقييم أداء كل كروموسوم. واستناداً إلى نتائج المجتمع الأوَّل تتطور الخوارزميَّة من خلال عمليات الانتقاء، والتَّقاطع، والطَّفرة (ويتم تحديد طول الكروموسوم وعددها في المجتمع بشكل عشوائي. حيث تعمل هذه العمليات على تعديل هذا المجتمع وتوليد مجتمع جديد. حيث تم بناء الخوارزمية بالاعتماد على البارامترات التالية:

Population size=50, Number pf generation=50, Elitism=5, number of iteration=11, Fitness Function=Roc/Vars , number of variable in chromosome=12, Crossover rate=0.7, Mutation Rate=0.03, maxiter=50.

2-2-4-خوارزمية الجار الأقرب (KNN)K-Nearest Neighbors:

هي خوارزمية بسيطة تخزن جميع الحالات المتاحة لديها من بيانات التدريب وتصنف الحالات الجديدة

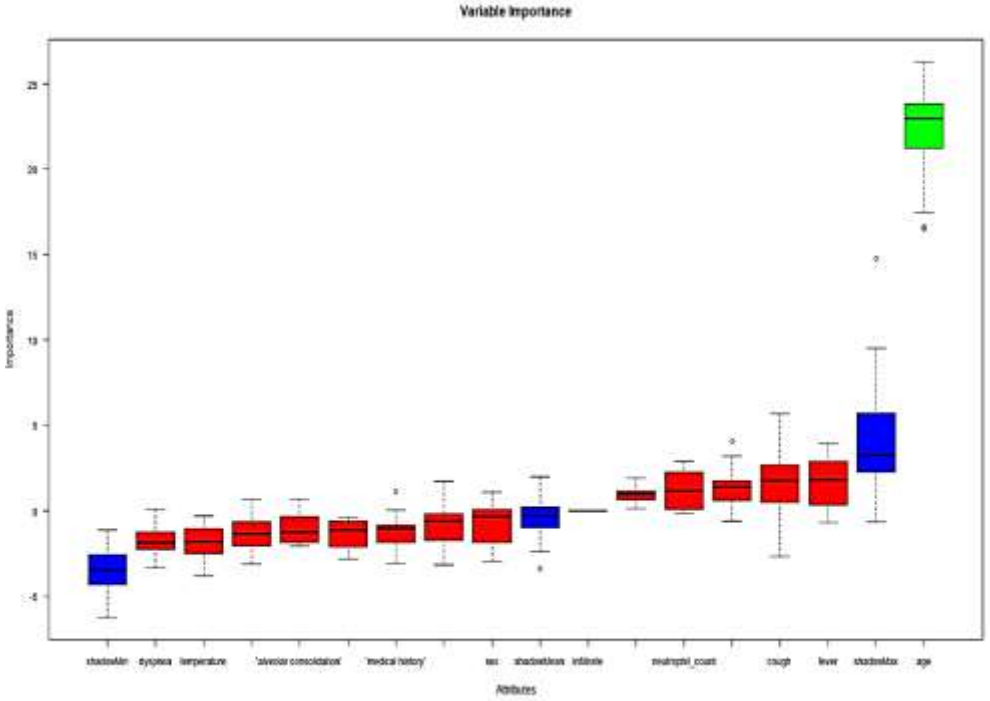
حسب أغلبية جيرانها، وذلك عن طريق وظيفة حسابية لقياس المسافة بينهم. يمثل مصنف الجار الأقرب كل كائن كنقطة بيانات في فراغ أبعاده d، حيث أن d هو عدد السمات. فإذا كان لدينا كائن اختبار. فإننا نحسب قرابته (proximity) إلى بقية نقاط البيانات في مجموعة الاختبار وذلك باستخدام مقاييس التشابه[13]. وقد تم بناء الخوارزمية باختيار k=33 وتابع التشابه هو المسافة الإقليدية.

3-النتائج ومناقشتها:

تم تخفيض عدد السمات الى عدد أصغر، واختلف العدد باختلاف الطريقة المستخدمة.

3-1-نتائج خوارزميةBoruta:

اختارت خوارزمية **Boruta** السمات التالية كأهم السمات التي تلعب دورا مهما في ظهور covid19 وهي العمر والسعال والحمى كما يبين الشكل (2).



الشكل (2) نتائج استخدام خوارزمية Boruta

3-2 نتائج الخوارزمية الوراثية (Genetic Algorithm):

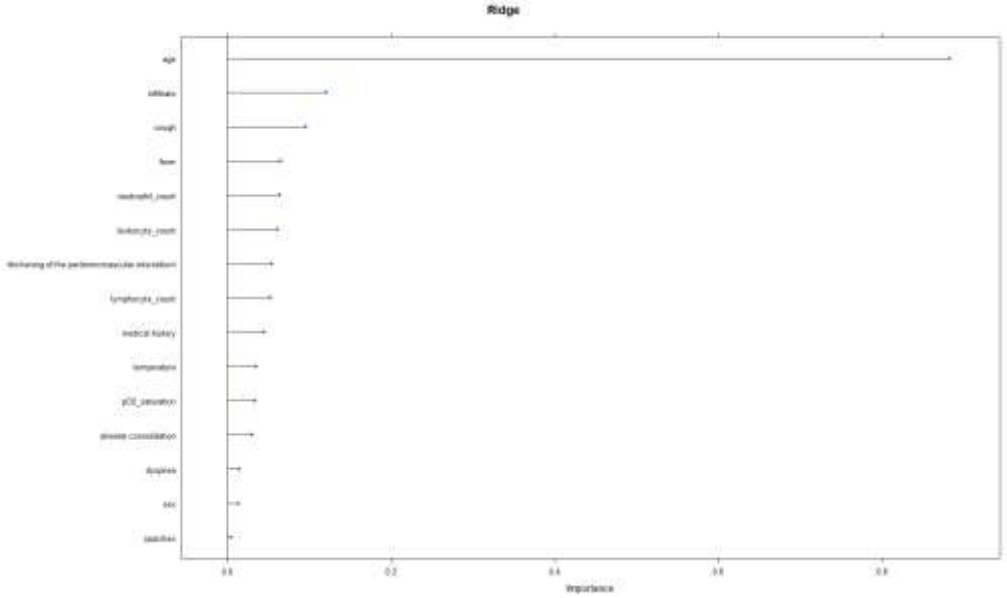
سمحت الخوارزمية الجينية بخفض عدد السمات الى خمسة سمات وهي العمر وارتشاح السوائل والحمى وعدد العدلات وتشبع po2 كما يظهر في الجدول (2).

الجدول(2) نتائج استخدام الخوارزمية الوراثية

features	importance
age	100%
infiltrate	100%
fever	66.7%
pO2_saturation	66.7%
neutrophil_count	66.7%

3-3 نتائج معامل Ridge:

تم خفض عدد السمات من خلال عامل مستوى الاهمية الى ثلاثة سمات بنسب أهمية متفاوتة. ويبين الشكل(3) مقدار أهمية السمات من خلال معامل Ridge وهي العمر(1) وارتشاح السوائل (0.17) والسعال (0.12)



الشكل (3) نتائج استخدام معامل Ridge

4-3 -نتائج خوارزمية الجار الأقرب KNN:

عند استخدام خوارزمية الجار الأقرب KNN تم ملاحظة خفض عدد السمات الى خمسة سمات بحسب مستوى الأهمية إلى العمر والسعال والحمى والقصة السريرية وعدد اللمفيات كما يظهر في الجدول (3).

الجدول(3) نتائج استخدام خوارزمية الجار الأقرب KNN

features	Importance
age	100.000
cough	47.902
fever	32.517
medical.history	23.776
leukocyte_count	22.378

5-3-العوامل المشتركة بين المورثات المستخلصة:

1. تتفق كل خوارزميات الاستخلاص بسمة العمر كأهم سمة لتزايد خطر الاصابة بفيروس كورونا.

2.إن ارتفاع درجة الحرارة سمة مشتركة بدرجات أهمية مختلفة بين خوارزميات الاستخلاص.

3. إن السعال وارتشاح السوائل وعدد العدلات تلعب دورا مهما في ازياد احتمال الاصابة بفيروس كورونا.

4- نتائج تصنيف الشبكات العصبونية قبل وبعد استخلاص السمات:

تم بناء الشبكة بحيث احتوت طبقة الدخل على 15عقدة والطبقة الخفية تضمنت طبقتين في كل منها العصبونات التالية على الترتيب (5 و2)عند عتبة قدرها 0.01 وطبقة خرج واحدة وذلك عند التعامل مع كامل قاعدة البيانات أي قبل استخلاص السمات أما بعد استخلاص السمات فقد تم بناء الشبكة بn طبقة دخل وهو عدد السمات الناتجة عن خوارزمية الاستخلاص و2 طبقة خفية في كل منها العصبونات التالية على الترتيب (5,1) وبعتبة قدرها(0.01). حيث تم تقسيم البيانات 70%تدريب و30%اختبار ويظهر الجدول (4) نتائج التصنيف

من خلال الشبكات العصبونية قبل وبعد استخلاص السمات .

الجدول (4) التصنيف من خلال الشبكات العصبونية قبل وبعد استخلاص السمات

	الخوارزمية Algorithm	الحساسية Sensitivity	النوعية Specificity	الدقة Accuracy
قبل استخلاص السمات	الشبكات العصبونيةMLP	0.95	0.95	0.956
بعداستخلاص السمات	الشبكات العصبونيةBoruta + MLP	**0.9333**	**1.0000**	**0.95**
	الشبكات العصبونيةGA +MLP	1	1	1
	الشبكات العصبونيةKNN +MLP	1	1	1
	الشبكات العصبونيةRidge +MLP	0.5	1	0.8

5- أداء تصنيف الشبكات العصبونية قبل وبعد استخلاص السمات:

تم الوصول الى دقة جيدة عند إدخال كافة السمات الى الشبكة العصبونية ولكن بعد تخفيض السمات بالطرق الموجودة تم الوصول إلى دقة أعلى، ومن ذلك نستنتج أن الصفات المستخلصة من أغلب الطرق ذات أهمية كبيرة وتعبر عن قاعدة البيانات هذا ما أدى بدوره إلى تخفيض الكلفة وزيادة سرعة المعالجة، إلا اننا نلاحظ انخفاض دقة خوارزمية الشبكة العصبونية بعد استخلاص السمات بطريقة ridge وذلك لأن عدد السمات كان قليل ثلاثة سمات فقط مقارنة مع باقي الطرق من ضمنها سمة لم تكن موجودة عند بقية الطرق المستخدمة ذات أهمية أقل مقارنة بباقي السمات.

6- خلاصة:

يسهم هذا البحث في تحديد أهم السمات لدرجة خطورة الإصابة بفيروس كورونا باستخدام الشبكات العصبونية مع خوارزميات تخفيض السمات، وأظهرت النتائج أن كل خوارزميات الاستخلاص تتفق بسمة العمر كأهم سمة لتزايد درجة خطورة حالة المرضى المصابين بفيروس كورونا، وأن ارتفاع درجة الحرارة سمة مشتركة بدرجات أهمية مختلفة بين خوارزميات الاستخلاص. كما أن السعال وارتشاح السوائل وعدد العدلات تلعب دورا مهما في ازياد احتمال درجة خطورة حالة المرضى المصابين بفيروس كورونا.

7–المراجع:

[1] World Health Organization. Novel Coronavirus (2019-nCoV) situation reports.

[2] Chung M, et al (2020) CT imaging features of 2019 Novel Coronavirus (2019-NCoV). Radiology, p. 200230. DOI.org (Crossref).

[3] Ozcan, T. (2020). A Deep Learning Framework for Coronavirus Disease (COVID-19) Detection in X-Ray Images.

[4] Trent McConghy, Bruce Pon, Eric Anderson (2020),"When does Hospital Capacity Get Overwhelmed in USA? Germany? A model of beds needed and available for Coronavirus patients" trent.st.

[5] Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine, 103792.

[6] Singh, D., Kumar, V., & Kaur, M. (2020). Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. European Journal of Clinical Microbiology & Infectious Diseases, 1-11.

[7] Pourhomayoun, M., & Shakibi, M. (2020). Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. medRxiv.

[8] د. م. فطمة الطراب. (2021). التنبؤ بدرجة خطورة حالة المرضى المصابين بالفيروس التاجي كورونا باستخدام الشبكات العصبونية .مجلة جامعة دمشق للعلوم الهندسية.37(4) ,

[9] https://github.com/UCSD-AI4H/COVID-CT.

[10] KURSA ,M., RUDNICKI,W., 2010- Feature Selection with Boruta Package, Journal of Statistical Software, vol. 36, 1–13pp.

[11] OGUTU, J. O., SCHULZ-STREECK, T., PIEPHO, H. P. 2012-Genomic selection using regularized linear regression models: ridge regression, LASSO, elastic net and their extensions. In BMC proceedings ,Vol. 6, No. 52, S10 p.

[12] L. Haldurai., T, Madhubala., R, Rajalakshmi ,2016- A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering,vol 4. 139-143pp.

[13] NEGNEVITSKY, M., 2005- Artificial Intelligence, A guide to intelligent systems. 2nd edition, pearson Education..

[14] U. S. N. L. o. M. R. P. M. U. National Center for Biotechnology Information, National Center for Biotechnology Information.

[15] Huang C, Wang Y, Li X, et al. (2020),Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet.395(10223):497-506.

[16] Lippi, G., & Plebani, M. (2020). Laboratory abnormalities in patients with COVID-2019 infection. Clinical Chemistry and Laboratory Medicine (CCLM), 58(7), 1131-1134