مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (5) العدد(10) 2021

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (5) No. (10) 2021

توليد تسلسلات DNA المستخدمة في تشفير النصوص والصور بالاعتماد على تقنية التعلم العميق باستخدام شبكات LSTM

د. كندة سليمان أبو قاسم *

م. تيســـير عزت ســـلمـان **

(تاريخ الإيداع 26 /8 /2021 . قُبِل للنشر في 17 /11 /2021)

□ ملخّص □

استُخدمت تسلسلات DNA في تشفير البيانات، حيث تم الاعتماد في الحصول على هذه التسلسلات من مصادر مختلفة كقواعد البيانات الجينية العامة أو من كائنات حية من خلال عمليات مخبرية معقدة. تركز هذه الدراسة على توفير التسلسلات DNA اللازمة لعملية التشفير للنصوص أو الصور، سواءً كمفاتيح تشفير أو تسلسلات تستخدم لعملية الفهرسة، وذلك بالاعتماد على تقنية التعلم العميق وتحديداً شبكات LSTM لتوليد تسلسلات DNA عشوائية، وذلك من خلال تدريب نموذج التعلم العميق على تسلسلات DNA طبيعية عشوائية تم أخذها من قاعدة بيانات جينية عامة. الهدف من الدراسة هو بناء مولد عشوائي يستخدم كبديل للتسلسلات المأخوذة من قواعد البيانات الجينية العامة لاستخدامها لاحقاً في عمليات تشفير النصوص والصور من خلال استخدام تسلسلات DNA الناتجة كمفاتيح تشفير أو لغرض فهرسة النصوص والصور في طرق التشفير التي تعتمد تقنية DNA. تم التأكد من عشوائية التسلسلات الناتجة بواسطة اختبارات NIST حيث اجتازت التسلسلات التي تم توليدها كافة الاختبارات العشوائية بنجاح وأثبتت فعاليتها عند استخدامها في تشفير النصوص والصور من خلال تطبيق المعايير الأمنية على النصوص والصور المشفرة الناتجة.

الكلمات المفتاحية: DNA، Deep Learning، LSTM ، NIST، Encryption

*أستاذ مساعد في قسم الحاسبات والتحكم الآلي من كلية الهندسة الميكانيكية والكهربائية بجامعة تشرين.

**طالب الدراسات العليا (دكتوراه) في قسم الحاسبات والتحكم الآلي من كلية الهندسة الميكانيكية والكهربائية بجامعة تشرين.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (5) العدد(10) 2021

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (5) No. (10) 2021

Generating DNA Sequences used for text and image Encryption based on Deep learning using LSTM networks

Dr. Kinda Abu Kassem *
Eng. Tayseer Izzat Salman **

(Received 26 / 8/ 2021 . Accepted 17 / 11/ 2021)

□ ABSTRACT □

DNA sequences have been used, for data encryption, acquired from different sources as from public genetic databases or sequences taken from living organisms via complex laboratory operations. This study focuses on providing DNA sequences as encryption keys or for indexing to encrypt either text or images based on deep learning technique more precisely LSTM networks for generating random DNA sequences by training the deep learning module on natural DNA sequences which taken from public genetic databases. The objective of this study is to build a random generator to be used as an alternative source for DNA sequences taken from genetic databases to be used later in text and image encryption as encryption keys or for the purpose of text and image indexing in encryption methods, which use DNA technique. Randomness of the generated sequences were tested using NIST tests. The generated sequences passed all randomness tests successfully and proved its effectiveness when applying security standards on encrypted texts and images.

Key words :DNA ،Deep Learning، LSTM ،NIST ،Encryption

1. مقدمة

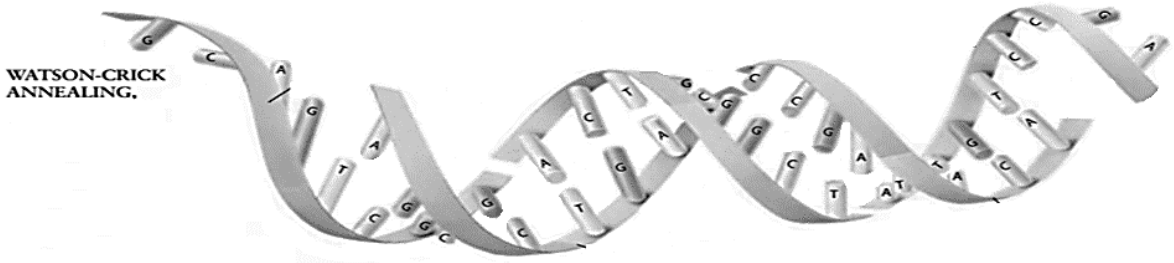
1.1. التشفير باستخدام DNA

تعتمد طريقة التشفير باستخدام DNA في البداية على ترميز النص أو الصورة بترميزات DNA، كما في الجدول (1) والذي يوضح عملية ترميز البيانات الثنائية برموز DNA أو بالعكس. بعد ذلك إجراء عمليات على النص المرمز وتشفير بالمفتاح الذي هو عبارة عن تسلسل DNA أيضاً. يتألف تسلسل DNA من أربع رموز تتالى بشكل غير منتظم الشكل (1). تم استخدام تسلسلات DNA في عمليات التشفير في العديد من الدراسات [1] و[2] و [3] و[4] و [5] و[6] حيث اعتمدت على الترميز برموز DNA وإجراء عمليات على النص المرمز وتشفيره. قدمت الدراسات [7] و [8] و [9] طرق لتشفير النصوص والصور بالاعتماد على الترميز برموز DNA وكذلك استخدمت مفاتيح للتشفير من تسلسلات DNA عشوائية مأخوذة من قواعد البيانات الجينية العامة [10].

الجدول-1: الحالات الممكنة لترميز DNA ثنائياً [11]

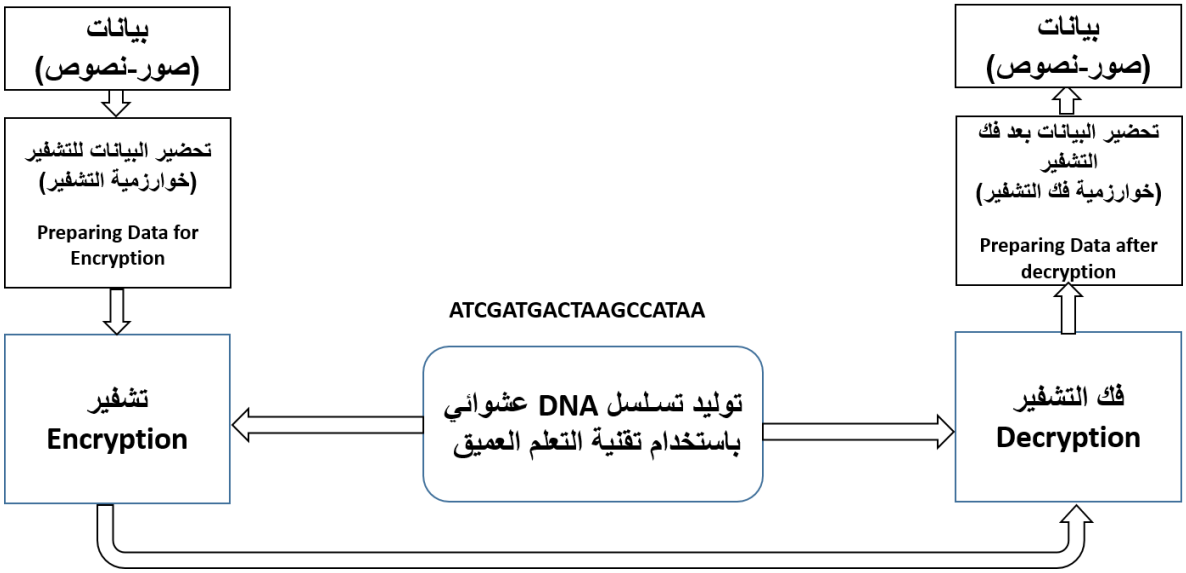
8	7	6	5	4	3	2	1	
11	11	10	10	01	01	00	00	A
00	00	01	01	10	10	11	11	T
10	01	11	00	11	00	10	01	G
01	10	00	11	00	11	01	10	C

تكمن المشكلة في تسلسلات DNA المأخوذة من قواعد البيانات الجينية العامة، بأن هذه التسلسلات تكون على شكل ملفات مضغوطة بحجوم تتراوح بين عدة كيلو بايت الى عدة ميغابايت، وقد تصل الى مئات الميغابايت وتحتاج الى فك ضغط ومن ثم الحصول على الملف الذي يحوي التسلسل وأخذ جزء منه كمفتاح للتشفير أو لأغراض الفهرسة وكذلك لا يوجد أي ضمان بتوافر هذا المصدر بشكل دائم. لذلك اعتمدت هذه الدراسة على إيجاد حل من خلال بناء نموذج للتشفير يستخدم مولد عشوائي للتسلسلات من خلال توظيف شبكات LSTM وذلك من خلال تدريب هذا النموذج على مجموعة بيانات مأخوذة حسب الدراسة [9].



الشكل 1: تمثيل فراغي لجزيء DNA حسب تصور واطسون كريك

تقترح هذه الدراسة عملية توليد تسلسلات DNAعشوائية لاستخدامها في عملية التشفير بحيث تغني عن قواعد البيانات الجينية العامة كمصدر للتسلسلات، حيث يتم دمج مولد التسلسلات مع مخطط التشفير كما في الشكل (2) والذي يبين بنية نظام التشفير المقترح حيث يتكون من جزء التشفير وفك التشفير والمولد العشوائي للتسلسلات.



الشكل 2: نموذج التشفير وفك التشفير مع مولد تسلسلات DNA عشوائي يعتمد تقنية التعلم العميق

2.1. الدراسات المرجعية

تم اقتراح اتجاهين لأنظمة توليد النص. الطريقة الأولى تحاول الحفاظ على إعادة الاستخدام وعموم الجملة بدون التركيز على هيكل الجملة. النهج الثاني يحاول الحفاظ على هيكل وقالب الجملة [12].

تم في الآونة الأخيرة استخدام نهج التعلم العميق بغرض توليد النصوص، من خلال استخدام المرمزات التلقائية المتباينة VAEs.[12] ، تكون الاستفادة من VAEs لترميز أمثلة البيانات إلى فضاء كامن ثم تمثيل عينات جديدة تم توليدها من تلك المساحة الكامنة. كانت هناك أعمال أخرى لتوليد نص السؤال والجواب باستخدام الرسوم البيانية المعرفية[13] . ينتج زوج السؤال والإجابة من رسم بياني للمعرفة Freebase كقاعدة معرفة.

هدفت الدراسة [14] إلى توليد تسلسلات باستخدام شبكات RNN على مستوى المحرف وذلك للتنبؤ بالمحرف التالي في تسلسل نصي، حيث اعتبرت الدراسة أن نموذجاً لغوياً أفضل على مستوى المحرف يمكن أن يحسن ضغط الملفات النصية ويسهل عملية التفاعل بين الأشخاص الذين لديهم إعاقات جسدية وأجهزة الحواسيب.

قدمت الدراسة [15] طريقة لتوليد النصوص بالاعتماد على شبكات التعلم العميق، حيث قامت بإجراء عمل بحثي على مهمة تلخيص النص، وصممت طريقة لتوليد الملخصات بالاعتماد على مجمع بحث محسن وقامت بإجراء التجارب لتحسين الحصول على محتوى ملخص من الانترنت.

قدمت الدراسة [16] طريقة لتوليد الأسئلة والاجوبة باستخدام شبكات LSTM من خلال إنشاء نموذج موحد يستخدم البيانات المهيكلة وغير المهيكلة باستخدام مفهوم جدول–إلى–نص والذي يهدف لتوليد توصيف بناءً على جدول معين.

قدمت العديد من الدراسات نماذج لغوية لتوليد النصوص باستخدام الشبكات العصبونية سواء RNN أو LSTM، حيث استخدمت RNN للتنبؤ وتوليد النصوص كونها قادرة على تعلم خاصية البيانات المتسلسلة مثل بيانات السلاسل الزمنية أو البيانات النصية. ومع ذلك، فإن RNN تعاني من مشكلة تلاشي التدرج Vanishing Gradient، حيث تختفي التدرجات وبالتالي لا يتم تحديث البارامترات أثناء عملية الانتشار الخلفي Back Propagation. تقدم LSTM حلاً لمشكلة التدرج المتلاشي RNN من خلال إدخال مفهوم البوابات [17].

تسمح خلايا LSTM للشبكة بتحديد ما الذي يجب تذكره أو نسيانه. أظهرت LSTM أداءً رائعًا في التعرف على الكلام [18] وترجمة اللغة [19].

اقترح [20] نموذجا لتوليد النصوص المعتمدة على السياق باستخدام LSTM، يتم تدريب النموذج المقترح على إنشاء نص لمجموعة معينة من كلمات الإدخال مع شعاع السياق. يشبه شعاع السياق شعاع الفقرة الذي يستوعب المعنى الدلالي (السياق) من الجملة. تم اقتراح عدة طرق لاستخراج أشعة السياق في هذا العمل. بسبب هذه البنية، يتعلم النموذج العلاقة بين كلمات الإدخال، شعاع السياق والكلمة الهدف. بالنظر إلى مجموعة من مصطلحات السياق، يتم إنشاء نموذج جيد التدريب متمركزاً حول السياق الذي تم تقديمه.

اهتمت العديد من الدراسات باستخدام تقنيات التعلم العميق في دراسة الحمض النووي DNA وكان معظمها منصباً على عمليات التصنيف فيما يخص مناطق على DNA إكسونات وانترونات وتحديد المناطق ذات الترميز والمناطق الخالية منها ومعاملات النسخ [21].

قدمت الدارسة [22] طريقة للتنبؤ بمعامل النسخ Transcript Factor (TF) ومواقع الربط أو التغليف والتنبؤ بوظائف المناطق غير المرمزة على DNA وذلك باستخدام شبكات التعلم العميق حيث استخدموا النموذج NCNet والذي يدمج التعلم العميق الباقي وشبكات التعلم Sequence–To–Sequence تسلسل–لتسلسل وذلك للتنبؤ بمعامل النسخ ومواقع التغليف والتي يمكن أن تستخدم للتنبؤ بوظائف المناطق غير المرمزة على DNA.

3. أهمية البحث وأهدافه

تأتي أهمية البحث كونه يقدم طريقة لتوليد تسلسلات DNA من تسلسلات حقيقية وذلك لاستخدام هذه التسلسلات لأغراض التشفير والفهرسة التي تم اعتمادها لتشفير النصوص باستخدام مفاتيح OTP من تسلسلات DNA مولدة عشوائياً [7] أيضا اعتمدت التسلسلات العشوائية لتشفير الصور [8]، وكذلك استخدمت تسلسلات DNA العشوائية لتشفير الصور الملونة عن طريق الفهرسة لتسلسلات DNA [9].

يهدف البحث إلى بناء نموذج توليدي يمكن استخدامه لتوليد تسلسلات DNA عشوائية وذلك بالاعتماد على شبكات التعلم العميق وبالتحديد شبكات LSTM كونها ملائمة للبيانات التسلسلية [23]، كما أنها تقدم الحل لمشكلة تلاشي التدرج الذي تعاني منه شبكات RNN والتي استخدمت على نطاق واسع في توليد النصوص.

4. منهجية البحث

استخدمت الدراسات المرجعية شبكات التعلم العميق LSTM لتوليد النصوص. تعتمد هذه الدراسة على استخدام LSTM لتوليد تسلسلات DNA عشوائية بناءً على مجموعة بيانات مأخوذة من قاعدة بيانات جينية عامة، هذه البيانات هي عبارة عن تسلسلات DNA حقيقية قد تم اختبار عشوائيتها من قبل [7] حيث بينت الدراسة بأنه ليس كل تسلسلات DNA الطبيعية عشوائية، وبرهنوا على ذلك من خلال اختبارات NIST ووصلوا إلى نتيجة مفادها أن 63% من تسلسلات DNA التي تم أخذها من قواعد البيانات الجينية العامة هي عشوائية والتي ستعتمد كبيانات لتدريب النموذج

الذي سيتم اعتماده في هذا البحث بغية الحصول على تسلسلات DNA عشوائية، كما سيتم تجريب هذه التسلسلات على تشفير النصوص والصور حسب الدراسات [7] و [9] وإجراء اختبارات امنية على ناتج التشفير.

5. الأدوات

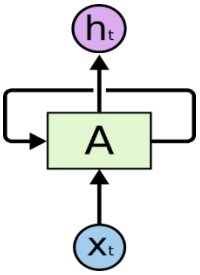
-تسلسلات DNA عشوائية مأخوذة من قواعد البيانات الجينية العامة والتي هي للجينوم البشري ممثلاً بالصبغيات من 1 إلى y . تم الحصول على التسلسلات من قاعدة بيانات جينية عامة من الموقع المجاني:

- https://www.ncbi.nlm.nih.gov
-شبكات LSTM.
-لغة برمجة Python.3.7 مع بيئة تطوير Pycharm.2019.3.1 (Community Edition)
-مكتبات python مثل Keras و Tensorflow و OpenCV و Matplotlib و numpy
-جهاز حاسب لابتوب HP بذواكر 8GB ومعالج i7-6500U@2.5Ghz(4CPUs) intel core(TM)

6. شبكات التعلم العميق المستخدمة مع البيانات النصية

1.6. الشبكات العصبونية العودية RNN

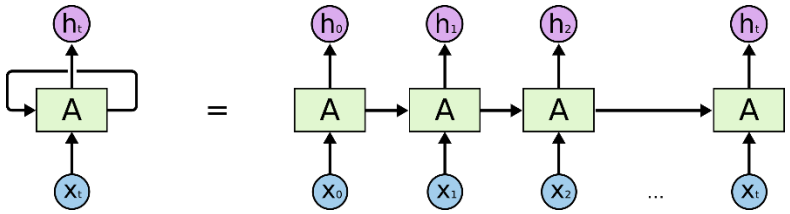
إنها شبكات تملك حلقات عودية بداخلها بحيث تمكن المعلومات من الديمومة. تقوم شبكات RNN بمعالجة التسلسلات من خلال التكرار عبر عناصر التسلسل وتحتفظ بالحالة التي تحوي معلومات عما تم رؤيته. شبكات RNN هي نوع من الشبكات العصبونية التي تملك حلقة داخلية الشكل (3). يتم إعادة وضع حالة RNN بين معالجة تسلسلين مختلفين مستقلين، حيث يمكن النظر الى أحد التسلسلات كنقطة بيانات أو كدخل مفرد الى الشبكة. والذي يتغير هو أن نقطة البيانات هذه لا تتم معالجتها في خطوة مفردة واحدة وبدلا من ذلك، تعمل الحلقة الداخلية للشبكة على عناصر التسلسل.



الشكل 3: شبكة RNN عودية مع حلقة

يمثل الشكل (3)، قطعة من الشبكة العصبونية والتي تبدو عند دخل x_t وتعطي قيمة على الخرج h_t. تسمح الحلقة للمعلومات بالمرور من خطوة واحدة للشبكة الى التالية. يمكن التفكير بشبكة RNN

وكأنها نسخ متعددة من نفس الشبكة، كل واحدة تقوم بتمرير رسالة الى التالية. والشكل التالي يوضح فيما لو قمنا بفرد أو نشر الحلقة الشكل (4).



الشكل 4: شبكة RNN عودية مع حلقة تم نشرها

هذه الطبيعة الشبيهة بالسلسلة تظهر بأن شبكات RNN على علاقة وثيقة بالتسلسلات والمصفوفات أحادية البعد أو القوائم. يظهر الشكل (4) أن الشبكة العودية الموضحة في الشكل (3) تكافئ سلسلة من الوحدات على التتالي، حيث خرج كل وحدة يكون دخل للوحدة التالية. إنها المعمارية الطبيعية من الشبكات العصبونية لتستخدم مع هذه البيانات وتم استخدامها بالتأكيد في السنوات القليلة الأخيرة، كان هناك نجاح منقطع النظير بتطبيق RNN على مسائل مختلفة كالتعرف على الكلام ونمذجة اللغات والترجمة وإضافة التعليقات على الصور [24]

1.1.6. مشكلة الاعتماديات طويلة الاجل

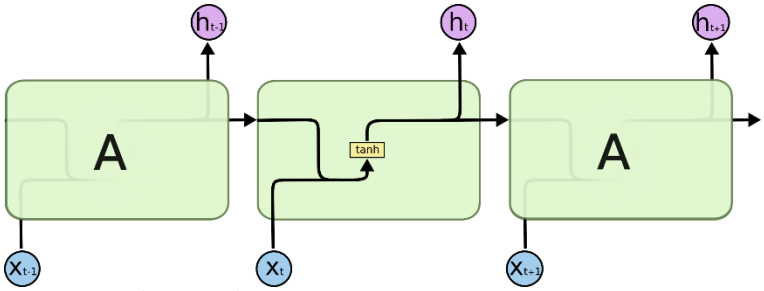
إحدى صفات شبكات RNN هي أنه يمكنها وصل المعلومات السابقة الى المهمة الحالية، كما في حالة استخدام إطارات الفيديو السابقة والذي يمكن أن يعطي معلومات لفهم الإطار الحالي. إن كان بإمكان RNN القيام بذلك عندها يمكن أن تكون فعالة الى حد بعيد لكن هذا لا يحدث. نحتاج فقط في بعض الأحيان الى النظر إلى المعلومات الأخيرة لإنجاز المهمة الحالية. كمثال نموذج لغة يحاول التنبؤ بالكلمة التالية بالاعتماد على الكلمات السابقة.

إذا كنا نحاول التنبؤ بالكلمة الأخيرة في عبارة " the planes are flying in the sky" فلا نحتاج الى سياق إضافي. من الواضح بأن الكلمة التالية ستكون "sky". في هذه الحالات، حيث تكون الفجوة صغيرة بين المعلومات ذات الصلة والمكان الذي نحتاجها فيها، عندها يمكن لشبكات RNN أن تتعلم كيف تستخدم المعلومات السابقة.

لكن هناك حالات أيضاً حيث نحتاج سياقات أكثر. كلما كبرت الفجوة، تصبح RNN غير قادرة على أن تتعلم كيف توصل المعلومات. يمكن لشبكات RNN نظرياً من التعامل مع "الإعتماديات طويلة الأجل" لكن في التطبيق العملي لا تبدو بأنها قادرة لتتعلمها. تم اكتشاف المشكلة من قبل [25] و [26] والذين وجدا بعض الأسباب الأساسية لماذا تكون صعبة. لكن هناك حالات أيضاً حيث نحتاج سياقات أكثر. كلما كبرت الفجوة، تصبح RNN غير قادرة على أن تتعلم كيف توصل المعلومات. فمثلا عندما نريد التنبؤ بالكلمة الأخيرة في النص "I grew up in **England**... I speak fluent **English**". تقترح المعلومات الأخيرة بأن احتمال أن تكون الكلمة التالية هي اسم اللغة، لكن إن كنا نريد حصر أية لغة، نحتاج سياق للبلد England. من الممكن للفجوة بين المعلومات ذات الصلة والنقطة حيث نحتاج هذه المعلومات أن تكون كبيرة جداً.

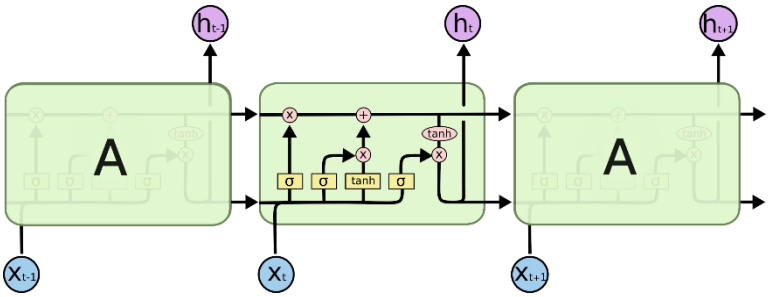
2.6. شبكات LSTM

شبكات الذواكر طويلة قصيرة الأجل هي نوع خاص من شبكات RNN قادر على تعلم الاعتماديات طويلة الأجل. تم تصميم شبكات LSTM لتتغلب على مشكلة الاعتماديات طويلة الأجل. إن تذكر المعلومات لفترات طويلة من الزمن هو التصرف الافتراضي لها وليس محاولة التعلم. كل الشبكات العودية لديها نفس شكل السلسلة من تكرار الوحدات للشبكات العصبونية. في شبكات RNN القياسية، هذا النموذج العودي سيكون له بنية بسيطة جداً مثل طبقة tanh مفردة الشكل (5).



الشكل 5: النموذج العودي في شبكات RNN القياسية يملك طبقة مفردة

تملك شبكات LSTM البنية الشبيهة بالسلسة، لكن النموذج المتكرر له بنية مختلفة. بدلاً من امتلاك طبقة شبكة عصبونية واحدة، هناك أربعة تعمل مع بعضها بطريقة خاصة الشكل (6).



الشكل 6: النموذج العودي في LSTM يحتوي أربع طبقات متفاعلة

الفكرة الأساسية بالنسبة لشبكات LSTM هي حالة الخلية، حالة الخلية هي حالة مشابهة لحالة الحزام الناقل حيث تسير مباشرة عبر كامل السلسلة، مع تفاعلات خطية بسيطة. تملك شبكة LSTM القدرة على حذف أو إضافة المعلومات الى خلية الحالة بشكل منظم بعناية من قبل بنى تسمى البوابات. البوابات هي طرق تسمح للمعلومات بالمرور بشكل اختياري. تتألف البوابات من طبقة شبكة عصبونية من sigmoid وعملية ضرب نقطية. تعطي طبقة sigmoid على خرجها أرقام ما بين 0 و 1 والتي تصف مدى السماحية لكل مركب بالعبور من خلالها. القيمة 0 تعني "عدم السماح لأي شيء بالعبور" بينما القيمة 1 تعني "السماح لكل شيء بالعبور". تملك شبكات LSTM ثلاثة من هذه البوابات للحماية والتحكم بحالة البوابة.

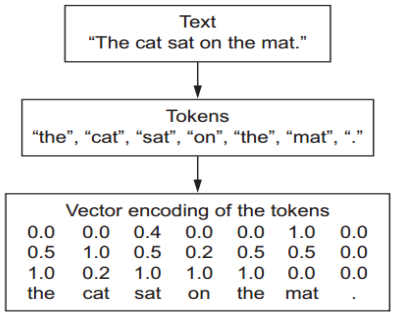
7. طريقة العمل

تم استخدام مجموعة بيانات لتسلسلات الجينوم البشري مأخوذة من قاعدة بيانات جينية عامة [10] والتي تكون موجودة على الموقع على شكل ملفات مضغوطة تم فك ضغط الملفات والحصول على التسلسلات ضمن ملفات نصية. تم اختيار مناطق من التسلسلات والتي تم أخذها حسب الدراسة [7]. استخدم ما مجموعه 80000 تسلسل كل منها بطول 512 نيكليوتيد والتي شكلت مجموعة البيانات، الشكل (7).

```
GGTGGCCGAATCGGCATAGAGGATCGATCAAATTTGCCCCGGCTACCCACACCAGAACTTTCAATTACCTAGCGGGCGCAA
TCTAAAACGGGGTCTAGGTTGCCCCCGGGACCCCTGTAATAAGGTGATTTTCGAGTTTAGCTATGCTTATAGGACTTACCAGTCATGT
TGGAGATGTTGTCCGATGGCACATCCCCTCGGTTCTCTTGCCAGGTCGGGGTTCTTACTCGCTCGCCGAATATTTCGGGTCGGGACC
GAACCCGTGAGCTTATGACGTATGCTACTCCAATGTGTATAAATCATCACCCGATGGGCGCTATTATTGCAGAGGGAATTCTCCAGCC
GCCGGCTCACGCAACTCTCGAGTCTGTCATAAACCCATTCCCAACTCAGTTACCTTCATTTATATTAGCTGAAGCTTTGACTAGATATC
ATTTACGGTCACTATGGTCGTCGAACGATATCCTGGTCGTCAACAGTTCACCCTCCGTAAACTATGTTCGATAAGTGCC
```

الشكل 7: عينة من التسلسلات العشوائية المستخدمة

يتم تحضير البيانات على شكل تسلسلات من النيكليوتيدات، والتي هي من أربع أحرف (A ,C,G,T) وتعتبر الأبجدية التي تكون كامل تسلسلات DNA. لا تستطيع شبكات التعلم العميق التعامل مع البيانات النصية مباشرة لذلك يجب تحويل البيانات النصية إلى أشعة من القيم الرقمية عبر عملية تسمى Vectorizing أي تحويل القيم النصية إلى قيم عددية أو تحويل أشعة النص سواء كلمات أم محارف إلى مصفوفات عددية tensors. أكثر الطرق المعتادة لتحويل الكلمات أو المحارف إلى أشعة من القيم العددية هي طريقة one-hot-vector، تجري عملية تفكيك النص إلى كلمات أو أحرف ثم إعطاء كل كلمة رمز أو token. وتسمى هذه العملية الترميز أو tokenization كما يوضح (الشكل 8). ثم تحويل كل علامة أو رمز إلى شعاع من القيم العددية. تتلخص طريقة one-hot-encoding بإعطاء كل كلمة فهرس ذو قيمة صحيحة (عدد صحيح) ومن ثم تحويل هذا الفهرس i إلى الصيغة الثنائية بحجم N والذي يعبر عن عدد المفردات في النص أو التسلسل المعطى والذي يوضحه (الشكل 7)



الشكل 8-a: تحويل النص إلى رموز ثم إلى أشعة [23]

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
A	G	T	C

الشكل 8-b: One-hot-encoding [23]

باستخدام طريقة الترميز One-hot-encoding، كما يوضح الشكل (8)، يكون تسلسل DNA الأولي بعد ذلك مرمز في مصفوفة بت مع تمثيل كل حرف في التسلسل كشعاع ثنائي رباعي العناصر. لتبسيط الأمر، في تجربة المحاكاة التي تم إجراؤها، وبافتراض أن التسلسل هو تسلسل بالتنسيق الجيني، مما يعني أن كل تسلسل له أربعة أنواع من الأحرف: A، C، G، T. تم ترميز كل حرف رقميًا كأحد المصفوفات الأحادية الثنائية: A= [1,0,0,0] , C= [0,1,0,0], T=[0,0,0,1] G = [0,0,1,0] [11] ثم يمكن بعد ذلك تمثيل

كل تسلسل كمصفوفة بتات مرمزة 512X4، وذلك بفرض طول التسلسل 512 رمز أو نيكليوتيد. مع أعمدة مقابلة لكل من A وC وG و T. بهذه الطريقة، يمكن الحفاظ على معلومات المواقع الحيوية لكل حرف في التسلسل.

يتكون نموذج مولد تسلسلات الحمض النووي DNA من طبقتين من LSTM مع 128 خلية عصبية مخفية الشكل (10). يقرأ حرفًا واحدًا في كل مرة ويتوقع التالي في التسلسل. يحدد حجم الدُفعة (B) في نموذج LSTM، كيف تتم معالجة العديد من التسلسلات بالتوازي في وقت واحد. طول التسلسل (S) يحدد طول كل منها (S = 512 في مجموعة البيانات الخاصة). بفرض أن ملف الإدخال إلى النموذج يحتوي على تسلسل DNA من 512 نيوكليوتيد فيكون N = K × 512. بعد ذلك، يتم تقسيم ملف الإدخال المكون من N حرف إلى قطع بيانات بحجم B × 512. بشكل افتراضي، استخدم 80٪ من التسلسلات للتدريب 20٪ تستخدم لتقدير خسارة التحقق من الصحة. يتم تقسيم ملف الإدخال إلى قطع من البيانات وتغذيتها إلى طبقات LSTM بالإعدادات الافتراضية. تم استخدام القيمة الافتراضية 512 لحجم الدفعة (B).

تم تدريب نموذج LSTM حسب المعادلة (1). حيث x_t هو شعاع يمثل النيكليوتيد t في تسلسل الدخل. فقط عنصر واحد من x_t يأخذ القيمة 1، والبقية تأخذ القيمة 0. y_t هو مؤشر صف من n_t معرف بالمعادلة (1). يقوم نموذج LSTM بحساب z_t للشعاع x_t حسب المعادلة (2). يقوم Softmax بتغيير z_t إلى شعاع قيم محصورة بين 0 و1 ومجموعها 1، وsoftmaxj هو العنصر j لخرج Softmax المعادلة (3). الخسارة loss هي المتوسط للوغاريتم السالب المحتمل للتنبؤ، المعادلة (4). تستخدم الخسارة لتحديث العصبونات في الطبقة المخفية باستخدام الخوارزمية RMSProp [27]. عند توليد التسلسل، يقوم النموذج بأخذ الشعاع (0.25، 0.25، 0.25، 0.25) ممثلا x_1 ويحسب Softmax(zt) ، والذي هو توزيع متعدد الأبعاد للنكليوتيدات. يتم أخذ محرف كعينة من التوزيع وشعاع المحرف يتم تغذيته ثانية إلى النموذج ممثلا x_2. يتم تكرار هذه العملية حتى الوصول إلى الطول المحدد مسبقاً للتسلسل.

$$y_t = \begin{cases} 1\ if\ n_t = A \\ 2\ if\ n_t = C \\ 3\ if\ n_t = G \\ 4\ if\ n_t = T \end{cases} \quad \text{حيث: } x_t = 4bit \text{ عدد يمثل النيكليوتيد و } n_t \in \{A, C, G, T\} \qquad (1)$$

$$z_t = LSTM(x_t) \qquad (2)$$

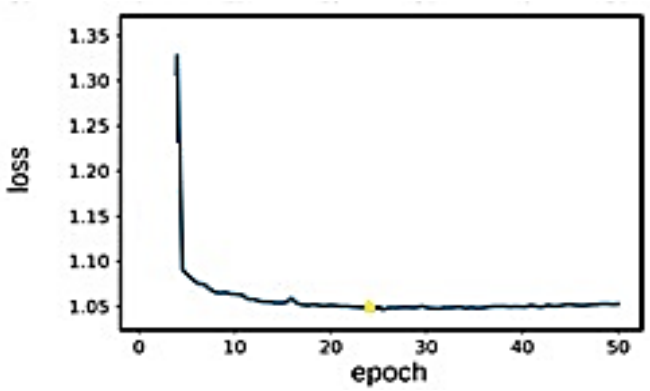
$$softmax_j(z_t) = \frac{e^{z_{tj}}}{\sum_{k=1}^{4} e^{z_{tk}}}\ , j \in \{1, 2, 3, 4\} \qquad (3)$$

$$loss = -\sum_{t=1}^{|x|} \ln\left(\frac{softmaxy_t(z_t)}{|x|}\right) \qquad (4)$$

بما أن تسلسلات DNA المستخدمة في تدريب النموذج كانت بطول 512 نيكليوتيد، فإن طول التسلسلات المولدة من قبل النموذج كانت أيضا بطول 512 نيكليوتيد. يمكن استخدام التسلسلات الناتجة

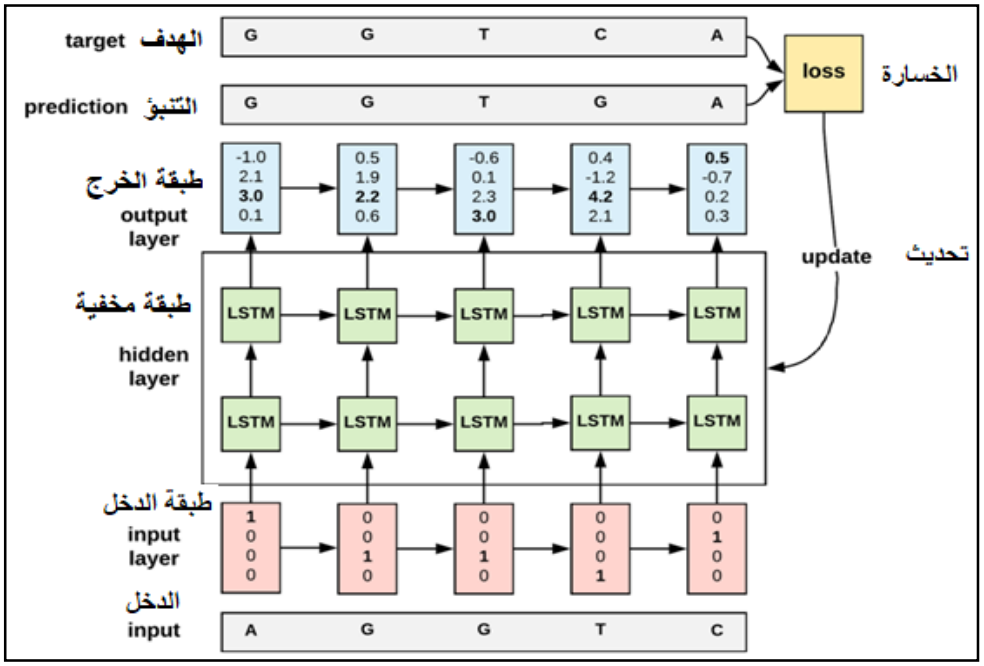
بطول 512 نيكليوتيد كمفاتيح تشفير، حيث أن كل نيكليوتيد يرمز برمزين ثنائيين حسب الجدول (1) وبالتالي يعطي مفتاح بطول 1024 بت. أما في حال استخدام التسلسلات لعملية الفهرسة وفي حال كان التسلسل المطلوب أكبر من 512 نيكليوتيد عندها يمكن أخذ عدة تسلسلات ودمجها معاً للحصول على الطول المناسب. عند تدريب النموذج، تم تقييم النتائج بالاعتماد على معيار الخسارة حسب المعادلة (4).

بعد تدريب النموذج على بيانات التدريب لخمسين تكرار 50 epochs، حيث بعد 50 تكرار وصل النموذج الى أقل خسارة. تم اختيار النموذج ذي قيمة الخسارة الأقل واستخدامه لتوليد تسلسلات DNA، حيث كانت أقل قيمة للخسارة بعد التكرار رقم 19 هي 0.95 كما يبين الشكل (9).



الشكل 9: منحني الخسارة Loss خلال 50 فترة Epochs من التدريب

النموذج الذي تم استخدامه للتدريب هو شبكة LSTM مع 128 عصبون و64 حجم الدفعة batch. تم اعتماد مفهوم درجة الحرارة Temperature وهو بارامتر يستخدم من أجل التحكم بكمية العشوائية في عملية أخذ العينات، والذي يوصف عشوائية التوزيع الاحتمالي المستخدم لأخذ العينات وكم سيكون اختيار المحرف التالي قابلاً للتنبؤ أم غير متوقعاً. بإعطاء قيمة لدرجة الحرارة، سيتم حساب توزيع احتمالي جديد من التوزيع الأصلي (خرج Softmax للنموذج) [23]. للإشارة إلى الانتروبية أو العشوائية للتسلسلات الناتجة عن التوليد مع القيم [0.2, 0.5, 1, 1,2]، حيث يوضح الجدول (2) محتوى التسلسلات من كل رمز من رموز DNA أو النكليوتيدات كنسبة مئوية من الحجم الكلي للرموز (النيكليوتيدات) التي تم توليدها على شكل تسلسلات DNA عند قيم درجات حرارة مختلفة.



الشكل 10: بنية مولد التسلسلات. تستخدم الخسارة لتحديث العصبونات في الطبقة المخفية باستخدام خوارزمية RMSProp [28]

الجدول 2: محتوى التسلسل من الرموز A–C–G–T كنسب مئوية لكل رمز من التسلسل الكلي وذلك عند نسب مختلة للبارامتر Temperature

محتوى التسلسل من الرموز كنسب مئوية %				Temperature
T	G	C	A	
25.10	23.81	23.81	22.31	0.2
24.89	29.61	23.60	22.53	0.5
24.09	28.14	19.62	29.45	1
24.60	26.82	23.42	23.81	1.2

من الجدول (2) تبين قيمة Temperature 1.2 أنها أقرب إلى التوزيع المتساوي بنسبة 25% لكل من النكليوتيدات الأربعة وهذا ما يجعل التسلسلات الناتجة تقترب من العشوائية حسب [7].

المراجع

[1] A. ASISH, S. RANJAN, D. SATYA and D. SATCHIDANANDA, "A Symmetric Key Cryptosystem Using DNA Sequence with OTP Key," *Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing 340 Springer India,* p. 207 – 215, 2015.

[2] Z. YUNPENG, L. XIN and S. MANHUI, "DNA based Random Key Generation and Management for OTP Encryption," *Biosystems, BIO 3753,* pp. 203 - 215, 2017.

[3] K. SHRUTI, K. HARLEEN and C. VICTOR, "DNA Cryptography and Deep Learning using Genetic Algorithm with NW algorithm for Key Generation," *Springer Science +Business Media,* vol. 45, no. 17, pp. 1-12, 2018.

[4] S. Stalin, P. Maheshwary, M. M. PK Shukla, B. Gour and A. Khare, "Fast and Secure Medical Image Encryption Based on Non Linear 4D Logistic Map and DNA Sequences (NL4DLM_DNA)," *Journal of Medical Systems,* vol. 43, pp. 1-17, 2019.

[5] X. Zhang, F. Han and a. Y. Niu, "Chaotic Image Encryption Algorithm Based on Bit Permutation and Dynamic DNA Encoding," *Hindawi,* pp. 1- 11, 2017.

[6] R. Guesmi, M. Farah, A. Kachouri and M. Samet, "A novel chaos-based image encryption using DNA sequence operation and Secure Hash Algorithm SHA-2," *Springer Science+Business Media Dordrecht* , pp. 1 -14, 2015.

[7] K. A. Kassem and T. Salman, "Text encryption using OTP keys from randomly generated DNA," *Tishreen University Journal for research and scientific studies,* vol. 41, 2019.

[8] K. A. Kassem and T. Salman, "Securing color image based on DNA encoding and chaos," *International Journal of Computer Science Trends and Technology (IJCST),* vol. 8, 3 2020.

[9] K. A. Kassem and T. Salman, "Encrypting colored images using DNA Sequence indexing," *Tishreen University Journal for research and scientific studies,* vol. 42, 12 2020.

[10] ncbi.nlm.nih.gov, "http:\\www.ncbi.nlm.nih.gov," [Online]. Available: http:\\www.ncbi.nlm.nih.gov. [Accessed 4 2021].

[11] T. Anwar, A. Kumar and S. Paul, "DNA Cryptography Based on Symmetric Key Exchange," *International Journal of Engineering and Technology (IJET),* vol. 7, no. 3, pp. 938 - 950, Vol 7 Jun-Jul 2015.

[12] D. Kingma and M. Welling, "Autoencoding variational bayes," *CoRRabs,* 2013.

[13] S. Indurthi, D. Raghu and M. Khapra, "Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model," in *EACL*, 2017.

[14] I. Sutskever, J. Martens and G. Hinton, "Generating Text with Recurrent Neural Networks," in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, 2011.

[15] S. Zhou, "Research on the Application of Deep Learning in Text Generation," *Journal of Physics,* pp. 1-7, 2020.

[16] H. Shahidi, M. Li and J. Lin, "Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data," pp. 2-7, May 2020.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput,* vol. 9, p. 1735–1780, 1997.

[18] A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2013.

[19] I. Sutskever, O. Vinyals and Q. Le, "Sequence to Sequence Learning with Neural Networks," *CoRR*, 2014.

[20] S. Santhanam, "CONTEXT BASED T EXT- GENERATION USING LSTM NETWORKS," pp. 1-10, 30 April 2020.

[21] M. Zhang, "Learning the Language of the Genome using RNNs," 2014.

[22] H. Zhang, C.-L. Hung, M. Liu, X. Hu and Y.-Y. Lin, "NCNet: Deep Learning Network Models for Predicting Function of Non-coding DNA," *Front. Genet*, vol. 432, 2019.

[23] F. Chollet, Deep Learning with Python, Shelter Island: MANNING, 2018.

[24] A. Karapathy, "github.com," 2016. [Online]. Available: https://github.com/karapathy/char-rnn. [Accessed June 2021].

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 8, pp. 1735-1780, 1997.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[27] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude.," 4 2012. [Online].

[28] G. H. T Tieleman, *rmsprop: divide the gradient by a running average of its recent magnitude*, 4 ed., COURSERA, 2012, p. 26–30.

[29] NIST, "http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-22r1a.pdf," 2010. [Online]. Available: http://www.nvlpubs.nist.gov. [Accessed 20 3 2021].

[30] B. Devi and K. Kumar, "A NOVEL TEXT ENCRYPTION ALGORITHM USING DNA ASCII TABLE WITH A SPIRAL APPROACH," *International Journal of Recent Scientific Research*, vol. 1, pp. pp. 23588-23595, 2018.

[31] C. Song and Y. Qiao, "A Novel Image Encryption Algorithm Based on DNA Encoding and Spatiotemporal Chaos," *Entropy*, no. 17, pp. 6954-6968, 2015.

[32] B. Norouzi, S. Seyedzadeh, S. Mirzakuchaki and M. Mosavi, "A novel image encryption based on hash function with only two-round diffusion process," *Multimedia Syst*, vol. 20, no. 1, p. 45–64, 2013.

[33] R. Enayatifar, H. Sadaei, A. Abdullah, M. Lee and I. F. Isnin, "A novel chaotic based image encryption using a hybrid model of deoxyribonucleic acid and cellular automata," *Opt. Lasers Eng*, vol. 71, p. 33–41, 2015.

[34] X. L. J Wu and B. Yang, "Color image encryption based on chaotic systems and elliptic curve ElGamal scheme," *Signal Process*, vol. 141, p. 109–124, 2017.

[35] J. Wang, "A Color Image Encryption Using Dynamic DNA and 4-D Memristive Hyper-chaos," *IEEE Access*, vol. 7, pp. 78367- 78379 , 2019.

[36] X. Wang, Y. Wang, X. Zhu and C. Luo, "A novel chaotic algorithm for image encryption utilizing one-time pad based on pixel level and DNA level," *Optics and Lasers in Engineering*, vol. 125, pp. 1 -12, 2020.

[37] C. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 3, p. 379–423, 1948.

[38] A. Jain and N. Rajpal, "A robust image encryption algorithm resistant to attacks using DNA and chaotic logistic maps," *Multimed Tools Appl* , vol. 75, pp. 5455-5473 , 2016.

[39] S. Agarwal, "A Chaotic Cryptosystem using Conjugate Transcendental Fractal Function," *I. J. Computer Network and Information Security*, vol. 2, no. 1, pp. 1-12, 2019.

[40] B. Norouzi, S. Mirzakuchaki, S. M. Seyedzadeh and M. R. Mosavi, "A simple, sensitive and secure image encryption algorithm based on hyper-chaotic system with only one round diffusion process," *Multimedia tools and Applications,* no. 71, p. 1469–1497, 2014.

[41] S. Stalin, P. Maheshwary, P. K. Shukla, M. Maheshwari, B. Gour and A. Khare, "Fast and Secure Medical Image Encryption Based on Non Linear 4D Logistic Map and DNA Sequences (NL4DLM_DNA)," *Journal of Medical Systems* , vol. 267, no. 43, pp. 1-17, 2019.

[42] C. T. Zhang, "Research on Image Encryption Based on DNA Sequence and Chaos Theory," in *Phys. Conf. Ser 1004 012023*, 2018.

[43] Y. Wang, K. Wong, X. Liao, T. Xiang and G. Chen, "A chaos-based image encryption algorithm with variable control parameters," *Chaos Solitons Fractals,* vol. 41, no. 4, pp. 1773-83, 2009.

[44] X. LI, C. Zhou and N. Xu, "A Secure and Efficient Image Encryption Algorithm Based on DNA Coding and Spatiotemporal Chaos," *International Journal of Network Security,* vol. 20, no. 1, pp. 110-120, 2018.

[45] T. Li, M. Yang, J. Wu and X. Jing, "A Novel Image Encryption Algorithm Based on a Fractional-Order Hyperchaotic System and DNA Computing," *Hindawi,* p. 13, 2017.

[46] K. Zhan, D. Wei, J. Shi and J. Yu, "Cross-utilizing hyperchaotic and DNA sequences for image encryption," *Journal of Electronic Imaging, vol. 26, no. 1, Article ID 013021, 2017.*, vol. 26, no. 1, p. 13, 2017.

[47] R. Speer, J. Chin and C. Havasi, "An open multilingual graph of general knowledge," in *Proceedings of AAAI*, Francisco, 2017.

[48] B. Peng and K. Yao, "Recurrent nural networks with external memory for language understanding," 2015.

[49] E. Dairai, "Deep Learning for NLP: An Overview of Recent Trends," 2018. [Online]. Available: http://medium.com/dair-ai/deep-learning-for-nlp-an-overview-ofrecent-trendsd0d8f40a776d/2018. [Accessed 04 April 2021].

[50] J. Elman, "Finding structure in time," *Cognitive Sci*, vol. 14, no. 2, p. 179–211, 1990.