



المجلة العراقية للعلوم الإحصائية

<http://stats.uomosul.edu.iq>



مقارنة بين أساليب الانحدار اللوجستي والشبكة العصبية الالتفافية و Kernel لتصنيف حركة الربداء الرشيقية

عمر أكرم محمد سعيد  أسامة بشير شكر 

قسم الاحصاء والمعلوماتية ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

الخلاصة

ان بيانات السلاسل الزمنية تستخدم على نطاق واسع في العديد من المجالات ومنها بيانات الاحياء المجهرية. ومن الضروري معرفة كيفية تصنيف الفئة التي تنتمي اليها كل مشاهدة وذلك باستخدام أساليب تصنيف إحصائية وخوارزميات التعلم الآلي والتعلم العميق. تعتبر دراسة حركة بعض أنواع الديدان الاسطوانية كأحد أنواع الأحياء المجهرية ومنها الربداء الرشيقية (CE) *Caenorhabditis Elegans* من الأمور المهمة لتحديد الأفعال وتأثيرها على حياة الديدان. في هذه الدراسة تمثلت بيانات السلسلة الزمنية لحركة CE بزوايا حركتها الموجية التي ستكون حالة الدراسة. في هذا النوع من البيانات فإن صفة اللاخطية وكذلك عدم اليقين تعد من أكثر المشاكل التي قد تؤدي الى تصنيفات لا ترقى الى ان تكون دقيقة. سيتم استخدام الشبكة العصبية الالتفافية (CNN) Convolutional Neural Network كأحد تقنيات التعلم العميق ويعتبر أسلوب غير خطي يستخدم لتصنيف حركة CE كمتغير معتمد في الحالات الثنائية بناء على صور زوايا الحركة الموجية كمتغير مستقل وان استخدامها سيؤدي الى نتائج دقيقة وذلك لانها أسلوب غير خطي ملائم للتعامل مع بيانات الدراسة لحل مشكلات عدم الخطية وكذلك فانه ملائم للتعامل مع مشكلة عدم اليقين من خلال التمثيل الصوري للبيانات الرقمية. كذلك تم استخدام الانحدار اللوجستي (LR) Logistic Regression وطريقة النواة Kernel أيضا لتصنيف زوايا حركة CE. وتمت الاستفادة من الانحدار الذاتي (AR) Auto Regressive بالاعتماد على رتبة نموذج الانحدار الذاتي (AR(p) في تحديد هيكليته الأساليب المستخدمة. ومن خلال مقارنة النتائج بين الأساليب المستخدمة يتبين ان أسلوب CNN يتفوق على الأساليب الأخرى المستخدمة باستخدام مقاييس الدقة للتصنيف. ولذلك فمن الممكن استنتاج ان استخدام أسلوب CNN والذي يعتمد على التصنيف الصوري يؤدي الى نتائج تصنيفية دقيقة مقارنة بالأساليب الأخرى المعتمدة على التصنيف الرقمي.

معلومات النشر

تاريخ المقالة:

تاريخ الاستلام: 1 اب 2023

تاريخ القبول: 18 تشرين الثاني 2023

تاريخ القبول: 28 تشرين الثاني 2023

متاح على الإنترنت: 1 كانون الاول 2023

الكلمات الدالة:

الشبكة العصبية الالتفافية، طريقة النواة

التصنيف، السلاسل الزمنية

الانحدار الذاتي، الربداء الرشيقية

المراسلة:

أسامة الحنون

drosamahannon@uomosul.edu.iq

q

DOI: [10.3389/IQJOSS.2023.0181225](https://doi.org/10.3389/IQJOSS.2023.0181225) , ©Authors, 2023, College of Computer Science and Mathematics, University of Mosul, Iraq.

This is an open access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. مقدمة Introduction

في هذه الدراسة تم التطرق الى دراسة تصنيف السلسلة الزمنية باستخدام خوارزميات التعلم تحت الاشراف supervised learning algorithms والتي تشترط توفر بيانات المتغير المعتمد (متغير الهدف Target variable) للحصول على أخطاء التعلم. تعتبر دراسة الديدان الاسطوانية الشفافة بصورة عامة

من الدراسات المهمة في علم الأحياء المجهرية كون خلاياها تشبه خلايا الإنسان وكذلك بسبب سرعة مراحل نموها. ان حركة الدودة تكون بصورة متتابعة خلال زمن معين أي على شكل سلسلة زمنية. إذ أن كل حركة تقوم بها الدودة أو تصل إليها تكون مرتبطة بالحركة التي قبلها لتمثل حالة الدراسة لمتغير سلسلة زمنية واحد لحركة الدودة وتكون خلال فترة زمنية لأجزاء من الثانية لكل حركة تقوم بها الدودة. في هذه الدراسة تم الحصول على بيانات عن الريداء الرشيق (CE) *Caenorhabditis Elegans* ككائن حي نموذجي مهم في دراسة علم الوراثة لزيادة فهم علم الوراثة السلوكية. البيانات تضم مشاهدات كل منها تعبر عن زاوية خاصة بحركة الريداء الرشيق 1 *Caenorhabditis Elegans Motion (CEM)*. ولأنه لا يمكن تحديد CEM تحديداً نقطياً ولكن يمكن حصره ضمن فترات فيكون من الضروري توصيفه ضمن فئات وهذا ما قد يعالج جزءاً من مشكلة عدم اليقين وعندها يتم تركيز الاهتمام بتصنيف المشاهدات الجديدة من خلال نموذج تصنيفي يتم بناؤه من خلال سلوك السلسلة الزمنية خلال فترة التدريب Training period. تعد بيانات CEM من السلاسل الزمنية الطويلة بعدد مشاهدات كبير جداً وبدليل زمني قد يكون بالثنائي أو أجزاء الثنائي مما قد يضيف عليها صفة عدم الخطية ومما قد يجعل من الصعب التعامل مع مثل هذه البيانات. وكذلك فإن طول السلسلة الزمنية الكبير جداً مما قد يكون سبباً رئيسياً لعدم التجانس والنتائج عن تعدد الخصائص والصفات والمركبات التي تمر بها البيانات من بداية السلسلة الى نهايتها والذي قد يجعل من التنبؤ النقطي point Forecast ذو نتائج غير دقيقة. وللتقليل من مشكلة عدم الخطية وعدم التجانس في البيانات وتحسين نتائج التنبؤ فيمكن تمثيل البيانات بالصور واستخدام التصنيف الثنائي Binary Classification بدلاً من التنبؤ النقطي لتحسين دقة النتائج مقارنة بنتائج التنبؤ بوجود عدم التجانس وعدم خطية البيانات لأنه يعالج بشكل غير مباشر مشكلة عدم اليقين.

هناك دراسات سابقة تناولت استخدام أسلوب الشبكة العصبية الالتفافية Convolutional Neural Network (CNN) والانحدار اللوجستي Logistic Regression (LR) و Kernel في مجال الأحياء المجهرية فيما يخص سلوك الديدان الاسطوانية. إذ قام الباحث (1) باستخدام أسلوب CNN لتمييز سلالات CE المتنوعة وراثياً وتصنيفها عن طريق تدريب النموذج على بيانات سلسلة زمنية لاضاع الدودة باستخدام عينات من صور حركة الدودة كمتغير ادخال وتم الحصول على نموذج قادر على تصنيف السلالات. كما استخدم (2) أسلوب CNN لتصنيف الصور المجهرية لتحديد نوع معين من الديدان الخيطية وكان النموذج المدرب يعمل بشكل جيد في التصنيف. وقام كذلك (3) باستخدام التعلم العميق من خلال اسلوبي CNN و Recurrent neural network (RNN) لحساب متوسط العمر المتوقع لبيانات CEM عن طريق تصنيفها انها على قيد الحياة او انها ميتة من خلال ملاحظة صور من حركة الدودة وحققت الطريقة المقترحة معدلات خطأ صغيرة مما يدل على جودتها. وقام (4) بتحديد اذا ما كان العمر والنمط الجيني لدودة الريداء الرشيق يؤثران على حركتها في جميع السلالات باستخدام نموذج الانحدار اللوجستي الثنائي إذ تم تقييم فروق الحركة بين السلالات من النوع البري من خلال الانحدار اللوجستي. وكان نموذج الانحدار اللوجستي مناسب تماماً إذ قام (5) بتوقع الحركة الامامية لدودة الريداء الرشيق باستخدام أسلوب Logistic Regression (LR) الانحدار اللوجستي من خلال التنبؤ بالخلايا العصبية المشاركة في سلوك الحركة. كانت نتائج التنبؤ فعالة باستخدام طريقة النواة وبمعدل خطأ قليل من خلال قيام الباحث (6) استخدام أسلوب النواة Kernel Method بالتنبؤ بعمل الجينات لدودة الريداء الرشيق للمساعدة في تمييز النمط الصحيح والتحقق من الإيجابيات والسلبيات.

2. المواد والطرق

2.1. نموذج الانحدار الذاتي (AR) Auto Regressive Model (AR)

إن السلاسل الزمنية Time Series هي مجموعة من المشاهدات التي تتولد بفترات زمنية متتابعة وتتميز بعدم الاستقلالية. إذ أن المشاهدات فيها ترتبط بسابقتها زمنياً إذ يمكن من خلالها التنبؤ بالسلاسل الزمنية المستقبلية يعتمد على مشاهدات لسلسلة زمنية وقعت في الماضي (7, 8). ان السلسلة الزمنية الحالية يمكن أن يعبر عنها باستخدام دالة الانحدار الذاتي لقيم السلاسل الزمنية السابقة ويمكن كتابة دالة الانحدار الذاتي من الرتبة p كما في المعادلة (1) ادناه.

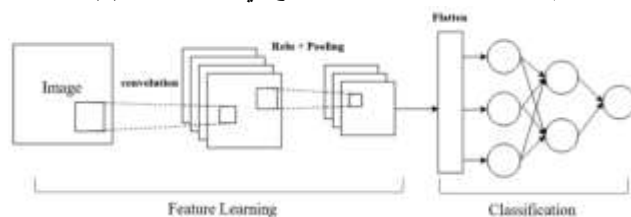
$$\begin{aligned}\phi(B)x_t &= e_t \\ \rightarrow (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)x_t &= e_t \\ \rightarrow x_t &= \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + e_t\end{aligned}\quad (1)$$

¹ أرشيف تصنيف السلاسل الزمنية UEA&UCR المتاح لجمهور الباحثين (UEA&UCR Time Series Classification Repository): <http://www.timeseriesclassification.com/description.php?Dataset=EigenWorms>

إذ أن ϕ_r هي معاملات الانحدار الذاتي و $r = 1, 2, 3, \dots, p$ وأن e_t الخطأ أو التغير العشوائي عبارة عن عملية تشويش أبيض بمتوسط صفر وتباين ثابت σ_e^2 إذ أن $e_t \sim i.i.d.N(0, \sigma_e^2)$ و $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)$. في تحليل بيانات السلسلة الزمنية يتم استخدام منهجية بوكس جنكنز بخطواتها الأربعة التعرف أو التحديد وتقدير المعلمات وإجراء الفحوص التشخيصية والتنبؤ. إن تحديد نموذج السلسلة الزمنية ومنها نموذج الانحدار الذاتي AR يتم غالبا من خلال الرسم باستخدام دالة الارتباط الذاتي Autocorrelation Function (ACF) ودالة الارتباط الذاتي الجزئي Partial Autocorrelation Function (PACF).

2.2. الشبكة العصبية الالتفافية (CNN) Convolutional Neural Network

هي أحد الأدوات الأساسية للتعلم العميق والتي تنضوي تحت مظلة الشبكات العصبية العميقة Deep neural network (DNN) التي تضم نوعا آخر من DNN وهي الشبكات العصبية المعادة Recurrent Neural Network (RNN). إن هيكالية CNN تتألف من جزأين أساسيين وهما طبقة التعرف على الميزات والتي تتم فيه عمليات الالتفاف والتجميع للتعرف على خصائص الصورة مثل الحواف وتدرج اللون وكذلك الطبقة المتصلة بالكامل التي تستقبل مخرجات طبقة التعرف على الميزات كمدخلات لينتج عنها عملية التصنيف كما موضح في شكل 1 أدناه (9).



شكل 1: هيكالية الشبكة العصبية الالتفافية.

يتم اختيار هيكالية CNN المناسبة عن طريق تحديد عدد الطبقات وهي طبقة الإدخال والطبقات الخفية وطبقة الإخراج. وبعد ذلك يتم تحديد حجم المرشح ويرمز له (f) وعدد المرشحات ويرمز لها (m) وأن المرشح عبارة عن أرقام عشوائية صغيرة تسمى الأوزان ويرمز لها (w) والتي تقوم بالالتفاف حول جميع نقاط الصورة للتعرف على التفاصيل التي تشكل الصورة. وعملية الالتفاف تكون بخطوات محددة تسمى stride باتجاه يسار الصورة. عند التفاف المرشح حول الصورة سوف يتقلص حجم الصورة ويتم فقدان العديد من البيانات والخصائص ولحل هذه المشكلة يتم إضافة الحشو Padding بصفوف وأعمدة على طول وعرض الصورة. إن حجم الإخراج الناتج من عملية الالتفاف حول الصورة يتم الحصول عليه عن طريق عملية رياضية مبسطة كما موضح في المعادلة (2) أدناه.

$$I \times J = \left[\left(\frac{H + 2p - f}{s} \right) + 1 \right] \times \left[\left(\frac{W + 2p - f}{s} \right) + 1 \right] \quad (2)$$

إذ أن H تمثل طول الصورة، W تمثل عرض الصورة، p تمثل الحشو، f تمثل أحد أبعاد المرشح، s تمثل الخطوة. بعد ذلك يتم جمع كل قيمة تحيز مع كل عنصر من عناصر المصفوفة التي تقابلها للحصول على مخرجات عملية الالتفاف كما في المعادلة (3) أدناه التي يمكن تسمية مخرجاتها بمخرجات الدالة الجمعية والتي تمثل المرحلة الأولى من الطبقة الخفية.

$$SUM_{ijk} = \sum_{i=1}^{f_2} \sum_{j=1}^{f_1} w_{f_1 f_2} x_{HW} + b_k \quad (3)$$

حيث أن I, J تمثلان أبعاد أو عدد الصفوف وعدد الأعمدة في كل صورة على التوالي، وأن f_1, f_2 تمثلان عدد الصفوف وعدد الأعمدة في كل مرشح على التوالي، وأن $k = 1, 2, \dots, m$ والتي ترمز إلى تسلسل ناتج كل مرشح. يتم استخدام أحد دوال التحويل ضمن الطبقة الخفية على مخرجات عملية الالتفاف ومن أكثر دوال التحويل استخداما في الشبكة العصبية هي كما يلي:

1. الدالة اللوجستية (Logistic Sigmoid function):

$$f(SUM) = \frac{1}{1 + e^{-f(SUM)}} \quad (4)$$

إذ أن SUM تمثل مدخلات دالة التحويل التي تمثل مخرجات الدالة الجمعية وتولد مخرجاتها ضمن الحدود (0,1).
2. دالة tan (Tan sigmoid function):

$$f(\text{SUM}) = \frac{2}{1 + e^{-2(\text{SUM})}} - 1 \quad (5)$$

إذ أن SUM تمثل مدخلات دالة التحويل التي تمثل مخرجات الدالة الجمعية وتولد مخرجاتها ضمن الحدود (-1,1).
3. الدالة الخطية (Pure Line function):

$$f(\text{SUM}) = \text{SUM} \quad (6)$$

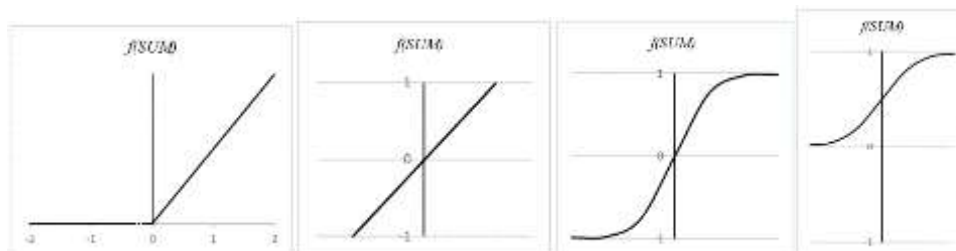
إذ أن SUM تمثل مدخلات دالة التحويل التي تمثل مخرجات الدالة الجمعية وتولد مخرجاتها ضمن الحدود (-1,1).
4. دالة الوحدة الخطية المصححة (Rectified Linear Unit function (Relu):

$$f(\text{SUM}) = \begin{cases} 0 & \text{SUM} \leq 0 \\ x & \text{SUM} > 0 \end{cases} \quad (7)$$

إذ أن x تمثل مدخلات دالة التحويل التي تمثل مخرجات الدالة الجمعية وتولد مخرجاتها أكبر أو يساوي (0).
5. دالة Softmax (Softmax function):

$$f(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^k \exp(x_j)} \quad (8)$$

إذ أن x_i تمثل مخرجات الطبقة المتصلة بالكامل، x_j تمثل فئات متغير الاستجابة، إذ أنه يتم تطبيق هذه الدالة على مخرجات الطبقة المتصلة بالكامل للحصول على نتائج بعدد فئات متغير الاستجابة وهي عبارة عن أرقام موجبة وكل رقم يمثل احتمالية انتمائه إلى فئة من الفئات ويكون مجموعهم يساوي واحد.



شكل 2: الدالة اللوجستية شكل 3: دالة tan شكل 4: الدالة الخطية شكل 5: دالة Relu

بعد عملية ادخال دالة التحويل على مخرجات الدالة الجمعية يتم تطبيق عملية التجميع pooling إذ أن المخرجات تقسم إلى عدد من المصفوفات الصغيرة ويتم ذلك بتحديد حجمها وعدد خطواتها. وكل مصفوفة صغيرة يتم اختزالها إلى قيمة واحدة فقط لتصبح المصفوفة في نهاية الطبقة الأولى الخفية بأبعاد أقل. ثم يتم تحويلها إلى متجه عن طريق عملية تسمى التسوية Flatten. إن المتجه يمثل مدخلات الطبقة المتصلة بالكامل Fully Connected Layer وتجمع هذه الطبقة كل الميزات التي تعلمتها الطبقات السابقة عبر الصور (10). يتم تحديد مخرجات الطبقة المتصلة بالكامل بعدد فئات التصنيف بالاعتماد على عدد فئات المتغير المعتمد ويتم أيضا تحديد الأوزان عشوائيا في كل عصبون وغالبا ما تتبع هذه الطبقة في حالة التصنيف بدالة التحويل Softmax. إن طبقة التصنيف Classification Layer يتم فيها تحديث قيم الوزن (w) والتهييز (b) وهي عملية تكرارية لحين الوصول إلى القيم المثلى (11).

2.3. الانحدار اللوجستي Logistic Regression

يعتبر الانحدار اللوجستي من أهم الخوارزميات المستخدمة في التصنيف ويستخدم عندما يكون المتغير المعتمد فئوي ثنائي ويسمى بالنموذج اللوجستي أو اللوجيت (Logit) وهو نموذج مفيد في تصنيف البيانات التي تتكون من متغير ثنائي للاستجابة ويعتمد على النموذج الخطي كما في المعادلات أدناه.

$$\ln \left[\frac{P(y=1|x_1, x_2, \dots, x_p)}{1-P(y=1|x_1, x_2, \dots, x_p)} \right] = \ln \left[\frac{\pi}{1-\pi} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta_0 + \sum_{j=1}^p \beta_j x_j \quad (9)$$

$$\frac{P(y=1|x_1, x_2, \dots, x_p)}{1-P(y=1|x_1, x_2, \dots, x_p)} = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p} = e^{\beta_0 + \sum_{j=1}^p \beta_j x_j} \quad (10)$$

إذ أن β_0 تمثل معلمة الحد الثابت، β_j تمثل معلمات الانحدار، x_j هي المتغيرات المستقلة، $\frac{\pi}{1-\pi}$ يمثل نسبة الترجيح (Odds) وهو الاحتمال الشرطي للنجاح مقسوم على الاحتمال الشرطي للفشل. وإن احتمالية وقوع الحدث يكون 1 إذا كانت قيمة الاحتمال أكبر أو يساوي 0.5 (12). إذ أن احتمال النجاح يمكن صياغته كما يلي.

$$\pi = P(y=1|x_1, x_2, \dots, x_p) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_j}} = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \quad (11)$$

2.4. طريقة النواة Kernel Method

تعتبر طريقة النواة من خوارزميات التعلم الآلي التي تستخدم في التصنيف الثنائي وخاصة في حالة البيانات غير الخطية وهي أكثر استخداماً على البيانات الضخمة التي تحتوي على مجموعات تدريب كبيرة وكذلك يمكن تطبيقها على مجموعات البيانات الأصغر ويستخدم أيضاً في الخوارزميات القائمة على النواة مثل الانحدار الخطي وخوارزمية آلة المتجه الداعم (Support Vector Machine (SVM). إن أساس عمل خوارزمية تصنيف النواة هو البحث عن أفضل طريقة لفصل البيانات إلى فئتين إذ تقوم بتحويل البيانات التي لا يمكن فصلها خطياً الموجودة في مساحة منخفضة الأبعاد إلى فضاء عالي الأبعاد ثم تعمل بملائمة نموذج خطي للبيانات في الفضاء عالي الأبعاد. هناك نقاط لا يمكن رؤيتها بشكل واضح أو الوصول إليها ولذلك فإن عملية فصل البيانات تمت في مساحة الإدخال الأصلية بصورة معقدة وغير خطية. في هذه الحالة يجب إعادة صياغة الزوايا والأطوال والمسافات إلى فضاء أو أبعاد أعلى لكي يتم فصل البيانات خطياً وبسهولة أكبر بواسطة النواة. إذ يتم فصل البيانات بشكل خطي وبسهولة بعد تحويل البيانات إلى فضاء عالي الأبعاد عن طريق دالة النواة وهي دالة تقوم بتحويل المتجهين (x_1 و x_2) إلى فضاء متجه جديد تعمل هذه الدالة كمقياس تشابه بتحديد التشابه أو الاختلاف بين كل نقطتين في الفضاء عالي الأبعاد وتعطي دالة النواة نتيجة 1 في حال التشابه الكبير بين النقاط وتعطي النتيجة صفر في حال الاختلاف الكبير بين النقاط. ولغرض حساب مدى اقتراب أو ابتعاد نقطة معينة عن عدد من النقاط لغرض التصنيف أو التنبؤ يتم هذا باستخدام دالة النواة (Gaussian Kernel) وكما في المعادلة أدناه (13).

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) = \exp\left(-\sum_{j=1}^n \frac{(x_i - x_j)^2}{2\sigma^2}\right) \quad (12)$$

إذ أن x_i و x_j تمثلان متجهين، وإن المقام في معادلة دالة النواة هو عبارة عن معلمة ضعف مربع σ وزياتها تؤدي إلى تقليل ما بداخل الدالة الأسية وهو ما يجعل النتيجة تقترب من 1 وإذا كانت σ قليلة مما يؤدي إلى زيادة ما بداخل الدالة الأسية فإن النتيجة تقترب من صفر.

مقاييس الدقة للتصنيف Classification Accuracy Measurement

تستخدم هذه المقاييس لقياس دقة أداء النموذج في التصنيف. وللتعرف على هذه المقاييس يجب معرفة مصفوفة الارتباك (Confusion Matrix (CM (14) كما موضح في المعادلة (14). ومن أبسط المقاييس المستخدمة في التصنيف مقياس الدقة التصنيفية Classification accuracy يتم فيه حساب نسبة الحالات المتوقعة المطابقة للحالات الفعلية إلى العدد الكلي لجميع الحالات المتوقعة والفعلية المطابقة وغير المطابقة وكما في المعادلة (13).

$$\text{Accuracy} = \frac{\text{Number of correctly classified}}{\text{Total Number}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (13)$$

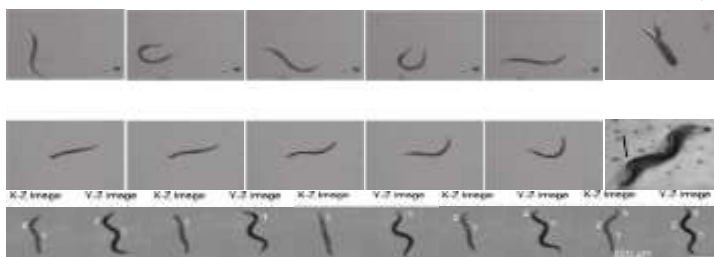
إذ أن:

$$CM = \frac{-1}{+1} \begin{bmatrix} \text{True Positive (TP)} & \text{False Positive (FP)} \\ \text{False Negative (FN)} & \text{True Negative (TN)} \end{bmatrix} \quad (14)$$

حيث ان (TP) تمثل عدد المشاهدات التي تم تصنيفها بشكل صحيح على انها ايجابية. وأن (TN) تمثل عدد المشاهدات التي تم تصنيفها بشكل صحيح على انها سلبية. وان (FP) تمثل عدد المشاهدات التي تم تصنيفها بشكل غير صحيح على انها ايجابية عندما كانت في الواقع سلبية. وان (FN) تمثل عدد المشاهدات التي تم تصنيفها بشكل غير صحيح على انها سلبية عندما كانت في الواقع ايجابية.

3. النتائج

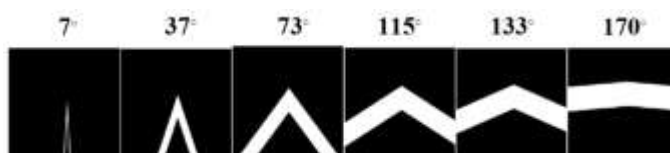
البيانات المستخدمة في الدراسة هي عبارة عن ظل زوايا CEM في حديقة طعام بكتيرية على لوحة اجار Agar plate بشكل سلاسل زمنية طويلة وبعدد مشاهدات كبيرة جداً. السلسلة الزمنية هي عبارة عن مقطع فيديو لحركة الدودة خلال ما يقارب 2.5 ساعة. تتراوح درجة زوايا جسم CE من 1° عندما تنحني على نفسها تقريباً الى الزاوية اقل من 180° كأكبر زاوية ممكنة عندما تصل الى الاستقامة تقريباً (15). إن شكل 6 أدناه يوضح انحناءات CE للعديد من الزوايا المختلفة تم تصويرها من ابعاد مختلفة.



شكل 6: سباحة وحركة دودة اليربوع الرشيق في زوايا وابعاد مختلفة (16).

في دراسات معينة يكون الاهتمام بدراسة متى تقطع الدودة مسافات أكبر في وقت أقل وذلك عندما تتحرك بسرعة أي عندما تكون حركتها بزوايا حادة. وهناك دراسات أخرى يكون الاهتمام فيها حول توقفات الدودة لوجود مشكلة ما أو بطئ حركتها أي عندما تكون حركتها بزوايا منفرجة نسبياً. ولذلك يتم تحديد الصفة الإيجابية والحالة السلبية حسب طبيعة الدراسة. ولأنه من الصعب التنبؤ بهذا العدد من القيم الرقمية لزوايا CEM لكون القيم كثيرة ومتقاربة من بعضها ولكن يمكن تصنيف هذه الزوايا فنوياً حسب سرعة الحركة الى حركة سريعة بزوايا حادة تمثل الصفة الإيجابية (+1) وحركة بطيئة بزوايا منفرجة تمثل الصفة السلبية (-1) أي تصنيف ثنائي binary classification. في هذه الدراسة تم الاعتماد على تحويل درجات زوايا CEM الى اشكال رسومية من خلال صور ثنائية البعد بتدرج الرمادي (Grayscale) كما موضح في

شكل 7 أدناه.



شكل 7: نماذج من تحويل درجات زوايا CEM الى اشكال رسومية.

بلغت عدد مشاهدات السلسلة الزمنية CEM (17984) مشاهدة ولخمس سلالات وهي (-unc, unc-1 strain, goa-1 strain, N2 reference strain, 38 strain, unc-63 strain)، تم اختيار سلسلتين زمنيتين عشوائياً من كل سلالة من السلالات الخمس وكل سلسلة زمنية تمثل CEM لدودة CE واحدة. المصفوفة رباعية البعد تمثل المتغير المستقل (X) وتم تصنيف هذه الصور الى فئتين (1 و -1) وهي الحادة والمنفرجة لغرض تمثيل المتغير المعتمد (Y). يعد أسلوب CNN مناسباً بشكل خاص لتحليل بيانات صورية لأشكال الزاوية التي تشكلها CE عند حركتها على شكل مشاهدة لكل وحدة زمنية ضمن المدى من ما يقارب (1°) الى ما يقارب (177°) وإن حد العتبة بين الزوايا الحادة والمنفرجة هو الزاوية (90) درجة للتصنيف الثنائي. لإنشاء متغير الإدخال لاسلوب CNN تم تحويل كل سلسلة زمنية من صيغتها الرقمية numerical وتشكيلها كصور ثنائية البعد وتم حفظ كل متغير سلسلة زمنية

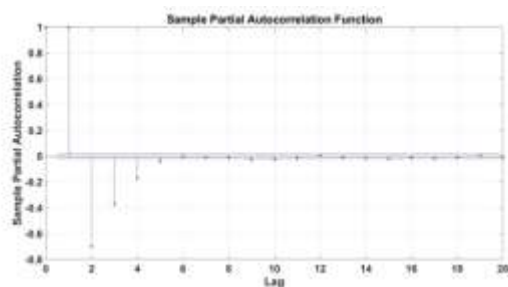
كمصفوفة رباعية الأبعاد. البعد الرابع للصورة فيمثل تسلسل الملاحظة التي تم التعبير عنها كصورة. في هذه الدراسة ظهرت صور CEM بشكل تلقائي بحجم (264×251) بكسل وبعدد مشاهدات 14400 تعادل تقريبا 80% من المشاهدات الكلية وعددها 17984 مشاهدة تقابل فترة التدريب و3584 مشاهدة أي تقريبا 20% من المشاهدات الكلية لفترة ولذلك فإن الحجم النهائي لصورة متغير الإدخال هو (246×251×1×17984) بكسل.

3.1. الانحدار الذاتي (AR) Auto Regressive (AR)

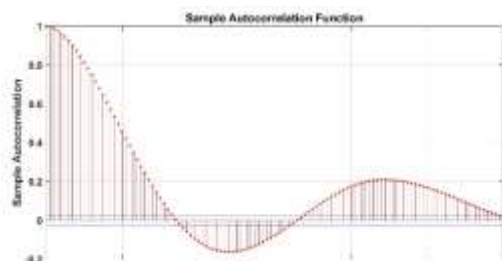
من خلال الدالتين (AR) Auto Regressive اعتمادا على مبدأ الانحدار الذاتي والارتباط الذاتي تم تحديد رتبة نموذج الانحدار الذاتي وبغض النظر عن استقرارية البيانات لأنه يؤدي الى Partial Autocorrelation Function (PACF) و Autocorrelation Function (ACF) فقدان جزء كبير من خصائص الانحدار الذاتي والارتباط الذاتي للسلسلة الزمنية الاصلية وبالتالي فقدان الخصائص الأساسية التي تتميز بها السلسلة. فمن لعينة السلسلة الأولى كما في (5) AR الممكن الاستدلال على أن افضل نموذج للانحدار الذاتي شكل 8 أدناه وذلك لأن ACF يعطي نمط مضمحل تدريجيا مع بطئ في الاضمحلال والذي يشير الى عدم استقرارية. في حين PACF تعطي نمط الانقطاع الفجائي بعد (5) من التخلفات الزمنية كما في شكل 8. وعليه بعد تطبيق دالة ACF على بيانات التدريب تم استخدام التخلفات الزمنية المشار إليها في جدول 1 أدناه.

جدول 1: نماذج الارتباط الذاتي الأنسب

السلسلة الأولى	السلسلة الثانية	السلسلة الثالثة	السلسلة الرابعة	السلسلة الخامسة
العينة الأولى	5	5	7	8
العينة الثانية	5	5	7	2



شكل 9: الارتباط الذاتي لعينة السلسلة الاولى (PACF)



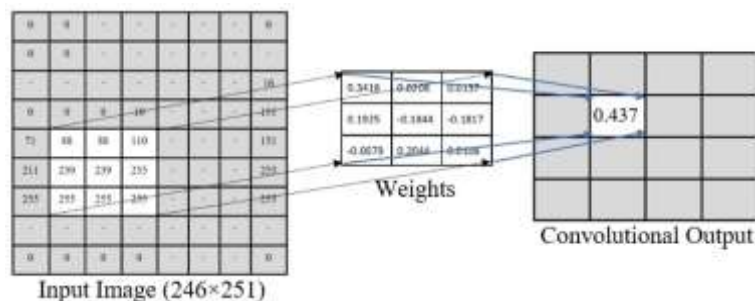
شكل 8: الارتباط الذاتي لعينة السلسلة الاولى (ACF)

3.2. الشبكة العصبية الالتفافية CNN

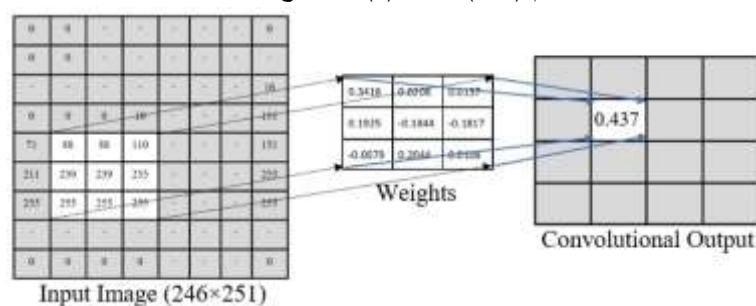
ان الاطار العام لخوارزمية تنفيذ CNN يتضمن تنفيذ عدة خطوات متسلسلة وكما يلي.

1. تحويل المشاهدات من حالتها الرقمية الى مصفوفة واحدة رباعية الأبعاد تجمع الصور مع بعضها.
2. تحديد الفتتين الإيجابية والسلبية لمتغير الهدف بصفتين للزاويا الحادة والمنفرجة.
3. تقسيم مشاهدات السلسلة الزمنية الى مجموعتين للتدريب والإختبار.
4. تحديد بنية الشبكة العصبية الالتفافية (هيكل الشبكة) طبقة الإدخال والطبقة الخفية عدد (2) وطبقة الإخراج أي أن اعداد الطبقات بشكل عام هي (1-2).

5. تحديد حجم المرشح (3×3) وعدد المرشحات بالاعتماد على جدول متغيرات الانحدار الذاتي الأمثل وخطوة واحدة في الطبقة الخفية الأولى اذ ان

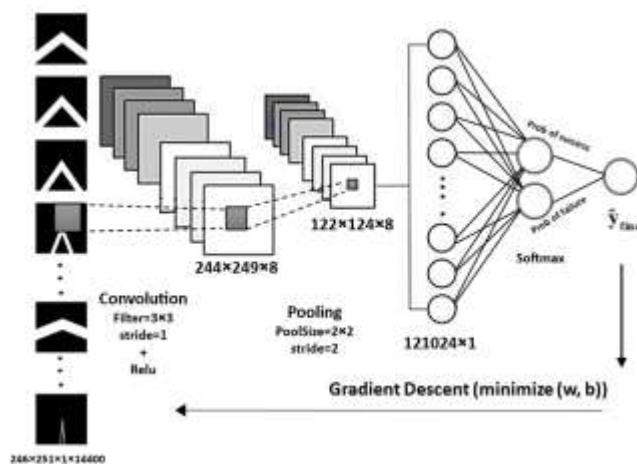


6. شكل 10 يوضح حجم الصورة وحجم المرشح وعملية التفاف المرشح على جزء من الصورة.
7. جمع قيمة التحيز مع عناصر المصفوفة (244×249) الناتج من الخطوة 5 والمعادلة (2) وتطبيق دالة ReLU عليه.
8. تم تطبيق متوسط التجميع Average Pooling بحجم (2×2) وخطوة (2) على ناتج دالة ReLU.



شكل 10 : عملية التفاف المرشح على جزء من صورة الادخال

9. جمع مخرجات متوسط التجميع في عمود واحد عن طريق عملية تسمى التسوية Flatten فيتم الحصول على vector يمثل الطبقة المتصلة بالكامل Fully Connected Layer ومن ثم تطبيق المعادلة (3) على مدخلات هذه الطبقة.
10. تطبيق دالة (Softmax) على مخرجات الطبقة المتصلة بالكامل.
11. تحديد خيارات التدريب ومنها عدد تكرار المحاولات (افتراضي) ومعدل التعلم (0.01).
12. انشاء الشبكة للتدريب. وبعد الحصول على القيم المثلى للأوزان والتحيز من خلال طبقة التصنيف فإن الاجراء النهائي الذي تقوم به CNN هو تصنيف السلسلة الزمنية بمقارنة المتغير $\hat{y}_{final}$ مع متغير القيم الأصلية وهو متغير الهدف فسيتم حساب دقة نموذج التصنيف بالنسبة للقيم الحقيقية للسلسلة الزمنية باستخدام مقياس تقييم دقة التصنيف بتطبيق المعادلة (13). شكل 11 أدناه يوضح خوارزمية الشبكة العصبية الالتفافية التي تم تطبيقها على صور زوايا حركة الدودة.



شكل 11: عملية تدريب الشبكة العصبية الالتفافية لبيانات العينة الاولى من السلسلة الرابعة

تم تطبيق معادلة (13) لقياس دقة نموذج التصنيف لبيانات التدريب والاختبار لمتغيرات السلاسل الزمنية وكانت النتائج كما في جدول 2 أدناه.

جدول 2: نتائج قياس دقة التصنيف بأسلوب CNN.

السلسلة الأولى	السلسلة الثانية	السلسلة الثالثة	السلسلة الرابعة	السلسلة الخامسة	
بيانات التدريب	99.80	99.56	99.74	99.6	99.78
بيانات الاختبار	99.58	99.72	99.92	99.69	99.83
بيانات التدريب	99.62	99.67	99.74	99.53	99.55
بيانات الاختبار	99.22	99.78	99.67	99.64	99.83

3.3. الانحدار اللوجستي

ان الاطار العام لخوارزمية تنفيذ LR يتضمن تنفيذ عدة خطوات متسلسلة وكما يلي.

1. استخدام متغيرات الانحدار الذاتي الأمثل اعتمادا على
2. جدول 1 لتحديد متغيرات الادخال لاسلوب LR.
3. تحديد الفئتين الإيجابية والسلبية لمتغير الهدف. وتقسيم مشاهدات السلسلة الزمنية الى مجموعتين للتدريب والاختبار.
4. تدريب نموذج الانحدار اللوجستي الثنائي على البيانات بواسطة اليعاز (fitglm) باستخدام المدخلات وتتضمن متغيرات الادخال والإخراج ونوع التوزيع. يتم تقييم الأداء للنموذج على بيانات الاختبار بواسطة اليعاز (predict) باستخدام المدخلات وتتضمن متغيرات الادخال بالإضافة الى النموذج LR.
5. تم تطبيق الانحدار المتدرج على متغيرات السلاسل الزمنية للإبقاء على المعلمات المعنوية فقط وكما في معادلات الانحدار اللوجستي أدناه.

معادلة (15) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلسلة الاولى

$$P(y = 1|x_1, x_4) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 30.504 - 0.4682x_1 + 0.1261x_4 \quad (15)$$

معادلة (16) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلسلة الثانية

$$P(y = 1|x_1, x_2, x_3, x_4) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 13.279 - 0.2746x_1 + 0.0263x_3 + 0.0198x_4 \quad (16)$$

معادلة (17) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلسلة الثالثة

$$P(y = 1|x_1, x_2, x_6) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 19.298 - 0.1791x_1 - 0.0694x_2 + 0.0322x_6 \quad (17)$$

معادلة (18) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلسلة الرابعة

$$P(y = 1|x_1, x_2, x_8) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 28.992 - 0.2731x_1 - 0.1025x_2 + 0.0484x_8 \quad (18)$$

معادلة (19) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلسلة الخامسة

$$P(y = 1|x_1, x_2, x_6) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 43.336 - 0.5245x_1 - 0.089x_2 + 0.1256x_6 \quad (19)$$

معادلة (20) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الثانية من السلالة الأولى

$$P(y = 1|x_1, x_4) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 29.505 - 0.4565x_1 + 0.1234x_4 \quad (20)$$

معادلة (21) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الثانية من السلالة الثانية

$$P(y = 1|x_1, x_3, x_5) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 15.181 - 0.2393x_1 - 0.0426x_3 + 0.0267x_5 \quad (21)$$

معادلة (22) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الثانية من السلالة الثالثة

$$P(y = 1|x_1, x_4) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 18.782 - 0.2657x_1 + 0.0565x_4 \quad (22)$$

معادلة (23) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلالة الرابعة

$$P(y = 1|x_1, x_7) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 29.551 - 0.386x_1 + 0.0549x_7 \quad (23)$$

معادلة (24) أدناه تمثل معادلة الانحدار اللوجستي لبيانات العينة الأولى من السلالة الخامسة

$$P(y = 1|x_1, x_2) = \frac{e^{Y'}}{1 + e^{Y'}}; Y' = 42.043 - 0.3703x_1 + 0.1062x_2 \quad (24)$$

وكانت نتائج قياس دقة نموذج التصنيف لبيانات التدريب والاختبار كما في جدول 3 أدناه.

جدول 3: دقة التصنيف للسلالات الخمس لبيانات التدريب والاختبار باستخدام نموذج الانحدار اللوجستي.

السلالة الأولى	السلالة الثانية	السلالة الثالثة	السلالة الرابعة	السلالة الخامسة	
بيانات التدريب	50.5486	52.1181	59.3611	48.9167	46.8819
بيانات الاختبار	59.0681	53.2645	13.8114	36.1328	56.1663
بيانات التدريب	54.8472	53.3403	49.1389	48.4375	40.3889
بيانات الاختبار	43.1920	52.2879	55.3571	49.2188	86.0770

3.4. طريقة النواة Kernel.

ان الاطار العام لخوارزمية تنفيذ Kernel يتضمن تنفيذ عدة خطوات متسلسلة وكما يلي.

1. استخدام متغيرات الانحدار الذاتي الأمثل لتحديد متغيرات ادخال Kernel. وتحديد الفئتين الإيجابية والسلبية.
2. تقسيم مشاهدات السلسلة الزمنية الى مجموعتين للتدريب والاختبار.
3. تدريب نموذج Kernel بواسطة اليعاز (fitckernel) باستخدام المدخلات وتتضمن متغيرات الادخال والإخراج.
4. استخدام النموذج الذي تم تدريبه لتصنيف بيانات الاختبار بواسطة اليعاز (predict) باستخدام المدخلات وتتضمن متغيرات الادخال بالإضافة الى النموذج. وكانت نتائج قياس دقة نموذج التصنيف كما في جدول 4 أدناه.

جدول 4: دقة التصنيف للسلالات الخمس لبيانات التدريب والاختبار باستخدام طريقة النواة.

السلالة الأولى	السلالة الثانية	السلالة الثالثة	السلالة الرابعة	السلالة الخامسة
----------------	-----------------	-----------------	-----------------	-----------------

99.4236	98.9306	98.5833	97.1806	99.1667	بيانات التدريب	العينة الأولى
99.5815	98.4096	99.6373	97.2377	97.3772	بيانات الاختبار	
68.5208	99.1319	99.1250	97.9792	99.1875	بيانات التدريب	العينة الثانية
35.9933	99.2467	98.8281	98.1585	98.2980	بيانات الاختبار	

4. المناقشة

في

جدول 1 تم تطبيق دالة الارتباط الذاتي ACF ودالة الارتباط الذاتي الجزئي PACF على بيانات التدريب وتم استخدام التخلقات الزمنية المشار إليها في الجدول في تحديد هيكلية الأساليب المستخدمة في التصنيف. وفي **Error! Reference source not found.** فان نتائج قياس دقة نموذج CNN تؤكد زيادة الدقة في التصنيف اذ انه في جميع السلالات كانت النتائج ممتازة في مرحلتي التدريب والاختبار. ومن خلال نتائج قياس دقة التصنيف لنموذج الانحدار اللوجستي لبيانات التدريب كما في وكانت نتائج قياس دقة نموذج التصنيف لبيانات التدريب والاختبار كما في جدول 3 أدناه.

جدول 3 يتضح انها تتراوح بين 40.3889 و 59.3611 لجميع السلالات اذ ان العينة الأولى من السلالة الثالثة تمثل اعلى دقة تصنيف بالنسبة لبيانات التدريب وبالنسبة لبيانات الاختبار فان النتائج تتراوح بين 13.8114 و 86.0770 اذ ان العينة الثانية من السلالة الخامسة كانت تمثل اعلى دقة في التصنيف وواضح ان هناك تقلب كبير وعدم استقرار في أداء النموذج في مرحلة الاختبار. وأخيرا في جدول 4 يتبين ان أداء النموذج باستخدام طريقة النواة في مرحلة التدريب حقق نتائج جيدة في جميع السلالات بنسبة تتراوح بين ما يقارب 97.1250 و 99.4236 ولكن في العينة الثانية للسلالة الخامسة كانت اقل دقة بنسبة 68.5208 تقريبا. وفي مرحلة الاختبار كان أداء النموذج جيدا في جميع السلالات ما عدا العينة الثانية من السلالة الخامسة كانت نسبة الدقة ضعيفة.

5. الخلاصة والاستنتاجات

في هذه الدراسة تم استخدام الاسلوب CNN والذي يختص دون غيرها بتحويل المشاهدات الرقمية الى صور كأسلوب مقترح لتحسين نتائج دقة التصنيف لبيانات السلسلة الزمنية لدودة اليربوع الرشيق CE. تم استخدام عينتين من البيانات كل منهما تحتوي خمس سلالات ووضحت النتائج تفوق الأسلوب المقترح CNN على أسلوب الانحدار اللوجستي و Kernel كأساليب تصنيف بديلة عند استخدام AR كأساس لتحديد عدد المتغيرات الداخلة الى الانحدار اللوجستي و Kernel. تم استخدام معيار دقة التصنيف لبيان جودة التصنيف. من الممكن استنتاج إمكانية استخدام CNN كطريقة مثلى مع بيانات السلاسل الزمنية بعد تحويل مشاهداتها الى صور لبيانات احد أنواع الديدان الاسطوانية والتي تحمل بصفحتها عدد كبير جدا من مشاهدات السلاسل الزمنية.

References

1. Javer, A., et al. *Identification of C. elegans strains using a fully convolutional neural network on behavioural dynamics*. in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.
2. Uhlemann, J., O. Cawley, and T. Kakouli-Duarte. *Nematode Identification using Artificial Neural Networks*. in *DeLTA*. 2020.
3. García Garvía, A., et al., *Towards lifespan automation for Caenorhabditis elegans based on deep learning: analysing convolutional and recurrent neural networks for dead or live classification*. *Sensors*. 2021; 21 (14): 4943.
4. Newell Stamper, B.L., et al., *Movement decline across lifespan of Caenorhabditis elegans mutants in the insulin/insulin-like signaling pathway*. *Aging cell*. 2018; 17 (1): e12704.

5. Maertens, T., et al. „*Multilayer network analysis of C. elegans: Looking into the locomotory circuitry*. Neurocomputing. 2021; 427: 238-261.
6. Le, Q., T. Sarlós, and A. Smola. *Fastfood-approximating kernel expansions in loglinear time*. in *Proceedings of the international conference on machine learning*. 2013.
7. Brockwell, P.J. and R.A. Davis, *Time series: theory and methods*. 2009: Springer science & business media.
8. Liu, L.-M., *Time Series Analysis and Forecasting*. 2nd ed. 2006, Illinois, USA: Scientific Computing Associates Corp.
9. Theobald, O., *Machine learning for absolute beginners: a plain English introduction*. Vol. 157. 2017: Scatterplot press.
10. Neapolitan, R.E. and X. Jiang, *Artificial intelligence: With an introduction to machine learning*. 2018: CRC Press.
11. Zhao, B., et al., *Convolutional neural networks for time series classification*. Journal of Systems Engineering and Electronics. 2017; 28 (1): 162-169.
12. Worster, A., J. Fan, and A. Ismaila, *Understanding linear and logistic regression analyses*. Canadian Journal of Emergency Medicine. 2007; 9 (2): 111-113.
13. Smola, A.J., S. Vishwanathan, and T. Hofmann. *Kernel methods for missing variables*. in *International Workshop on Artificial Intelligence and Statistics*. 2005. PMLR.
14. Luque, A., et al., *The impact of class imbalance in classification performance metrics based on the binary confusion matrix*. Pattern Recognition. 2019; 91: 216-231.
15. Yemini, E., et al., *A database of c. elegans behavioral phenotypes*. Nature Methods. 2014; 10 (9): 877-879.
16. Bilbao, A., et al., *Roll maneuvers are essential for active reorientation of Caenorhabditis elegans in 3D media*. Proceedings of the National Academy of Sciences. 2018; 115 (16): E3616-E3625.

Comparison of Logistic regression, Convolution Neural Network, and Kernel Approaches for Classifying the Caenorhabditis Elegans Motion

Omar Akram Malaa

Osamah Basheer Shukur

Department of Statistics and Informatics\ Faculty of Computer Sciences and Mathematics\ University of Mosul\ Mosul\ Iraq.

Abstract: Time series data are widely used in many fields including microbiology data. It is necessary to know how to classify the category to which observation belongs by using statistical classification methods and machine learning and deep learning algorithms. The study of the movement of some types of nematodes as one of the types of microorganisms including *Caenorhabditis elegans* (CE) is important to determine the actions and their impact on the life of the worms. In this study the CE motion time series data were represented by its wave motion angles which would be the study case. the non-linearity and uncertainty will be among the most common problems in this type of data that may lead to classifications that are not accurate. Convolutional Neural Network (CNN) will be used as one of the deep learning techniques and it is a non-linear method used to classify CE movement as a dependent variable in binary cases based on images of wave motion angles as an independent variable and its use will lead to accurate results because it is a suitable non-linear method to deal with Study data to solve nonlinearity and uncertainty problems through digital data visualization. Logistic regression (LR) and kernel method were also used to classify CE angles of movement. The AR(p) rank was used to determine the structure of the used methods. And by comparing the results between the methods used, it was found that the CNN method is superior to the other methods used. Therefore, it is possible to conclude that the use of the CNN method, which is based on pictorial classification, leads to accurate classification results compared to other methods based on numerical classification.

Keywords: Logistic Regression (LR), Convolutional Neural Network (CNN), Kernel Method, Classification, Time series, Autoregressive (AR), *Caenorhabditis elegans* (CE).