مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

مقاربة علمية للمساهمة في كشف التزييف العميق لمحتوى مرئي

دعاء مهنا *

نورا كويس **

(تاريخ الإيداع 2022/11/9 . قُبِل للنشر في 2022/12/14)

□ ملخّص□

تعتمد تقنية التزييف العميق (Deepfake) على استبدال صورة وجه شخص بوجه شخص آخر مستهدف، أو استبدال صوت شخص بصوت شخص آخر مستهدف، لتبدو مقاطع الوسائط المرئية أو الصوتية المزيفة حقيقيّة، وعلى الرغم من أهمية هذه التقنية إلا أنها تملك أضرار كبيرة تتمثل بالسماح باستغلال هذه التقنيات لصنع محتويات غير صحيحة تؤدي الى خلق مشاكل أمنية و اجتماعية.

الهدف من البحث هو تقديم مقترح لمنهجية قادرة على كشف التزييف الحاصل على المقاطع المرئية من خلال العمل على تدريب مجموعة من خوارزميات الذكاء الصنعي لاستخلاص الوجوه بأعلى دقة ممكنة ثم استخلاص ميزات هذه الوجوه بشكل دقيق وتصنيف المقطع المرئي (حقيقي/مزيف) مع مراعاة تسلسل البيانات الواردة في المقطع(المحور الزمني).

يقدم البحث منهجية علمية مقترحة لتصنيف مقطع مرئي مع إيضاح دقة هذا التصنيف، وقد اعتمدت هذه المنهجية على:

- ✓ الرؤية الحاسوبية من خلال مكتبة opencv لتحويل المقطع المرئي الى سلسة من الصور.
- ✓ الشبكة العصبونية الالتفافية متعددة المهام MTCNN لاستخراج الوجوه.
- ✓ شبكة الرواسب ResNeXtCNN-50 (32x4d) لاستخراج ميزات الوجوه.
- ✓ الشبكة العصبونية المتكررة ذات الذاكرة طويلة قصيرة الأمد لإعطاء التصنيف النهائي للمقطع مع دقة التنبؤ.

الكلمات المفتاحية: كشف التزييف العميق، الشبكة العصبونية الالتفافية متعددة المهام ، الشبكة الرواسب، الرؤية الحاسوبية ، الذكاء الصنعي، الشبكة العصبونية المتكررة ذات الذاكرة طويلة قصيرة الأمد.

*ماجستير في قسم تكنولوجيا المعلومات-كلية هندسة تكنولوجيا المعلومات والاتصالات-جامعة طرطوس-طرطوس-سوريا
**ماجستير في قسم تكنولوجيا الاتصالات-كلية هندسة تكنولوجيا المعلومات والاتصالات-جامعة طرطوس-طرطوس-سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

A scientific approach to contributing to the detection of deepfakes in visual content

D'uaa mhnaa *
Noura kuays **

(Received 9/11/ 2022 . Accepted 14/12/ 2022)

□ ABSTRACT

Deepfake technique is based on replacing the image of a person's face with the face of another target person, or replacing a person's voice with the voice of another target person, So that the fake visual or audio clips appear real, In spite of the importance of this technology, it has great damages represented in allowing the exploitation of these techniques to create incorrect content that leads to the creation of security and social problems.

The aim of the research is to present a proposal for a methodology capable of detecting fake video clips by working on training a set of artificial intelligence algorithms to extract faces with the highest possible accuracy and then extract the features of these faces accurately and classify the video clip (real/fake) taking into account the sequence of incoming data in the segment (time axis).

The research presents a proposed scientific methodology for classifying a video clip with an explanation of the accuracy of this classification, this methodology relied on:

- ✓ Computer vision through the opencv library to convert a video clip into a series of images.
- ✓ Multitasking convolutional neural network MTCNN for face extraction.
- ✓ ResNeXtCNN-50 (32x4d) network for feature extraction of faces.
- ✓ Short-term long-term memory recurrent neural network to give the final segment classification with prediction accuracy.

Keywords: Deepfake detection, Multitasking convolutional neural network, Residual network, Computer vision, Artificial intelligence, short-term long-term memory recurrent neural network.

*Postgraduate و Department of Information Technology, Faculty of Information and Communication Technology, University of Tartous, Syria
**Postgraduate و Department of Communication Technology, Faculty of Information and Communication Technology, University of Tartous, Syria

1-المقدمة

ساهم تطور التكنولوجيا الكبير لكاميرات الهواتف الذكية وتوافر الإنترنت في جميع أنحاء العالم إلى سهولة الوصول لوسائل التواصل الاجتماعي وبالتالي سهولة مشاركة الوسائط الرقمية .

هذا التطور أدى الى زيادة القوة الحسابية لتقنيات التعلم العميق مما جعلها قادرة على إنشاء محتوى رقمي جديد بالاعتماد على الشبكات التوليدية التنافسية "Generative Adversarial Networks" [1] والتي أدت الى ظهور تحديات جديدة أطلق عليها اسم التزييف العميق (Deepfake).

يتم تصنيف المحتوى المزيف إلى فئتين أساسيتين[4] :

- ✓ التزييف السطحي Shallowfakes:

هو طريقة لمعالجة محتوى الوسائط دون استخدام أساليب التعلم الآلي والأنظمة الحسابية بحيث لاتتضمن هذه التقنية استخدام أنظمة التعلم العميق وفي المقابل يتم تطبيق برنامج لتحرير وتعديل محتوى الوسائط بشكل يدوي ويتضمن هذا النوع من التزييف:

1.مقاطع فيديو ذات حركة بطيئة: بحيث يتم استخدم برنامج يعمل على إبطاء سرعة الكلام دون تغير طبقة الصوت، ويقصد من ذلك الإشارة إلى وجود خلل في الشخص المستهدف من خلال الفيديو أو التشديد على كلمات معينة أو نبرة الصوت لتزييف وجهات نظر محددة ولترك انطباعا خاطئ لدى الجمهور.

2. تغير التواريخ والمواقع: التلاعب بالتواريخ والمواقع لتظهر مقاطع الفيديو على أنها حديثة وفي أماكن مختلفة، مما يؤدي إلى انتشار أخبار كاذبة تضر بسلامة المجتمع والأفراد.

- ✓ التزييف العميق Deepfakes:

تعتمد تقنية التزييف العميق (Deepfake) على استبدال صورة وجه شخص بوجه شخص آخر مستهدف، أو استبدال صوت شخص بصوت شخص آخر مستهدف، لتبدو مقاطع الوسائط المرئية أو الصوتية المزيفة حقيقيّة، من خلال استخدام الشبكة التوليدية التنافسية(Generative Adversarial Networks) والتي تتضمن الخوارزمية التوليدية والتي يتم فيها إدخال بيانات عشوائية لتحويلها إلى صورة، ثم تضافُ هذه الصورة المصطنعة ضمن سلسلة من الصور الحقيقية ثم يتم إدخالها في الخوارزمية الثانية المعروفة باسم خوارزمية التمييز " Discriminator" ومع بدء العملية لا تبدو الصور التي يتم إنتاجها على أنها صور وجوه، إلا أن تكرار العملية عدة مرات وإجراء التعديلات يؤدي إلى تحسين أداء خوارزميتي التمييز والتوليد وبعد تنفيذ عدد كاف من الدورات والملاحظات تبدأ الخوارزمية في إنتاج وجوه واقعية تماما لأشخاص غير حقيقيين [2] وعلى الرغم من أهمية هذه التقنية إلا أن تملك أضرار كبيرة تتمثل بالسماح باستغلال هذه التقنيات لصنع محتويات غير صحيحة تؤدي الى خلق مشاكل أمنية و اجتماعية.

يكمن التحدي الرئيسي في إمكانية اكتشاف حدوث تزييف في المحتوى الرقمي حيث اعتمدت بعض الدراسات على نظام يقوم بتحليل قرنية العين باستخدام نماذج الشبكة العصبية العميقة(Deconvolutional Neural Network) تعتمد هذه الطريقة على اكتشاف وميض العين في مقاطع الفيديو ، وهي إشارة فيزولوجية لا يتم عرضها بشكل جيد في مقاطع الفيديو المزيفة المركبة على اعتبار أن القرنية لها سطح يشبه المرآة تقوم بتولد أنماطاً عاكسة عند سقوط الضوء عليها، حيث أن الانعكاس على العينين سيكون متشابهاً في صورة الوجه حقيقي الذي تم التقاطه بواسطة كاميرا لأنهما تشاهدان نفس الشيء على عكس الصور التي يتم تزييفها في الاختبارات التي أجريت على الصور كانت الأداة فعالة بنسبة 94% في اكتشاف التزييف العميق[3].

من منظور آخر تم استخدام كبسولات الشبكة العصبونية "capsule network " لاكتشاف الصور ومقاطع الفيديو المزيفة [5] حيث تم استخدامها لاكتشاف التزييف والمعالجة في سيناريوهات مختلفة ، مثل اكتشاف هجوم إعادة التشغيل واكتشاف الفيديو الذي تم إنشاؤه عن طريق الحاسب بالاعتماد على الشبكة العصبونية الالتفافية مع الاشارة الى استخدام الضوضاء العشوائية في مرحلة تدريب الشبكة مما يجعل النموذج غير فعال مع البيانات في الوقت الحقيقي.

في هذه البحث تم العمل على:

- ❖ إجراء معالجة مسبقة للمقطع المرئي.
- ❖ استخلاص الوجوه
- ❖ تدريب الشبكة العصبونية لاستخراج ميزات الوجوه
- ❖ معالجة التسلسل التكراري للمقطع الجديد وتصنيفه(حقيقي/مزيف).
- ❖ عرض النتيجة.

2-أهمية البحث وأهدافه

يشكل اكتشاف التزييف العميق تحدياً كبيراً نظراً لصعوبة تدريب خوارزميات الذكاء الصنعي على تحديد اتجاه معين لذلك تم استخدامه بشكل كبير لإحداث تأثيرات سياسية واقتصادية ودينية واجتماعية تؤثر على ثقة الجمهور بمصادر المعلومات المختلفة الخاصة بالفيديو بالإضافة الى إمكانية انتحال هوية أشخاص آخرين بهدف التشويه أو التزوير أو الاختلاس أو الابتزاز

يعمل التزييف العميق على جمع البيانات من ملفات الصوت والفيديو بواسطة تقنيات ذكاء التعلم الاصطناعي بشكل دقيق، وتتضمن مجالات التزييف العميق استبدال وجه شخص بآخر و تزامن تحريك الشفاه إذ يمكن ضبط فم المتحدث على ملف صوتي مختلف عن الصوت الأصلي من أجل استخدامه لقول أشياء أخرى، وانطلاقاً من هذا يهدف البحث الى اقتراح منهجية تؤدي الى كشف حدوث تغير في المحتوى المرئي من خلال إظهار كيفية الحصول على الصور من مقاطع الفيديو وآلية استخلاص ميزات الوجه من هذه الصور بالاعتماد على الشبكة العصبونية Residual Network ثم تصنيف المحتوى الرقمي باستخدام الشبكة العصبونية المتكررة ذات الذاكرة طويلة قصيرة الأمد.

3-طرائق البحث ومواده

أنجز هذا البحث للمساهمة في إيجاد منهجية تعمل على كشف التزييف العميق الحاصل على محتوى مقطع مرئي من خلال الاعتماد على عدم التناسق الحاصل بين منطقة الوجه المستبدلة والمنطقة المحيطة بها وفق التالي:

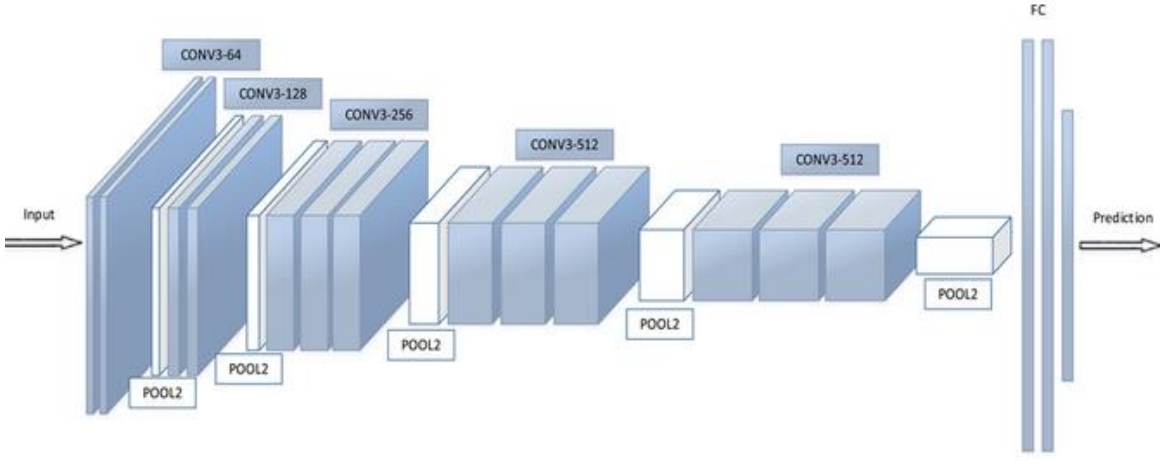
- إجراء معالجة للمقطع المرئي المراد التنبؤ به وتتضمن هذه المعالجة تقسيم الفيديو الى إطارات واستخلاص الوجوه الموجودة ضمن هذه الأطارات.
- استخلاص ميزات الوجوه بالاعتماد على الشبكة العصبونية ResNext CNN.
- استخدام الشبكة العصبونية المتكررة ذات الذاكرة طويلة قصيرة الأمد RNN(LSTM) لتصنيف مقطع الفيديو(مزيف/حقيقي) بالاعتماد على التحليل الزمني للفيديو بالمقارنة مع لحظات زمنية مختلفة.

هذا المنهج نُفذ باستخدام المحاكاة الحاسوبية من خلال:

- برنامج Anaconda ومن خلاله تم العمل على برنامج Jupyter Notebook لتوفيره المكتبات الداعمة لبحثنا.
- الاعتماد برمجياً على لغة البايثون Python الإصدار 3.6 بسبب احتوائها على المكتبات الداعمة لمعالجة الفيديو والصور
- مجموعة بيانات تتضمن المقاطع مرئية المطلوبة لتدريب واختبار الشبكات العصبونية.

3 -1 الشبكات العصبونية الالتفافية CNN:

تعمل هذه الشبكة الموضحة في الشكل (1) مع الصور ضمن مجال التعلم العميق[8]، بحيث يكون دخل الشبكة العصبونية هي عبارة عن صور أو بشكل أكثر تحديد مصفوفة ثلاثية الأبعاد ،وتتألف هذه الشبكات من سلسلة من الطبقات والتي تتعلم استخراج السمات المميزة من أي صورة .



الشكل (1) الشبكة العصبونية الالتفافية

- **طبقة التفافية (CONV):**تعتبر الأساس في هذا النوع من الشبكات العصبية ، والتي تقوم بتطبيق سلسلة من مُرشحات الصور (filters) المختلفة على الصورة المدخلة، وهذه المُرشحات تستخرج سمات مختلفة من الصورة مثل حواف الأجسام والزوايا والتدرجات اللونية،وبينما تتدرب شبكات الطبقة الالتفافية العصبية فهي تقوم بتحديث الأوزان ضمن هذه الطبقة باستخدام الانتشار العكسي (backpropagation) وهذه الأوزان بدورها تحدد نوعية مُرشح الصورة، والناتج النهائي

هو مُصنف (classifier) يتألف من العديد من الطبقات الالتفافية والتي بدورها تعلمت كيفية ترشيح الصورة لاستخراج السمات (features) المهمة منها.

❖ **طبقة التجميع(pool):** تعد هذه الطبقات اختيارية في تصميم الشبكة، وفي حال وجودها سيكون موقعها بعد كل طبقة من الطبقات الالتفافية وتهدف إلى تخفيض عدد العينات أو العصبونات حيث ستقوم باختصار كل مجموعة من عصبونات الدخل بحجم معين إلى عصبون واحد، ويحدد هذا الحجم ضمن تصميم الشبكة وتكون قيمته المثلى 2×2 لأن تكبيرها قد يؤدي إلى ضياع في المعلومات ويتم التخفيض بعدة طرق منها :

القيمة العليا المشتركة (max pooling) :تأخذ القيمة الأعلى بينها.

المعدل المشترك (average pooling): تأخذ معدل جميع القيم

❖ **طبقة الاتصال الكامل(FC):**

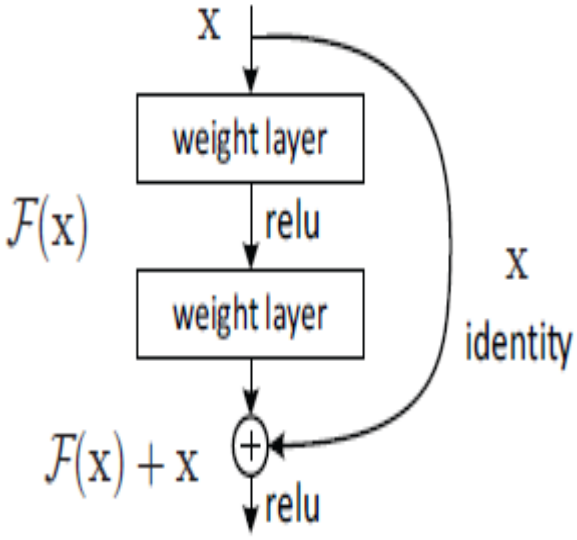
بعد عدة طبقات من النوعين السابقين تأتي هذه الطبقات لتربط كل عصبونات الطبقة السابقة (مهما كان نوعها) وتجعلها دخل لكل عصبون فيها كما في الشبكات العصبونية العادية، لا يشترط أن تكون بعدد معين ولكن غالباً يوجد منها طبقتان متتاليتان كالطبقات الأخيرة في الشبكة إذ لا يمكن أن تأتي قبل طبقة من النوع الالتفافي.

3 -2 الشبكات العصبونية Residual Network:

تعتبر شبكة الرواسب Residual Network (ResNET) [9] من أكثر الشبكات الرائدة في مجال الرؤية الحاسوبية والتعلم العميق، وذلك بعد التقدم الكبير الذي حققته شبكة AlexNet (شبكة التفافية عصبونية CNN عميقة تتألف من ثماني طبقات عصبونية خمسة منها هي طبقات التفافية وثلاثة هي طبقات اتصال كامل Fully Connected) في تصنيف مجموعة البيانات المرئية، حيث حافظت الشبكة الراسبة ResNET على الأداء العالي في تدريب مئات وآلاف الطبقات.

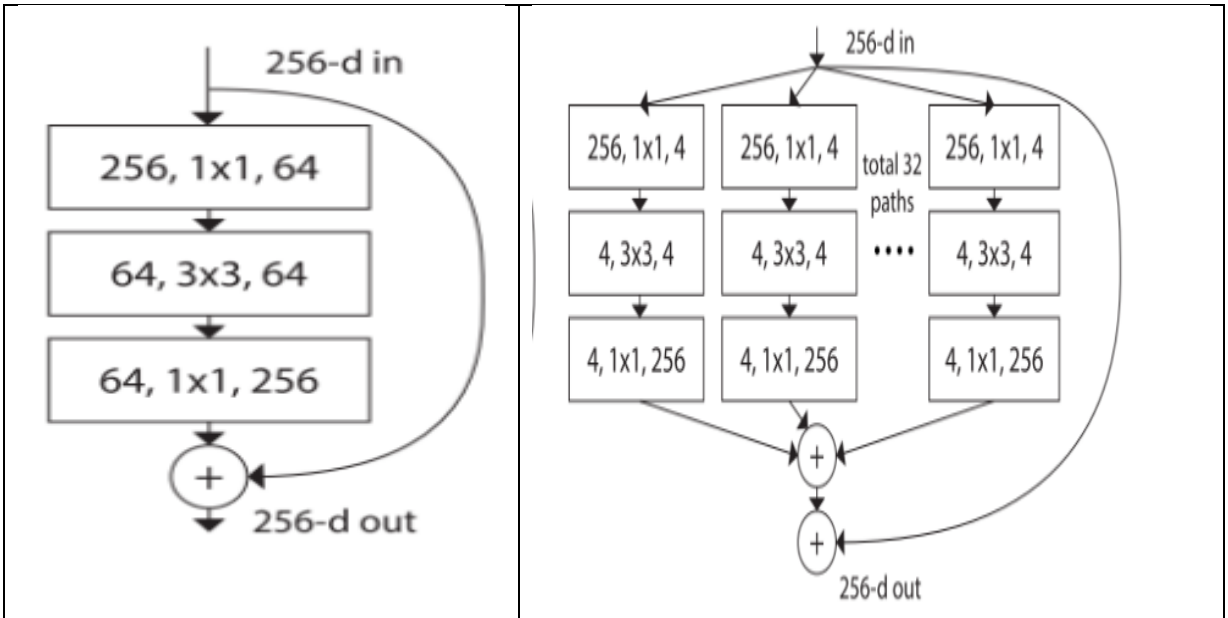
ساهمت القدرة التمثيلية القوية لشبكة الرواسب في رفع أداء العديد من تطبيقات الرؤية الحاسوبية مثل التعرف على الوجوه واكتشاف الكائنات مقارنة بالأداء في تطبيقات تصنيف الصور.

يعتبر أساس شبكة الرواسب ما يسمى بـ وصلة الاختصار (identity shortcut connection)التي تسمح بتجاوز طبقة أو أكثر كما هو موضح في الشكل (2) والذي يمثل كتلة البناء أي أن انتشار الإشارة في كتل االبناء في شبكة الرواسب أماماً وخلفاً يتم من كتلة بناء إلى أي كتلة أخرى، وبالتالي سهولة تدريب هذه الشبكة بالاعتماد على مفهوم محدد التخطّيط (identity mapping)كوصلات لتجاوز بعض الكتل .



الشكل (2) كتلة البناء

تم تطوير شبكة رواسب مختلفة ResNeXt [10] لها كتلة بناء بحيث عدد المسارات في كتلة البناء لشبكة ResNeXtتساوي 32، ولها نفس درجة التعقيد مقارنة مع كتلة البناء لشبكة الرواسب ResNET حيث يتم عرض الطبقة في الشكل (3) من خلال توضيح قنوات الدخل وقنوات الخرج وحجم المرشح



كتلة البناء في شبكة ResNeXt	كتلة البناء في شبكة ResNET

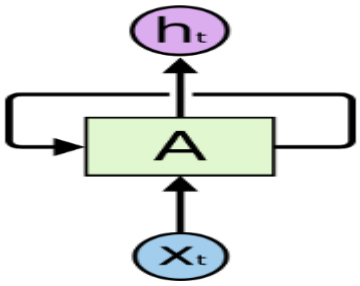
الشكل (3) بنية كتلة البناء في شبكات الرواسب

تعتمد كتلة البناء في شبكة ResNeXt على طريقة التقسيم – النقل – الدمج، بحيث يتم في هذا النوع من الشبكات دمج خرج المسارات المختلفة بإضافتها إلى بعضها البعض كما تحتوي شبكة ResNeXt على متغير قابل للضبط (hyper–parameter) يسمى بعدد العناصر وهو عدد المسارات المستقلة والذي يعطي طريقة جديدة في ضبط سعة النموذج، حيث أثبتت التجارب أن الدقة تزداد بزيادة قيمة هذا البارامتر أكثر مما تزداد بالذهاب عمقاً أو عرضاً في الشبكة.

بالإضافة الى أن هذه البنية أسهل في التكيف مع المعطّيات والمهام الجديدة لأنها ذات نموذج بسيط ولها بارامتر واحد فقط يحتاج الضبط.

3 -3 الشبكات العصبونية المتكررة ذات الذاكرة طويلة قصيرة الأمد (LSTM)RNN :

الشبكات العصبية التكراريةRNN [11]هي إحدى أنواع الشبكات التي يمكن استخدامها للتعامل مع تسلسل البيانات(المحور الزمني). يعتمد خرج هذا النوع من الشبكات على المدخلات الحالية و حالة النظام ، كما في الشكل(4) و يشبه هذا المنطق التسلسلي في الإلكترونيات الرقمية ، حيث أن الناتج يعتمد أيضًا على "flip–flop"(وحدة ذاكرة أساسية في الإلكترونيات الرقمية).



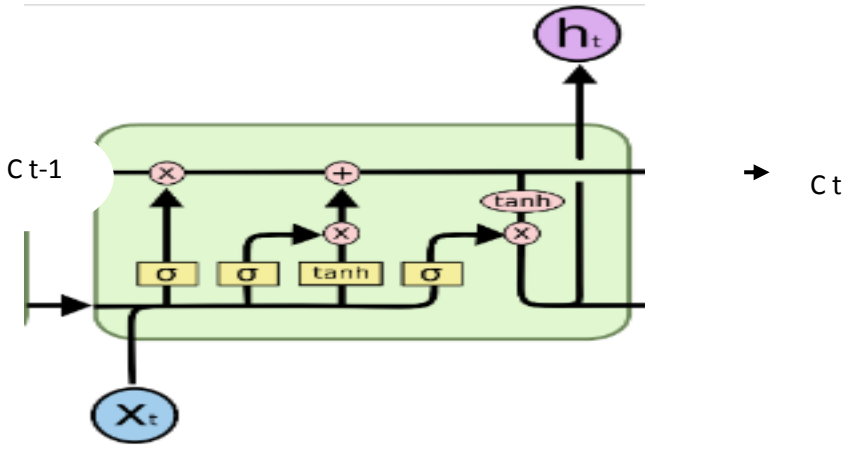
الشكل (4) بنية الشبكة العصبونية المتكررة

تتضمن الشبكات العصبونية التكراريةRNN نوعين من المشاكل:

1. مشكلة الاعتماديّة طويلة المدى long-term reliability problem : تتميز الشبكات التكرارية بالقدرة على تذكر الأحداث ولكن ومع ازدياد طول الأحداث المترابطة تنشأ مشكلة تسمّى بالاعتماديّة طويلة المدى، وهنا تفقد الشبكات التكرارية البسيطة قدرتها على التعلّم والأداء بفاعليّة عالية.
2. مشكلة انعدام التدرج vanishing gradient problem: تظهر هذه المشكلة مع ازدياد طول الشبكة، حيث تبدأ قيمة تابع الخسارة بالانعدام تدريجيّاً خلال عملية الانتشار الخلفيّ، وتصبح عملية تعديل الأوزان في الطبقات الأولى من الشبكة غير فعّالة، مما يجعل عملية التعلم بطيئة جداً وغير مُجدية.

تعتبر الشبكاتُ العصبونية التكرارية ذات الذاكرة طويلة قصيرة الأمد (LSTM)RNN نوعاً محسّناً من الشبكات العصبونيّة التكرارية، ويعتبر الهدف الرئيسيّ من تصميمها هو تفادي مشاكل الشبكات العصبونيّة االتكرارية البسيطة للحصول على نتائج أفضل.

مفتاح عمل الشبكات ذاتِ الذاكرة طويلة قصيرة الأمد هي خلية الحالة (cell state) ، فهي تعمل على المرور على طول السّلسلة بالكامل و تطرأ عليها تغيرات طفيفة وبالتالي تعتبر وسيلةً جيدة للحفاظ على المعلومات بدون تغيير.



الشكل (5) خلية الحالة

يوضح الشكل (5): خليّة الحالةالسابقة ct−1 وخلية الحالة الحالية ct ضمن الشبكة، حيث تملك الشبكات ذات الذاكرة طويلة قصيرة الأمد القدرة على تغيير المعلومات ضمن خلية الحالة عن طريقِ بنيةٍ تعتمد على البوّابات المنطقيّة وهذه البوّابات مكوّنة من مجموعة من طبقة عصبونيّة تنتهي بتابع التفعيل الأسي الجيبي Sigmoid (ծ) ومجموعة من عمليّات الضرب الموجبة،ويكون خرج طبقة تابع التفعيل الأسي الجيبي Sigmoid بين الصّفر والواحد، تحدّد قيمته كمية المعلومات الواجب السماح بمرورها من كل عنصرٍ من عناصر الخلية.

خطوات عمل الشبكاتِ ذاتِ الذاكرة طويلة قصيرة الأمد LSTM:

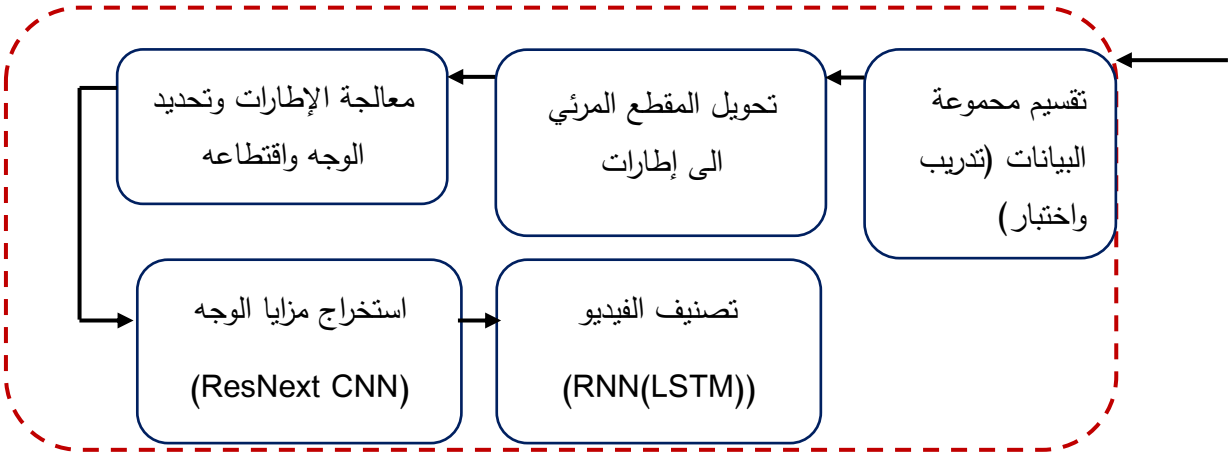
1. اتخاذ القرارِ حول المعلومات الواجب الاحتفاظ بها والمعلومات التي من الأفضل حذفها من خلية الحالة، وتتمّ هذه العمليّة ضمن طبقة تابع التفعيل الأسي الجيبي والتي تسمّى بطبقة بوّابة النسيان forget gate layer
2. تحديد المعلومات الواجب تخزينها في خلية الحالة، و تتكوّن من جزئين:

- طبقة تابع تدعى طبقة بوّابة الدّخل input gate layer مسؤولة عن تحديد القيمة المتغيرة
- طبقة تنتهي بتابع التفعيل الأسي الظلي Tanh تشكّل شعاعاً من القيم الجديدة المرشحة C~tمن المحتمل إضافتها إلى خلية الحالة

3. تعديل قيمة خلية الحالة السّابقة Ct−1إلى القيمة الجديدة Ct
4. تحديد الخرج النهائيّ، وهو مبنيّ على خرج خلية الحال

4-المناقشة

نُفذ البحث بتطبيق سلسلة من الخطوات على المقطع المرئي المراد كشف تزييفه ويمثل المخطط التالي المبين بالشكل (6) المنهجية المقترحة



يتم العمل بدايةً على مجموعة البيانات المعتمدة لتدريب الشبكات العصبونية ثم يتم استخلاص الوجه من الفيديو الهدف لإجراء خطوات استخلاص الميزات ليتم اقتطاع الوجه فقط الموجود في الفيديو بهدف إمكانية تصنيف الفيديو بحيث يمكن الاعتماد على هذا التصنيف في تحديد مدى صحة هذا المقطع المرئي.

4-1 مجموعة البيانات:

تم الاعتماد على مجموعة بيانات متنوعة تتكون من مقاطع مرئية تم جمعها من مصادر مختلفة :

- FaceForenise++: عبارة عن مجموعة بيانات مزيفة تتكون من 2000 مقطع مرئي [6].
- deepfake-detection-challenge: مجموعة بيانات تحدي اكتشاف التزييف العميق تتكون من 3000 مقطع مرئي.[7]
- Deepfake Dataset: مجموعة بيانات المستخرجة من ال تتضمن 1000 مقطع مرئي حقيقي.

احتوت مجموعة البيانات هذه على 50٪ من المقاطع المرئية الأصلية و 50٪ من المقاطع المرئية المزيفة وتم تقسيم مجموعة البيانات إلى مجموعة تدريب 70٪ (4200مقطع مرئي) ومجموعة اختبار 30٪(1800 مقطع مرئي).

4-2 المعالجة المسبقة للمقطع المرئي:

✓ **تقسيم الفيديو الى إطارات:**

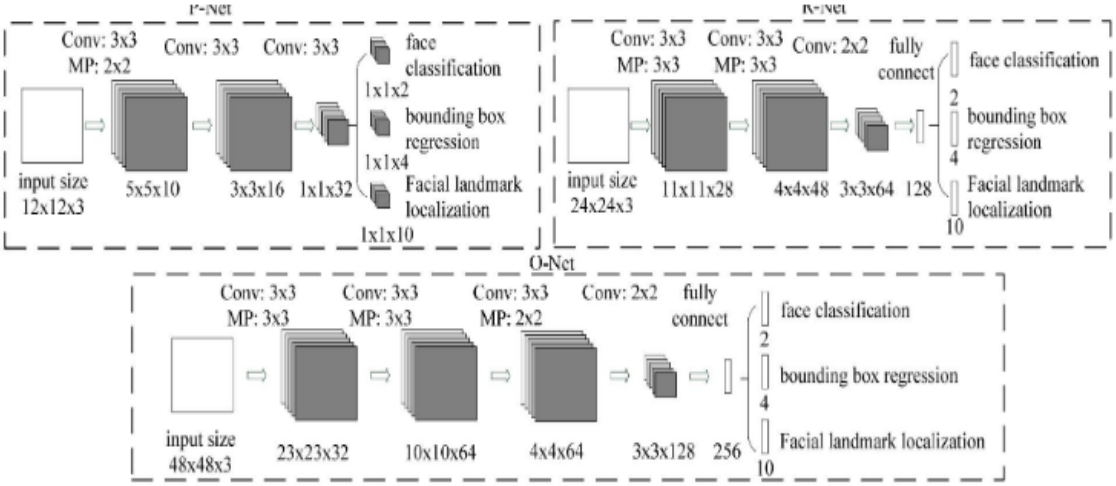
تطلب العمل على المقطع المرئي تقسيمه الى سلسلة من الصور، والتي تسمى إطارًا (الإطار هوعبارة عن فاصل زمني ثابت يتم الحصول عليه من الفيديو) وتسمى السرعة التي يتم الحصول على الإطار بها معدل الإطارات، و تم الاعتماد على مفهوم الرؤية الحاسوبية لغرض التعامل مع الفيديو وتحويله الى سلسلة من الصور، ومعالجتها من خلال الاعتماد على مكتبة opencv التي تتضمن التوابع المؤدية لهذا الغرض.

ضمن بحثنا كانت المدة الزمنية للمقطع المرئي 10 ثوانٍ أي بمعدل 30 إطارًا في الثانية ، فسيكون العدد الإجمالي للإطارات 300 إطار مما سيتطلب الكثير من المعالجة والزمن، و لذلك تجريبياً قمنا باستخدام أول 100 إطار فقط ضمن برنامج Jupter Notebook.

✓ **اكتشاف الوجه:**

❖ تم استخدام الشبكة العصبونية الالتفافية متعددة المهام MTCNN الشكل (7) لاكتشاف منطقة الوجه والتي توفرها لغة البايثون بشكل جاهز وتتضمن الطبقات التالية [8]مع مهامها:

1. Proposal Network (P-Net): تقوم هذه الطبقة بتحديد حجم الصورة ب 12*12التي تحتوي العامل البشري بالتالي تقلل من عدد الصور المدخلة اليها.
2. Refine Network(R-Net): تقوم بإجراء المزيد من الاستبعاد للصور المدخلة بحيث تحافظ على الصور الأكثر دقة.
3. Output Network(O-Net): تتعرف على منطقة الوجه تعيد نقاط التي تتضمن ملامح الوجه للشخص.



الشكل(7) بنية الشبكة العصبونية الالتفافية متعددة المهام

تكمن الأهمية من تطبيق شبكة MTCNN في الحصول فقط على الإطارات التي تتضمن الوجوه بأعلى دقة ممكنة واستبعاد باقي الإطارات.

ثم يتم اقتطاع الوجه المكتشف بواسطة الشبكة العصبونية MTCNN ولتوحيد عدد لإطارات الجديدة مع الإطارات التي تتضمنها المقاطع المرئية ضمن مجموعة البيانات التي سيتم تدريب النموذج عليها تم حساب المتوسط لتلك المجموعة وإنشاء مقطع مرئي جديد يتضمن أطر للوجه مساوية للمتوسط وتخزينه.

4-3 نموذج التدريب:

تم استخدام الشبكة العصبونية ResNeXtCNN-50 (32x4d) الشكل(8) لاستخراج مزايا الوجه الرئيسة بحيث:

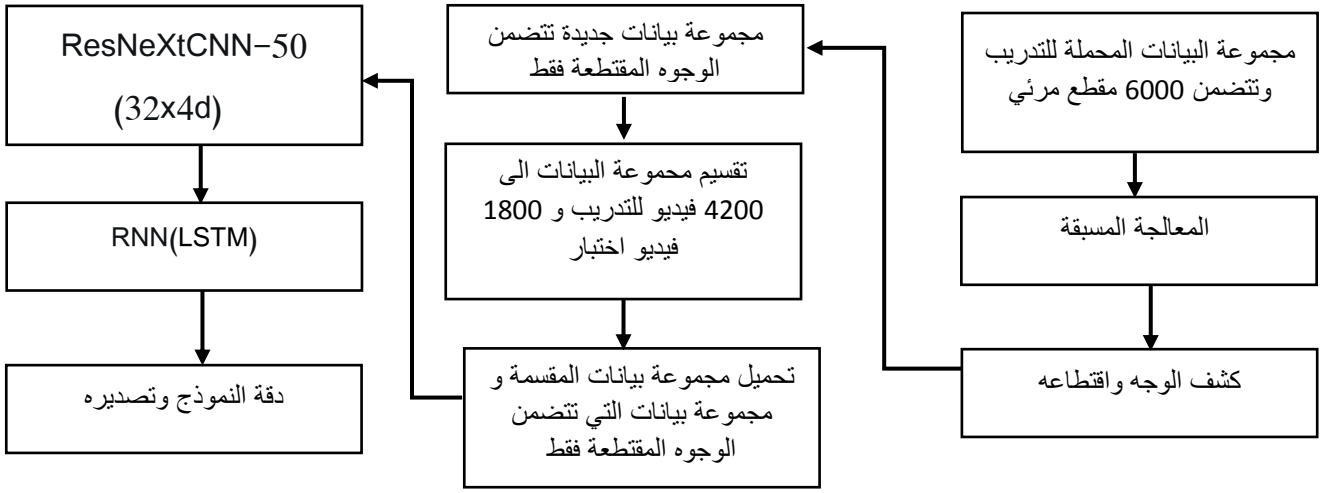
- يمثل العدد 50 عدد الطبقات الالتفافية و طبقات الاتصال الكامل.
- يمثل العدد 32 عدد الطبقات المخفية.
- يمثل 4d عدد المسارات في كل طبقة مخفية بالتالي عدد مسارات الكتلة هو 32*4=128

stage	output	ResNeXt-50 (32×4d)
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 [1×1, 128 3×3, 128, C=32 1×1, 256] ×3
conv3	28×28	[1×1, 256 3×3, 256, C=32 1×1, 512] ×4
conv4	14×14	[1×1, 512 3×3, 512, C=32 1×1, 1024] ×6
conv5	7×7	[1×1, 1024 3×3, 1024, C=32 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax
# params.		25.0×10^6

الشكل(8) بنية الشبكة العصبونية ResNeXt-50 (32x4d)

استخدمت المرشحات بحجم 1x1 ، 3 x3 ، 1 x1 لضغط الأبعاد ومعالجة الالتفاف واستعادة الأبعاد حيث عملت الشبكة العصبونية على إيجاد النقاط الفاصلة (المميزة)للوجه وتتواجد في (أعلى الدقن ، الحافة الخارجية للعين ، الحافة الداخلية للعين ، الحاجب ، حول الفم ، حول الأنف) وتم إيجاد هذه النقاط من خلال تدريب الخوارزمية بالاعتماد على الوجوه التي تم الحصول عليها من الشبكة العصبونية MTCNN بحيث تم الحصول على متجهات الميزات بأبعاد 2048.

تم استخدام الشبكة العصبونية التكرارية ذات الذاكرة طويلة قصيرة الأمد RNN(LSTM) لتصنيف هذه المتجهات الى حقيقية أو مزيفة وتطلب ذلك التعامل مع الإطارات بطريقة متسلسلة بحيث يمكن إجراء التحليل الزمني للمقطع المرئي لذلك تم إدخال متجه الميزات ذات الأبعاد 2048 الى طبقة من LSTMوالتي تعمل على مقارنة هذه الميزات ضمن الإطار عند الثانية"t" بالميزات الموجودة ضمن الإطار عند الثانية "t-n" حيث يمكن أن يكون n أي عدد من الإطارات قبل t مع فرصة تسرب بمقدار 0.4(الهدف من استخدام فرصة التسرب هو استبعاد بعد العصبونات في الطبقات المخفية بهدف تجاوز حدوث حالة over-fitting) و استخدام تابع التفعيل Relu.

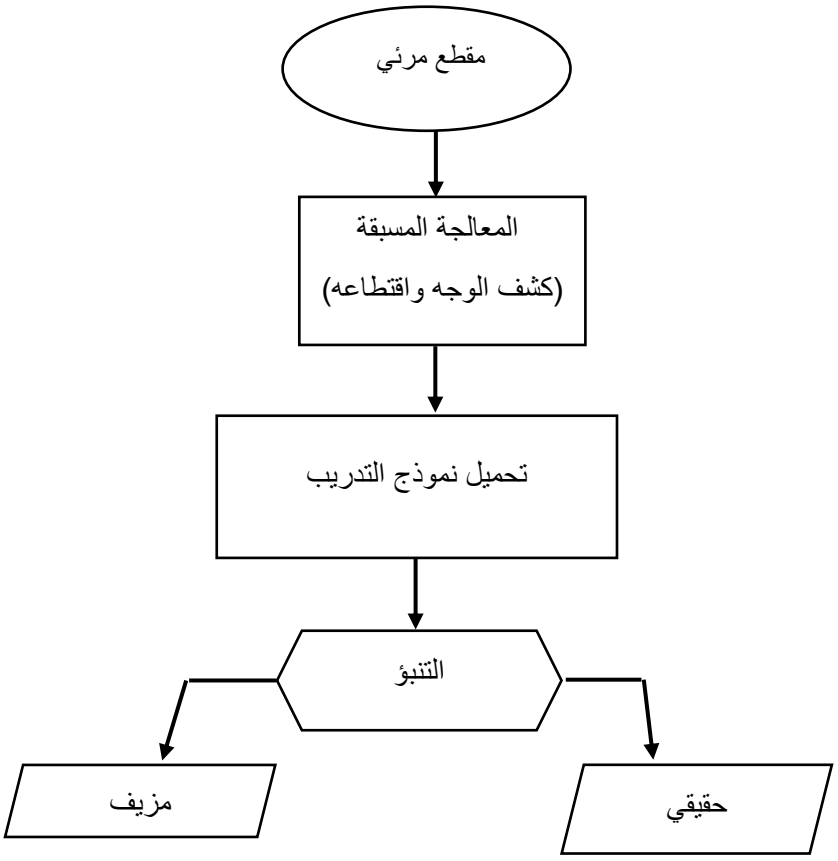


الشكل(9) نموذج التدريب

يظهر الشكل (9) ملخص لما تم شرحه سابقاً حول الخطوات المتبعة في نموذج التدريب.

4-4 نموذج التنبؤ:

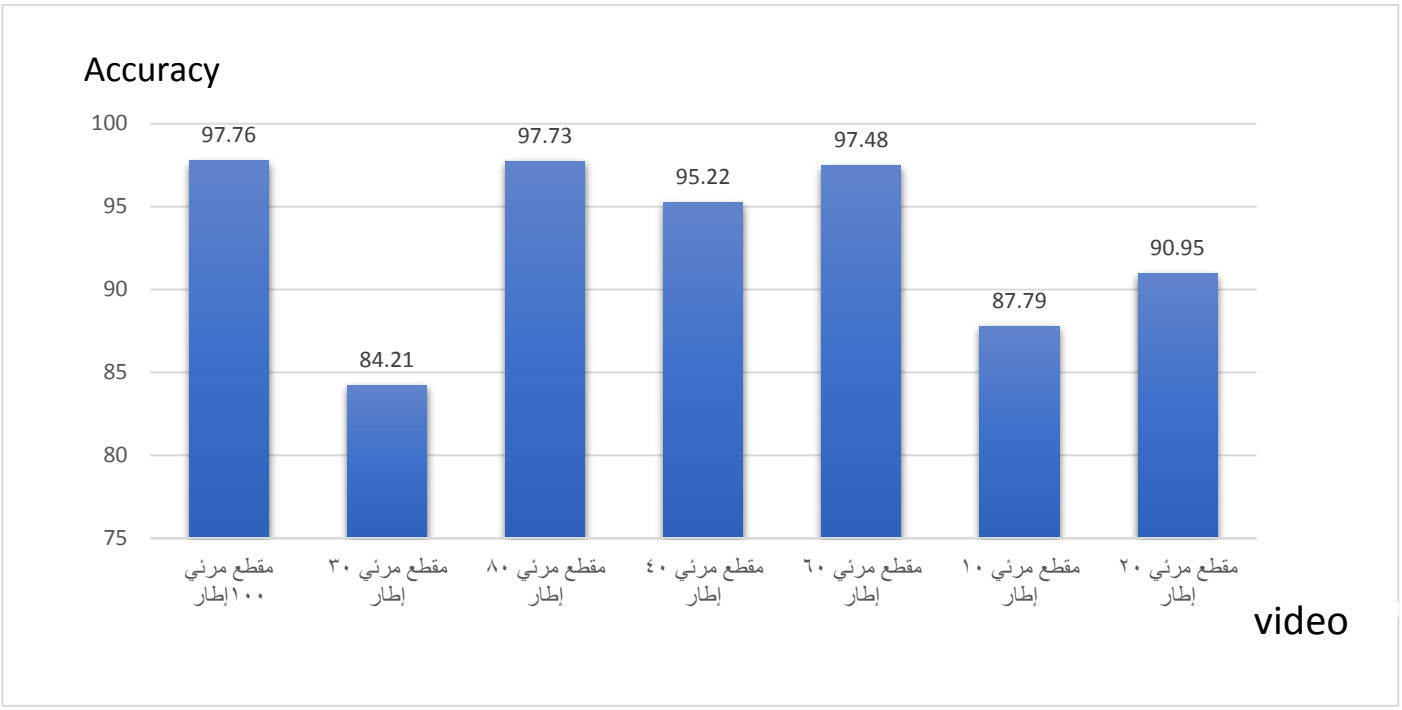
يظهر الشكل(10) نموذج التنبؤ الذي تم العمل عليه ضمن البحث بحيث يتم إجراء عملية المعالجة المسبقة للمقطع المرئي الجديد من خلال تقسيمه الى أطارات واكتشاف الوجه واقتطاعه ولكن بدلاً من تخزين الاطارات الجديدة تم تمريرها الى نموذج التدريب المشروح مسبقاً لتصنيفها.



الشكل(10) نموذج التنبؤ

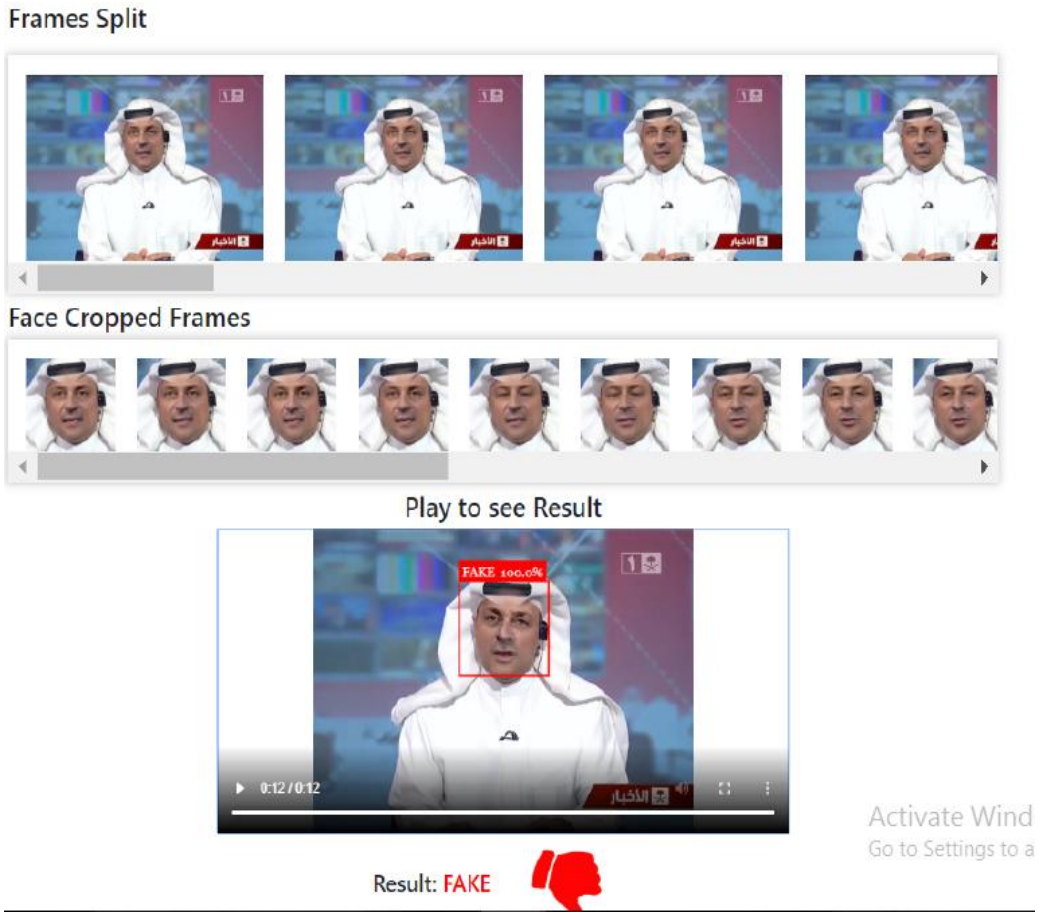
5-4 النتائج:

يظهر الشكل(11) دقة المنهجية المقترحة ويمكن الملاحظة أنه كلما ازداد عدد إطارات المقطع المرئي كلما ازداد دقة التنبؤ عن وجود تزييف للمقطع المرئي حتى دقة 97% بالتالي استخدام عدد الاطارات الكلي يعطي أعلى دقة ممكنة وهي أفضل بالمقارنة مع أداة كشف التزييف العميق المعتمدة على الشبكة العصبونية العميقة (DNN) والتي بلغت 94%و كبسولات الشبكة العصبونية التي لاتعمل بشكل فعال مع البيانات في الزمن الحقيقي.



الشكل(11) دقة خوارزمية كشف التزييف العميق

الشكل (12) التزييف اللعميق



الشكل (13) كشف التزييف اللعميق

يظهر الشكل (12) و(13) النتيجة النهائية التي ستظهر للمستخدم بعد تطبيق خطوات الخوارزمية السابقة.

5-الاستنتاجات والتوصيات

- استخدام الشبكة العصبونية متعددة المهام (MTCNN) أدى الى استخلاص صورة الوجه بأعلى دقة ممكنةوبشكل أفضل من الشبكة العصبونية الالتفافية.
- استخدام الشبكة العصبونية ResNeXtCNN أدى الى سهولة تدريب النموذج وبالتالي القدرة على استخلاص ميزات الوجه بزمن معالجة أقل.
- إظهار فعالية الشبكة العصبونية RNN(LSTM) في معالجة تسلسل البيانات في المقطع المرئي.
- استخدام تقنية التسرب أدى الى التقليل من تعقيد النموذج وبالتالي التقليل من زمن تصنيف الفيديو .

يمكن المتابعة في البحث من خلال:

- العمل على كشف التزييف العميق للصوت.
- العمل على كشف التزييف العميق وفق وميض العين .
- استخدام كبسولات الشبكة العصبونية "capsule network " لاكتشاف المقاطع المرئية المزيفة ومقارنة النتائج مع المنهجية المقترحة.

المراجع

[1] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V. and Nguyen, C.M., 2022. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, p.103525.

[2] Brownlee, J., 2019. Generative adversarial networks with python: deep learning generative models for image synthesis and image translation. Machine Learning Mastery.

[3] Li, Y., Chang, M.C. and Lyu, S., 2018, December. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International workshop on information forensics and security (WIFS) (pp. 1-7). IEEE.

[4]What is the Difference between A Deepfake And Shallowfake?, deepfakenow.com, APRIL 21, 2020, https://deepfakenow.com/what-is-the-difference-between-a-deepfake-and-shallowfake/.

[5] Nguyen, H.H., Yamagishi, J. and Echizen, I., 2019, May. Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2307-2311). IEEE.

[6] https://github.com/ondyari/FaceForensics

[7] https://www.kaggle.com/c/deepfake-detection-challenge/data

[8] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), pp.1499-1503.

[9] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[10] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K., 2017. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

[11] Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network