مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (8) العدد(3) 2024
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

استخدام تقنيات التعلم الآلي والتعلم العميق في الكشف عن سرطان الرئة بالاعتماد على مثيلات الحمض النووي

د. يعرب ديوب *

سناء علي إبراهيم **

(تاريخ الإيداع 2023/12/19 . قُبِل للنشر في 2024/3/11)

□ ملخّص □

السرطان سبب رئيسي للوفاة في جميع أنحاء العالم وهو الخطر الأكثر شيوعاً الذي يهدد صحة الإنسان، والتشخيص المبكر مطلوب للتشخيص المناسب. أظهرت العديد من الدراسات أن السرطان ناجم بشكل رئيسي عن طفرات ضارة في الجينات المسرطنة الأولية وجينات كبت الورم وجينات منظم دورة الخلية، مما يعزز التعرف على السرطان بناءً على المعلومات الجينية. على الرغم من أن العديد من الجينات التي تم العثور عليها لها دور رئيسي في حدوث وانتشار السرطان، إلا أن الآليات المسببة للأمراض للطفرات الجينية والتفاعلات بين الجينات غير معروفة إلى حد كبير. يُعد تحديد نوع السرطان الفرعي أحد أفضل الحلول للكشف عن أمراض السرطان وتحسين اقتراح العلاج الشخصي. في هذه الدراسة، تم توظيف تقنيات تعلم الآلة في بيئة عمل TensorFlow بهدف تحديد الخلية المصابة بسرطان الرئة بناء على بيانات مثيل الحمض النووي من منصة البيانات k450 التي تم الحصول عليها من موقع جينوم أطلس السرطان The Cancer Genome Atlas TCGA. تم استخدام تقنيات التعلم الآلي، تحديداً تقنية الغابة العشوائية Random Forest، لتقليص عدد الميزات بهدف تقليل حجم البيانات وجعل بنية التعلم العميق المطبقة باستخدام الشبكة العصبية التلافيفية CNN قابلة للتطوير، حيث حسنت هذه الخطوة دقة النموذج بشكل كبير حيث بلغت 91.28%، في حين أنه بدون تطبيق هذه المرحلة كانت دقة النموذج منخفضة جداً إذ بلغت 40.84% فقط.

الكلمات المفتاحية: مثيل الحمض النووي، السرطان، المعلوماتية الحيوية، الغابات العشوائية، التعلم العميق، الشبكات العصبية التلافيفية

* استاذ في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

** طالبة ماجستير في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (8) العدد(3) 2024
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

Using of Machine Learning Techniques and Deep Learning for identifying Lung Cancer based on DNA methylation

*** Dr. Yarub Dayoub**
**** Sanaa Ali Ibrahim**

(Received 19/12/2023 . Accepted 11/3/2024)

□ ABSTRACT □

Cancer is a major cause of death worldwide, and an early diagnosis is required for a favorable prognosis. Numerous studies have shown that cancer is mainly caused by harmful mutations in proto-oncogenes, tumor suppressor genes and cell cycle regulator genes, potentiating the identification of cancer based on genomic information. Although many genes that have been found have major roles in the occurrence and spread of cancer, the pathogenic mechanisms of gene mutations and interactions between genes are largely unknown. The identification of cancer subtype is one of the best solutions for detecting cancer and improving personalized treatment proposal. In this study, machine learning techniques were employed using TensorFlow platform to identify the lung cancer cell based on DNA methylation data from the 450k data platform obtained from The Cancer Genome Atlas TCGA. Machine learning technique, specifically Random Forest, was used to reduce the number of features in order to reduce data volume and make deep learning model implemented by Convolutional Neural Network scalable. This step greatly improved the model's accuracy, reaching 91.28%, whereas without applying this stage, the model's accuracy was low, reaching only 40.84%.

Key Words: DNA Methylation, CpG, Cancer, Bioinformatics, Random Forest, Deep Learning, Convolutional Neural Network

* Professor, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.
** Master student, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

١. المقدمة:

تعد مثيلات الحمض النووي DNA methylation مهمة في تطور السرطان وتقدمه نظراً لدوره في تثبيط الجينات الكابتة للورم أو تعزيز تعبير الجين الورمي [1]. تنطوي هذه العملية اللاجينية على إضافة مجموعة الميثيل CH3 إلى موضع زوج قاعدة السيتوزين في الحمض النووي للكائن الحي مما يؤثر بشكل مباشر على التعبير الجيني [2].

يمكن حساب بيانات المثيلة باستخدام تقنيات تسلسل عالية الإنتاجية [3] حيث يمكن أن يكون لدى متبرع واحد أكثر من 850.000 علامة مثيلة قابلة للاكتشاف (يعبر عنها بـ CpGs وهي تعني موقع السيتوزين بالنسبة للغوانين في شريط الحمض النووي [4] cytosine–phosphate–guanine) عبر الجينوم البشري. تسمح تقنيات التسلسل الأحدث الآن بتقييم المثيلة في كل موقع جينوم باستخدام بيانات تسلسل الجينوم الكاملة.

بالنظر إلى نسبة العلامات (CpG markers) المرتفعة للعينات في مجموعات بيانات السرطان، فمن الضروري إنشاء إطار عمل آلي قادر على معالجة مثل هذه الكمية الهائلة من المعلومات، وتقليل تحيز التنبؤ. يجب أن تحقق أطر العمل الخاصة بالتنبؤات نسب عالية من الدقة والحساسية والخصوصية، ولإنجاز ذلك يجب أن تتم معالجة مجموعات البيانات بشكل دقيق لتحقيق وقت تدريب أقل مما يسهل عملية اكتشاف المعرفة الأفضل عبر مجموعات البحث وأنواع السرطان.

يعتبر التعلم العميق [5،6] (مجموعة فرعية من التعلم الآلي) أحد الحلول لمشكلة البيانات الكبيرة [7] التي اكتسبت زخماً كبيراً بسبب قدرتها على استخراج مجموعة فرعية ذات معنى من الميزات من مجموعات البيانات المختلفة [8،9] دون أي معالجة مسبقة. بالرغم من أن خوارزميات التعلم العميق تتفوق في التنبؤ، يمكن أن تكون أيضاً ذات تعقيد حسابي عالي، أي أنها يمكن أن تتطلب، بالنسبة لحجم إدخال معين، عدداً كبيراً نسبياً من الخطوات لإكمالها. يمكن أن يشكل ذلك، بالإضافة إلى مجموعات بيانات عالية الأبعاد، تحدياً كبيراً في تحقيق دقة عالية في التدريب والتنبؤ [10]. للتغلب على هذه العقبات يمكن استخدام تقنيات استخلاص السمات قبل تطبيق خوارزميات التعلم العميق [11] بحيث تزيل هذه التقنيات المعلومات المتكررة من البيانات الكبيرة وبالتالي تقليص قيود الذاكرة.

نُشرت أول مقالة بحثية لاستخدام التعلم العميق على مجموعات بيانات المثيلة في عام 2016 [12] حيث قدمت نموذجاً للتعلم العميق قادراً على التنبؤ بحالة مثيلة الحمض النووي من علامات CpG باستخدام الخلايا K562. منذ ذلك الوقت، اقتصرت الأبحاث في هذا المجال على مجموعات البيانات الصغيرة أو على أنواع معينة من السرطانات. في دراسة أجراها Angermuelle وآخرون [13]، تم تدريب وحدات DNA وCpG من بيانات تسلسل بيكبريتيت الخلية المفردة (scBS–seq) وبيانات تسلسل بيكبريتيت مخفضة الخلية المفردة (scRRBS–seq) الخاصة بـ Mus musculus باستخدام التعلم العميق للتنبؤ بحالات المثيل في خلية ما.

استخدم الباحثون في [14] التعلم العميق لاستخراج حالات مثيلة الحمض النووي من قراءات تسلسل Nanopore ووجدوا أن دقة التنبؤ أفضل من التقنيات التقليدية مثل نماذج ماركوف المخفية (HMM).

في حين استخدم ليو وآخرون. [15] التعلم الآلي لاستخراج علامات مثيلة CpG لـ 27 نوعاً من السرطان من إجمالي 13526 عينة، حيث كانت 10140 عينة سرطانية و3386 عينة طبيعية. استخدم المؤلفون الإحصائيات لاستخراج أفضل 2000 علامة CpG من أصل 485000 موقع CpG. تمت تصفية العلامات المختارة أيضاً بناءً على Random Forest وتم استخدام 12 علامة فقط لتدريب نموذج التعلم العميق، بينما استخدم هذا البحث مجموعة بيانات أكبر بكثير، فإن عملية استخراج الميزات يدوياً لاختيار أفضل 2000 علامة قضت على أكثر من 99% من

ا
ل
ص
ف
ح

مواقع CpG. تيان وآخرون. [16]، استخدموا بيانات تسلسل ثنائي كبريتيت الجينوم الكامل (WGBS) الخاصة بالخلايا الجذعية البشرية للتنبؤ بما إذا كانت عينات الإدخال غير مثيلة أو مثيلة مفرطة أو متوسطة المثيلة.

استنادا إلى العمل السابق في هذا المجال تبيّن لدينا محدودية الدراسات التي قامت بتطبيق خطوة استخلاص السمات أثناء بناء نموذج التنبؤ الخاص بها والذي يعتمد على تقنيات التعلم العميق، لذلك كانت أهمية هذا البحث هي بناء نموذج للتنبؤ بسرطان الرئة، ومقارنة الدقة الناتجة بين حالتين، الأولى هي بدون إجراء استخلاص للسمات الاكثر أهمية، والحالة الثانية هي في حال تم تطبيق هذا الإجراء.

٢. أهمية البحث وأهدافه:

تأتي أهمية هذا البحث من كونه يقدم نموذج تنبؤي جديد في عملية التعرف على الخلية السرطانية الرئوية وتمييزها عن الخلية الطبيعية وذلك بناءً على تحليل مثيلات الحمض النووي. أيضاً تتجلى أهمية البحث بدمجه لإحدى أهم تقنيات التعلم الآلي مع التعلم العميق، إذ يتجلى ذلك أولاً في مرحلة استخلاص السمات باستخدام نموذج الغابة العشوائية RF في الحصول على أكثر مواقع مثيل الحمض النووي التي تلعب دور في تحديد الخلية المصابة بالورم، ومن ثم استخدام الشبكة العصبية التلافيفية Convolutional Neural Network CNN في عملية التنبؤ مما يزيد من دقة النموذج المنجز.

٣. طرائق البحث ومواده:

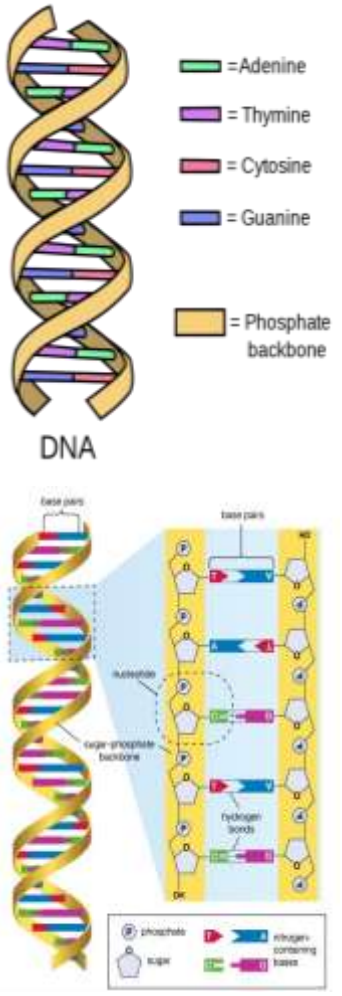
يقدم هذا البحث مراحل تطوير نموذج يعتمد تقنيات التعلم الآلي والتعلم العميق لتحديد الخلية المصابة بسرطان الرئة وذلك بالاعتماد على تحليل بيانات مثيل الحمض النووي DNA Methylation.

٣,١ الحمض النووي DNA Deoxyribonucleic acid:

يعرّف الحمض النووي (كما هو موضح في الشكل (1) الجانبي) بأنه جزيء المعلومات الذي يحمل تعليمات لصنع جزيئات أكبر أخرى، تسمى البروتينات. يتم تخزين هذه المعلومات داخل كل خلية من خلايا الانسان، موزعة على 46 بنية طويلة تسمى الكروموسومات Chromosomes. هذه الكروموسومات تتكون من آلاف الأجزاء الأقصر من DNA، والتي تسمى الجينات Genes. كل جين يخزن اتجاهات صنع أجزاء البروتين، البروتينات الكاملة أو عدة بروتينات محددة [17].

يتكون جزيء الحمض النووي من نيوكليوتيدات Nucleotides (جزيئات صغيرة) مرتبطة الواحدة تلو الأخرى في سلسلة طويلة جداً. تتألف أبجدية الحمض النووي من 4 أحرف فقط، 4 نيوكليوتيدات، وهي A, T, C ,G وهي اختصار لـ Adenine, Thymine, Cytosine, Guanine على التوالي [17].

كل نيوكليوتيد مبني من ثلاثة أجزاء جزيئية وهي: السكر deoxyribose، مجموعة الفوسفات Phosphate group، قاعدة النواة Nucleobase، يتم تصنيع كل نيوكليوتيد عن طريق لصق مجموعة فوسفات وقاعدة نواة على السكر، يجدر بالذكر أن السكر والحمض acid في جميع النيوكليوتيدات الأربعة متماثلان. روابط الفوسفات في شريط الحمض النووي، دائماً تربط الكربون الثالث 3' من النيوكليوتيد الأول مع الكربون 5' من النيوكليوتيد الثاني، وتسمى هذه الروابط بروابط phosphodiester ويعبر عنها بـ 3'–5'. ترتبط النيوكليوتيدات A دائماً بالنيوكليوتيدات T، بينما ترتبط C مع G بشكل دائم [17].

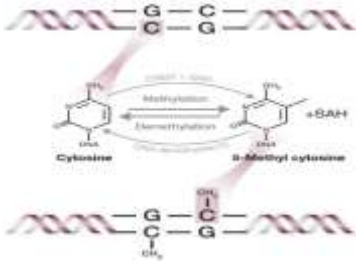


الشكل (1) بنية الحمض النووي DNA

٣,٢ مثيل الحمض النووي DNA Methylation:

مثيلة الحمض النووي هي تعديل لاجيني يتضمن إضافة مجموعة المثيل (CH_3) إلى جزيء الحمض النووي. إنها عملية أساسية وقابلة للعكس reversible وتلعب دوراً حاسماً في تنظيم التعبير الجيني gene expression والتحكم في العمليات البيولوجية المختلفة في الكائنات الحية، بما في ذلك البشر [18].

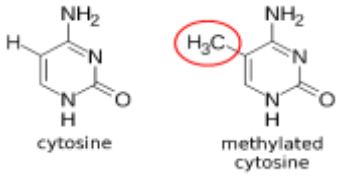
فيما يلي بعض النقاط الرئيسية حول مثيلة الحمض النووي:



الشكل (2) مثيلة الحمض النووي

ا
ل
ص
ف
ح

- مثيلة السيتوزين Methylation of Cytosine: تتضمن مثيلة الحمض النووي، موضحة في الشكل (3)، بشكل رئيسي مثيلة نيوكليوتيدات السيتوزين C. على وجه الخصوص، يتم حدوث المثيلة في موضع الكربون 5 في حلقة السيتوزين، مما يؤدي إلى 5−methylCytosine (5mC) [18].



الشكل (3) مثيلة السيتوزين mC5

- تنظيم التعبير الجيني: يمكن أن يكون لمثيلة الحمض النووي تأثيران رئيسيان على التعبير الجيني:

- فرط المثيل Hypermethylation: عندما تكون المنطقة المحفزة للجين مثيلية بشكل كبير، فإنها عادة ما تؤدي إلى إسكات الجينات gene silencing، مما يمنع نسخ الجين والتعبير عنه [18].
- نقص المثيل Hypomethylation: يمكن أن يؤدي انخفاض المثيل أو إزالة المثيل لمحفزات الجينات إلى زيادة التعبير الجيني [18].

٣,٣ العلاقة بين الحمض النووي DNA ومرض السرطان:

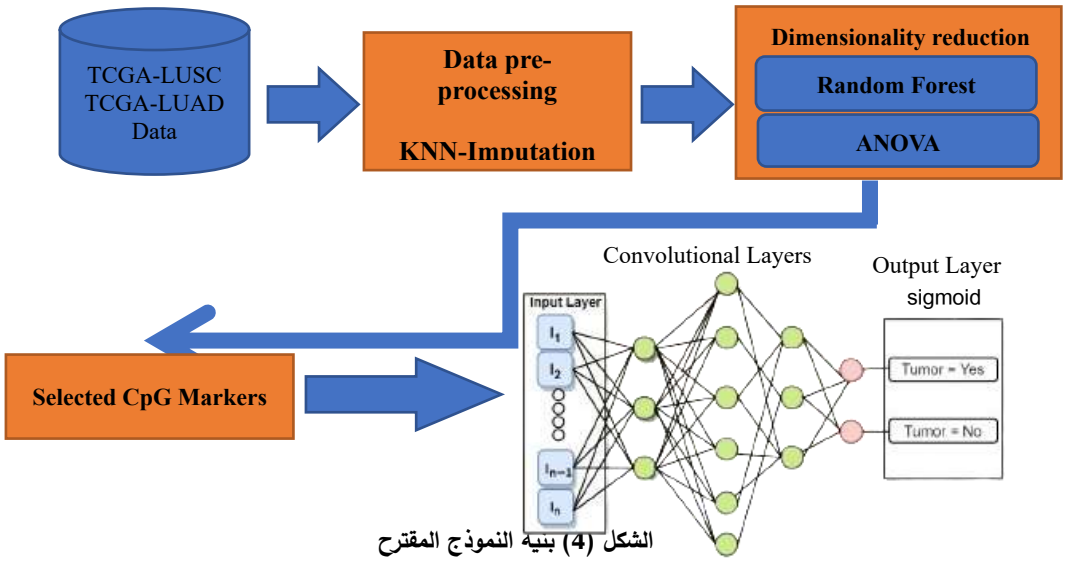
السرطان هو مجموعة معقدة من الأمراض التي تتميز بالنمو الغير طبيعي والغير منضبط للخلايا في الجسم. العلاقة بين DNA والسرطان منشأية، لأنها تنطوي على تغيرات جينية ولاجينية تؤدي إلى بدء السرطان وتطوره [19].

فيما يلي بعض الجوانب الرئيسية للعلاقة بين الحمض النووي والسرطان:

- الطفرات الجسدية Somatic mutations: تنجم عنها العديد من أنواع السرطان، وهي عبارة عن تغيرات جينية تحدث في الحمض النووي للخلايا الفردية خلال حياة الشخص. يمكن أن تحدث هذه الطفرات بسبب عوامل مختلفة، مثل التعرض للمواد المسرطنة، أو أخطاء النسخ، أو التغيرات الجينية التلقائية [19].
- التغيرات اللاجينية: مثل مثيلة الحمض النووي وتعديلات الهيستون. يمكن أن تؤدي أنماط مثيلة الحمض النووي الشاذة وتعديلات هيستون إلى إسكات الجينات الكابتة للورم وتنشيط الجينات المسرطنة [18].

٣,٤ بنية النموذج المقترح:

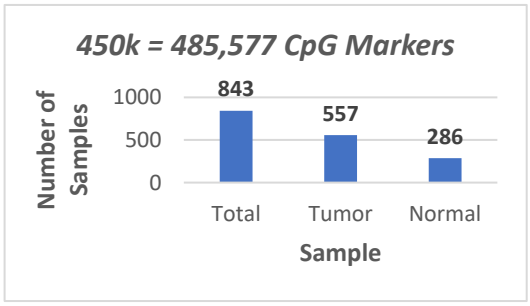
تبدأ بنية النموذج المقترح بمرحلة معالجة البيانات، ثم مرحلة استخلاص السمات وأخيراً تطبيق الشبكة العصبية التلافيفية CNN كما هو موضح في الشكل (4).



الشكل (4) بنية النموذج المقترح

٣,٤,١ مجموعات البيانات Datasets:

تم استخدام مجموعتين من بيانات مثيلات الحمض النووي الخاصة بسرطان الرئة من موقع TCGA [20]، وهي بيانات TCGA-LUSC و TCGA-LUAD تشملان عينات من منصة Illumina 450K والتي توفر تحليل مثيلة الحمض النووي في أكثر من 450,000 موقع CpG [4]. تم تحميل هذه البيانات باستخدام برنامج GDC Data Transfer Tool الخاص بالموقع نفسه [20]. يبين الشكل (5) توزع بيانات المثيلة المستخدمة.



الشكل (5) توزع العينات المدروسة

تكون مجموعة بيانات مثيل الحمض النووي عند الحصول عليها مؤلفة من سمتين هما (الجدول (1)):

١. PROBE_ID: وهي تحدد موقع CpG

٢. Beta_value: قيمة بيتا β وهي تمثل مستوى المثيلة وتكون قيمتها بين 0 و1

الجدول (1) نموذج لمجموعة البيانات

	A	B
1	probe_id	beta_value
2	cg22501393	0.025827074
3	cg18895155	0.013117202
4	cg27126442	0.13958522
5	cg15264255	0.102252424
6	cg18464559	0.023592437
7	cg20379125	0.013719298

•موقع CpG: المعروف أيضاً باسم موقع CG أو CpG Islands، هو تسلسل محدد من الحمض النووي حيث يتبع نيوكليوتيد السيتوزين C نيوكليوتيد الغوانين G، مرتبطان بمجموعة فوسفات p (phosphodiester bond) [2].

-تعتبر مواقع CpG حاسمة في علم الوراثة اللاجينية لأنها غالباً ما تكون هدف مثيلة الحمض النووي.

-CpG Islands: لا تتوزع مواقع CpG بالتساوي في جميع أنحاء الجينوم. فهي تميل إلى التجمع في مناطق تعرف باسم جزر CpG والتي هي مناطق من الحمض النووي. يمكن أن تؤثر ميثيلة مواقع CpG داخل جزر CpG على نسخ الجينات، وعندما يتم فرط المثيل، يمكن أن يؤدي إلى إسكات الجينات [2].

•قيمة Beta [3]: تعد قيمة بيتا β، مقياساً شائعاً يستخدم في سياق تحليل مثيلة الحمض النووي، إذ تمثل درجة الميثيلة في موقع CpG محدد. يتم حساب قيمة بيتا كنسبة السيتوزينات الميثيلية (mC-5) إلى

ا
ل
ص
ف
ح

العدد الإجمالي للسيتوزينات (الميثيلية وغير الميثيلية) في موقع CpG معين. ويتم التعبير عنها عادةً كقيمة عشرية بين 0 و1، حيث:

-تشير $\beta = 0$ إلى عدم مثيل أي من السيتوزينات الموجودة في موقع CpG.

-يمثل $\beta = 1$ مثيلة كاملة، مما يشير إلى أن جميع السيتوزينات الموجودة في موقع CpG تمت ميثلتها.

يتم الحصول على هذه القيم عادة من خلال تقنيات تحديد مثيلة الحمض النووي، مثل bisulfite sequencing أو microarray-based assays [3].

4.4.1.1 المعالجة المسبقة للبيانات Data pre-processing:

•**تشكيل قاعدة البيانات:** في البداية، تم تحويل تنسيق البيانات لتكون ممثلة بأعمدة تعبر عن الميزات (مواقع CpG) وأسطر تمثل العينات (المرضى). لتصبح على الشكل التالي:

الجدول (2) تحويل تنسيق البيانات

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	cg00000029	cg00000108	cg00000109	cg00000165	cg00000236	cg00000289	cg00000292	cg00000321	cg00000363	cg00000622	cg00000658	cg00000714	cg00000721	cg00000734
2	0.251236977	0.345898803	0.906956032	0.238860644	0.910125222	0.812338267	0.829333801	0.216151448	0.788299409	0.013582165	0.841872954	0.239718433	0.951906125	0.065546095
3	0.102621221	0.960120323	0.91757923	0.85254165	0.917259076	NA	0.664057464	0.577938023	0.840576714	0.013349902	0.886087061	0.074612164	0.943630406	0.047023058
4	0.288162419	0.959272227	0.926184102	0.173343886	0.924995307	NA	0.629480067	0.640094968	0.411057103	0.019406842	0.882170561	0.170136594	0.960494033	0.103333852

-ثم تم إجراء عملية labeling، وإضافة عمود التسميات إلى قاعدة البيانات ليتم استخدامه كمتغير هدف للخرج، حيث تم إسناد القيمة 1 للخلية السرطانية والقيمة 0 للخلية الطبيعية.

-كما تم التوضيح مسبقاً أن البيانات المهمة في مجموعة بيانات مثيل الحمض النووي هي قيمة بيتا، لكن هناك بعض القيم التي تعيق الشبكة من العمل وهي عندما تكون: β=NA ,β=0 لذلك يجب أن يتم معالجة هذه البيانات قبل أن يتم إدخالها إلى نموذج التصنيف. تمت المعالجة باستخدام آلية cut-off ومن ثم تطبيق تقنية التضمين K-NN imputation K-nearest neighbors.

-Cut-off [21,22]: هي تقنية إحصائية يتم استخدامها في تحديد نسبة البيانات المفقودة لسمة معينة في قاعدة البيانات، والتي من الممكن أن تتسبب في تحيز النتائج وتؤثر سلباً على دقة التنبؤ. لذلك وبسبب حجم البيانات الكبير الذي نتعامل معه، فإنه توجد احتمالية كبيرة لوجود العديد من علامات CpG المفقودة والتي يمكن أن تؤثر على أداء الشبكة ولا يكون لها أي أهمية في زيادة دقة العمل خصوصاً إذا كانت هذه العلامات مفقودة في معظم العينات التي نتعامل معها. لذلك تم التأكد بشكل يدوي أولاً من وجود العديد من السمات في قاعدة البيانات المختارة التي تحتوي على عدد كبير من القيم المفقودة، وبعضها لا يحوي أي قيمة في أي سجل من البيانات. لهذا السبب ومن خلال تحليل البيانات قمنا بتطبيق آلية القطع cut-off بنسبة 30%، تم الحصول على النتائج الموضحة في الجدول التالي:

الجدول (3) عدد المواقع CpG بعد تطبيق آلية cut-off على مجموعات البيانات

Illumina Platform	Total CpG markers	null CpG	Cut-off applied	Remaining CpGs
450K	485,577	89,671	30%	395,722

بعد تطبيق آلية القطع وإزالة السمات ذات القيم المفقودة بشكل كبير، سيتم تضمين القيم المفقودة المتبقية باستخدام تقنية التضمين وفق الجار الأقرب K-NN، تم الاعتماد على هذه التقنية في التضمين بعد التجريب مع تقنيات أخرى مثل Mean imputation وZero imputation، حيث كانت تقنية الجار الأقرب K-NN ذات النتائج الأفضل والأقل خطأً.

-تطبيق تضمين K-nearest neighbors على مواقع CpG المتبقية [23]: يتم فيه تقدير القيم المفقودة بناءً على قيم أقرب مواقع CpG المجاورة لـ K. تأخذ هذه الطريقة في الاعتبار تشابه أنماط المثيلة عبر العينات.

لأجل كل عينة ذات قيمة مفقودة، يتم حساب المسافة بينها والعينات الأخرى، وقد تم اعتماد القياس الاقليدي في حساب هذه المسافات. بالمرحلة الثانية يتم تحديد K-NN لكل عينة ذات قيم مفقودة، وذلك بالاعتماد على المسافات المحسوبة حيث يتم اختيار العينات المجاورة k الأصغر مسافة عن العينة ذات القيمة المفقودة (عدد الجيران k التي تؤخذ بعين الاعتبار تحدد بالتجريب لعدة قيم لـ k). أخيراً يتم ملئ القيم المفقودة وفقاً لقيم هذه الميزات في العينات المجاورة. يوضح الجدول -4- مقاييس الخطأ والانحراف المعياري الناتجة عن عملية التضمين [23].

تم تطبيق التضمين باستخدام مكتبة scikit-learn، والتابع KNN-imputer، وتم تحديد قيمة k=2 (بعد التجريب بين قيم k=1,2,3).

جدول (4) مقياس الخطأ والانحراف المعياري بعد عملية التضمين

Metric	KNN-Imputation
Mean Square Error MSE	0.017253
Standard Deviation STD	0.005286

٣,٤,٢ استخلاص الميزات Feature selection:

من الشائع في مرحلة استخلاص السمات أن يتم استخدام تقنيات مثل ANOVA، و Tukey's HSD وغيرها. لكن في هذا البحث تم اقتراح طريقة جديدة للحصول على أكثر السمات أهمية في عملية التمييز بين الخلية الطبيعية والخلية السرطانية. لذلك بهدف تحسين جودة البيانات، تم تطبيق عمليتين هما: الغابة العشوائية Random Forest RF و تحليل التباين أحادي الاتجاه one-way ANOVA كما هو موضح في مخطط العمل الشكل (4).

ANOVA: هي عبارة عن تقنية يمكن استخدامها لاختيار الميزات أثناء تحليل بيانات مثيل الحمض النووي لتحديد مواقع CpG التي تظهر اختلافات كبيرة في مستويات المثيلة عبر مجموعات مختلفة. الهدف من إجراء تحليل ANOVA هو تقليل عدد مواقع CpG حيث تمت إزالة العلامات ذات القيمة p الأكبر من قيمة العتبة 0.05 من إجمالي الميزات، حيث أن p تعبر عن احتمال يقيس الدليل ضد فرضية العدم [24]. يعبر عن هذه التقنية بالعلاقة التالية [24]:

$$F = MSB / MSE$$

ا
ل
ص
ف
ح

حيث أن: **MSB** يعبر عن متوسط مجموعالمربعات بين المجموعات، **MSE** يعبر عن متوسط مربعات الأخطاء، يعطى كل منهما بالعلاقات التالية:

MSB = SSB / (K-1) , MSE = SSE / (N-K)

حيث أن: K هو عدد المجموعات (في الدراسة هنا K=2)، و N يعبر عن عدد العينات الاجمالية.

$$SSB = \sum n_j (\bar{X}_j - \bar{X})^2 \qquad SSE = \sum\sum (X - \bar{X}_j)^2$$

nj: هو حجم العينة للمجموعة j، -

$\bar{X}$ $\bar{X}$j: متوسط المجموعة j، **$\bar{X}$j - X**: المتوسط الاجمالي، **X**: قيمة العينة في كل مجموعة.

تم تطبيق تقنية ANOVA بشكل منفصل على 395722 علامة من مجموعة بيانات K450، وكنتيجة لذلك تم تقليل عدد الميزات إلى 125949 موقع CpG.

Random Forest RF: تعد الغابات العشوائية [25] من أهم خوارزميات التعلم الآلي الأكثر شيوعاً، يمكن من خلالها استخلاص أهمية كل متغير في شجرة القرار. بمعنى آخر، من السهل حساب مقدار مساهمة كل متغير في عملية اتخاذ القرار. تعمل هذه الخوارزمية عن طريق إنشاء العديد من أشجار القرار في وقت التدريب ثم يتم جمع التوقعات للحصول على نتائج أكثر دقة، مما يساعد على تقليل الـ overfitting وتحسين التعميم. تقوم خوارزمية الغابة العشوائية بحساب أهمية الميزة من خلال المعادلات الرياضية التالية:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

حساب أهمية كل عقدة (عقدة الجذر، عقد القرار، والعقد الأخيرة leaf من أشجار القرار)

$$n_i = \frac{N_t}{N}[impurity - (\frac{N_{t(right)}}{N_t} * right\ impurity) - (\frac{N_{t(left)}}{N_t} * left\ impurity)]$$

N_t: عدد الصفوف في كل عقدة، N: عدد الصفوف الإجمالية، $N_{t(right)}$: عدد العقد في العقدة اليمنى، $N_{t(left)}$: عدد العقد في العقدة اليسرى.

في حين أن قيمة الشائبة impurity لكل عقدة يعبر عنها بالعلاقة التالية:

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2$$

Pi: هي عبارة عن مجموع مربعات احتمالات الفئة الموجبة والسالبة.

▪باستخدام مكتبة sklearn في لغة Python، تم بناء نموذج RF (حيث تم تعيين عدد الأشجار إلى القيمة 100) وتدريبه على السمات التي خضعت لعملية المعالجة المسبقة، وكما في مرحلة استخلاص الميزات باستخدام ANOVA، تم تطبيق هذه الخوارزمية بشكل منفصل على مجموعة البيانات.

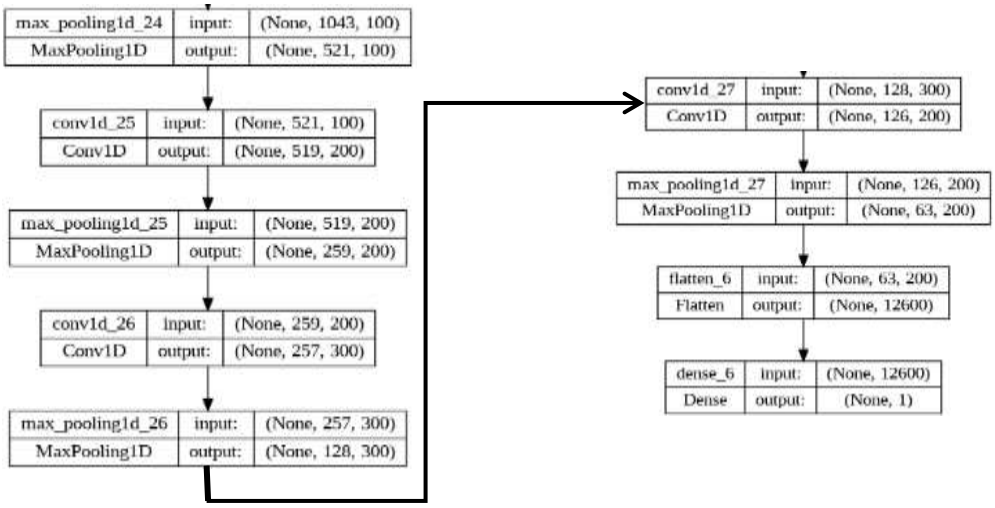
▪نتج عن نموذج RF معدلات أهمية كل ميزة (موقع CpG) في تحديد نوع الخلية، بناءً على درجات الأهمية هذه تم استخلاص الميزات الأكثر أهمية والتي تلعب الدور الأكبر في عملية تحديد الخلية السرطانية.

النتيجة النهائية من مرحلة استخلاص الميزات:

لوحظ وجود تقاطع كبير بين السمات الناتجة عن كلا التقنيتين المستخدمتين، لذلك لدمج النتائج، تم تطبيق الميزات المخفضة الناتجة عن تقنية ANOVA إلى نموذج RF. أدى ذلك إلى تخفيض الميزات إلى 1045 موقع CpG يلعب دور أساسي في تحديد الخلية السرطانية، ستمثل هذه المواقع دخل الشبكة العصبية التلافيفية.

٤. النتائج والمناقشة:

تم بناء النموذج المقترح باستخدام الشبكة العصبية التلافيفية CNN، في بيئة عمل TensorFlow، بنية الشبكة موضحة في الشكل (6). تشكل مجموعة التدريب 40% من إجمالي البيانات، مجموعة التحقق 30% ومجموعة الاختبار 30%.



الشكل (6) بنية الشبكة CNN

-تم تطبيق أربع تقنيات لتحسين عملية التدريب، كما يلي:

١- Xaiver method: بهدف تهيئة أوزان الشبكة، وذلك لتجنب مشاكل التلاشي vanishing وexploding gradients. تضمن هذه التقنية تهيئة أوزان الشبكة بحيث لا تكون صغيرة جداً ولا كبيرة جداً مما يعني أن الشبكة يمكن أن تبدأ بتعلم الميزات المهمة من بداية عملية التدريب بالتالي الوصول إلى تقارب أسرع.

٢- Adam optimization: تم استخدامه بهدف تحسين عملية التدريب، حيث يقوم بضبط معدل التعلم لكل بارامتر أثناء التدريب.

ا
ل
ص
ف
ح

٣- Learning rate decay: تساعد هذه التقنية على استقرار التدريب عن طريق جعل معدل التعلم أصغر مع اقتراب النموذج من التقارب. تعمل على تحديث معدل التعلم أثناء عملية التدريب وتقليله تدريجياً (في هذا البحث، تم تعيين معدل الخطوات إلى 1019 خطوة، ومعدل انحياز 0.9).

٤- Mini-batch training: وهو التدريب على جزء من البيانات، حيث يلعب التدريب على دفعات صغيرة من البيانات دور كبير في تحسين تدريب الشبكة العصبية. يمثل حجم الدفعة الصغيرة أحد المعلمات الفائقة hyperparameter التي يمكن أن تؤثر على سرعة التدريب وتقارب النموذج. تم اختبار 3 دفعات (32, 64 ,128) في عملية التدريب ومراقبة أداء النموذج خلال التدريب على كل دفعة.

لتطبيق التقنيات الأربعة، تم اعتماد مكتبة keras في بيئة عمل TensorFlow بلغة البرمجة Python.

-بسبب عدد البيانات الكبير، تم بناء نموذج يسمح بتمرير الكثير من المعلومات من مجموعة البيانات، إذ يحتوي على 4 طبقات مخفية تحتوي كل منها على عدد معين من العصبونات موضحة في الجدول التالي:

الجدول (5) عدد العصبونات في الطبقات المخفية للنموذج

Hidden Layer 1	Hidden Layer 2	Hidden Layer 3	Hidden Layer 4
100 neurons (filters)	200 neurons	300 neurons	200 neurons

-تم تمرير هذه العصبونات (في كل طبقة مخفية) من خلال تابع التنشيط غير الخطي ReLU.

-أثناء عملية التدريب وبهدف عدم وصول النموذج إلى حالة التجاوز Overfitting، تم تحقيق الخطوات التالية:

- ✓ تدريب كل متغير في النموذج لمدة 10 epochs، أيضاً تم تحديث أوزان النموذج باستخدام الخسارة (تم استخدام الخسارة نوع cross-entropy) التي يتم الحصول عليها من مجموعة بيانات التحقق من الصحة cross validation-5.
- ✓ تطبيق منهج التدريب على حجم عينات ثابت وضبط أداء النموذج من خلال مراقبة الأداء على مجموعة التطوير (التحقق من الصحة). تم اعتماد mini-batch 128، بمعنى أنه تم تدريب النموذج على 6 دفعات بيانات صغيرة، تحتوي كل منها 128 عينة أما الدفعة الأخيرة تحتوي على 75 عينة، حيث تم تقديم مجموعة بيانات التدريب بأكملها للنموذج أثناء التدريب 10 مرات.
- ✓ بالنسبة لمعدل التعلم، تم تخفيض معدل التعلم بمعامل قدره 0.5 في حال لم تتحسن دقة التحقق بعد 5 epochs، مع معدل تعلم أدنى يبلغ 0.0001.
- ✓ تم تطبيق dropout=0.25 (باستخدام توابع مكتبة keras) بعد كل طبقة مخفية. هذه التقنية تعمل بشكل عشوائي على إلغاء تنشيط جزء صغير من الخلايا العصبية في الطبقة لكل تمريرة للأمام والخلف بحيث لا تساهم هذه الخلايا العصبية المعطلة في حساب هذا المرور مما يمنع النموذج من الوصول إلى حالة التجاوز.

الجدول (6) جدول توضيحي لعدد بيانات التدريب والزمن المستغرق في كل مجموعة

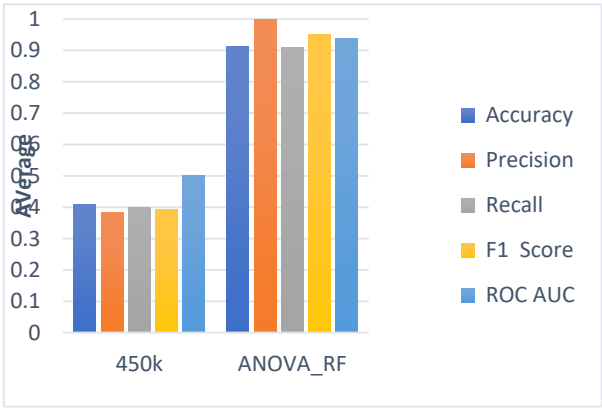
Dataset	Features	Sample size	Tumor samples	Normal samples	Runtime
All CpG markers (450K)	395,722	843	557	286	1:44:08 s
Dataset after applying ANOVA_RF	1045	843	557	286	36:42 s

تم تقييم عمل النموذج في مجموعة بيانات الاختبار لتمييز الخلية السرطانية الرئوية عن الخلية الطبيعية، النتائج موضحة في الجدول (7) والشكل (7):

الجدول (7) مقاييس تقييم اختبار النموذج 5-fold cross-validation

	A	B	C	D	E	F	G
	Dataset	Trials	Accuracy	Precision	Recall	F1 Score	ROC AUC
	All CpG markers	1	0.95769	0.95769	1	0.97839	0.5
		2	0.04231	0	0	0	0.5
		3	0.04231	0	0	0	0.5
		4	0.95769	0.95769	1	0.97839	0.5
		5	0.04231	0	0	0	0.5
		Avg.	0.40846	0.38307	0.4	0.39135	0.5
		St.dev.	0.50137	0.52455	0.54772	0.53588	0
	ANOVA_RF	1	0.93089	0.99685	0.93078	0.96268	0.93206
		2	0.89563	0.99672	0.89396	0.94255	0.91365
		3	0.86601	1 0.86009	0.92478	0.3422	0.93004
		4	0.95628	1 0.95435	0.97664	0.63885	0.97717
		5	0.91537	0.99839	0.91311	0.95385	0.93989
		Avg.	0.91283	0.99839	0.91046	0.9521	0.93856
		St.dev.	0.03431	0.00161	0.0359	0.01971	0.0236

الشكل (7) متوسط مقاييس التقييم



يوضح الجدول (7)، مقاييس دقة النموذج المقترح خلال مراحل التحقق 5-fold cross، ومتوسط الدقة والانحراف المعياري في كل مرحلة بالنسبة لمجموعتي البيانات:

- All CpG markers (450k): وهي قاعدة البيانات التي تشتمل على السمات كاملة من المنصة K450 بدون إجراء عملية استخلاص السمات.
- ANOVA_RF: وهي قاعدة البيانات بعد أن تم تطبيق استخلاص السمات باستخدام التقنيتين ANOVA والغابات العشوائية RF.

بمقارنة النتائج نلاحظ وجود تحسن كبير في دقة التنبؤ في حال تم استخلاص السمات الأكثر أهمية قبل أن يتم إدخالها على الشبكة CNN.

في هذه الورقة، تم تطوير نموذج يجمع ما بين التعلم الآلي والتعلم العميق للتنبؤ بالخلية المصابة بسرطان الرئة وتمييزها عن الخلية الطبيعية بناءً على كمية كبيرة من بيانات مثيل الحمض النووي من 843 مريض. حقق النموذج المقترح أداءً عالياً كما هو موضح في مرحلة التقييم، حيث حقق دقة أعلى 91.28% واسترجاع 91.04%.

في هذه الورقة، تم إثبات أن خطوتي تضمين البيانات المفقودة واستخلاص السمات مهمتان قبل التحليل في تطبيقات التعلم العميق في المعلوماتية الحيوية، وبالنتيجة تم تحسين دقة التنبؤ بحالات سرطان الرئة باستخدام مجموعة بيانات المثيلة من k450.

ا
ل
ص
ف
ح

تشمل نقاط القوة في هذه الدراسة حجم عينة كبير، ومعالجة مسبقة ثابتة وموحدة لبيانات المثيلة، واحتساب البيانات المفقودة. ساعد كل ذلك على تحسين دقة وموثوقية التحليل. لكن من ناحية أخرى لم يتم تجميع العينات بناءً على سمات المرض، وتعد عملية المثيلة عملية ديناميكية قد تتقلب بمرور الوقت، مما يحد من قدرتنا على تحديد تغييرات المثيلة المسؤولة عن تطور سرطان الرئة مقابل التغييرات الناتجة عن وجود سرطان الرئة.

٥. الاستنتاجات والآفاق المستقبلية:

في هذه الورقة تم تقديم جزء من النموذج المقترح والذي يستخدم في عملية صنع القرار بيانات مثيل الحمض النووي فقط. في المرحلة الثانية من البحث سيتم إضافة بيانات المريض السريرية (مرحلة الورم) وبعض المعلومات الديموغرافية مثل العمر، الجنس والعرق. ليتم مكاملتها مع بيانات مثيل الحمض النووي واستخدامها في بيئة العمل بهدف زيادة دقة النموذج. بالإضافة إلى أنه سيتم إجراء خطوة معالجة جديدة للبيانات قبل أن يتم إدخالها للشبكة وهي موازنة البيانات بهدف تحسين عملية التنبؤ ومنع التحيز نحو نتيجة محددة وفقاً للاختلاف الكبير نسبياً بين عدد العينات لنوعي الخلايا السرطانية والطبيعية.

٦. المراجع:

1. Xiao, C.L.; Zhu, S.; He, M.; Chen, D.; Zhang, Q.; Chen, Y.; Yu, G.; Liu, J.; Xie, S.Q.; Luo, F. (**2018**). *N6-methyladenine DNA modification in the human genome. Mol. Cell.*

2. Gardiner-Garden, M.; Frommer, M. (**1987**). *CpG islands in vertebrate genomes.*

3. Levin, J.Z.; Yassour, M.; Adiconis, X.; Nusbaum, C.; Thompson, D.A.; Friedman, N.; Gnirke, A.; Regev, A. (**2010**). *Comprehensive comparative analysis of strand-specific RNA sequencing methods.*

4. IlluminaHumanMethylation450kmanifest: *Annotation for Illumina's 450k Methylation Arrays.* https://bioconductor.org/

5. O'Shea, K.; Nash, R. (**2015**). *An introduction to convolutional neural networks.*

6. Zaremba,W.; Sutskever, I.; Vinyals, O. (**2014**). *Recurrent neural network regularization.*

7. Halevy, A.; Norvig, P.; Pereira, F. (**2009**). The unreasonable effectiveness of data. IEEE Intell. Syst.

8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. (**2022**). *Imagenet classification with deep convolutional neural networks.*

9. Johnson, R.; Zhang, T. (**2014**). *Effective use of word order for text categorization with convolutional neural networks.*

10. Verleysen, M.; François, D. (**2005**). *The curse of dimensionality in data mining and time series prediction. In International Work-Conference on Artificial Neural Networks*.

11. Ahsan, M.; Gomes, R.; Chowdhury, M.; Nygard, K.E.*(**2021**). Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector*.

12. Wang, Y.; Liu, T.; Xu, D.; Shi, H.; Zhang, C.; Mo, Y.Y.; Wang, Z. (**2016**). *Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks.*

13. Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. (**2017**). *DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning*.

14. Ni, P.; Huang, N.; Zhang, Z.; Wang, D.P.; Liang, F.; Miao, Y.; Xiao, C.L.; Luo, F.; Wang, J. (**2019**). DeepSignal: Detecting DNA methylation state from Nanopore sequencing reads using deep-learning.

15. Liu, B.; Liu, Y.; Pan, X.; Li, M.; Yang, S.; Li, S.C. (**2019**). DNA methylation markers for pan-cancer prediction by deep learning.

16. Tian, Q.; Zou, J.; Tang, J.; Fang, Y.; Yu, Z.; Fan, S. (**2019**). MRCNN: A deep learning model for regression of genome-wide DNA methylation.

17. Alberts B; Johnson A; Lewis J; et al. *Molecular Biology of the Cell*. 4th edition, Garland Science, New York, **2002**. https://www.ncbi.nlm.nih.gov/books/NBK26821/

18. Tim B. *DNA methylation*. **2023**; Available from: https://biomodal.com/blog/the-fascinating-world-of-dna-methylation/

19. *Genetic of cancer*. **2022**; Available from: https://www.cancer.gov/about-cancer/causes-prevention/genetics

20. The Cancer Genome Atlas TCGA; https://portal.gdc.cancer.gov/

21. Paul D; Rahael H; Kate T; Jon H. (**2019**). *The proportion of missing data should not be used to guide decisions on multiple imputation.*

22. Lingbing F; Gen N; T.J. O'Neill; A.H. Welsh. (**2014**). *CUTOFF: A spatio-temporal imputation method*

23. Kaushik R,C. *KNNImputer: A robust way to impute missing values (using Scikit-Learn). 2020.*

https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/

24. Weisstien, E.W. *ANOVA*. **2020**; Available from: https://mathworld.wolfram.com/ANOVA.html

25. C. Aldrich, L. Auret. (**2010**). *Fault detection and diagnosis with random forest feature extraction and variable importance methods.* https://doi.org/10.3182/20100802-3-ZA-2014.00020

ا
ل
ص
ف
ح