Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Face Deepfake Detection Model Based on Machine Learning

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By

Duha Amir Sultan Abd_Al Qadir

Supervised by

Prof. Dr. Laheeb Mohammad Ibrahim Mohammad

2024 A.D. **1446 A.H.**

Abstract

The revolutionary advancement and development of deepfake technology has led to the emergence of a new era of dazzling and anxiety in the world of digital media. Deepfake, that fed by advanced algorithms of machine learning have showed a surpassing ability in manipulating and fabricating digital content, so that the differences between the real and the fake ones faded away. As these fabricated media techniques progressed, the possibility for using them as a mean for fraud and cheating has increased and this may cause a harm to many people.

In response to the raising menace, several researches and studies have been made to understand how deepfake works and how to detect these generated fake medias. In this thesis, a new model for fake video detection has been considered, named as Video_Audio Deepfake Detection (VADD). The model took advantages of the visual and audial features of the video, based on the consideration that a well-fabricated video should be manipulated in both visual and audial aspects.

VADD works as on two major strategies, Unimodal: two separated models were designed for checking videos based on their visual and audial features separately. Multimodal: the two separated unimodals were combined together to produce a single video_audio multimodal for deepfake detection, enabling checking the video through both visual and audial features.

First of all, each video was braked into 10, 20, and 30 frames with 0, 1, and 3 intervals between frames. Faces were detected using MTCNN algorithm. Facial features were extracted using either of two models:
FaceNet+PCA+HeadPose estimation,
VGGFace+PCA+HeadPose estimation,
while the audio features were extracted using the mel-frequency cepstral coefficients (MFCC) machine learning function.

The models are trained on two types of datasets. These datasets are: FaceForensics++(FF++), which contains 1000 real and 1000 fake of silent videos, and FakeAVCeleb which contains 500 real and 500 fake of speaking videos.

In all cases, a LazyClassifier was used to determine whether the video was fake or not. The best obtained accuracy values were: 0.9309462 for the Visual model using VGGFace+PCA+HeadPose estimation applied on 20 frames per video with 1_interval taken between frames extracted from

the FF++ dataset and 0.86 on the FakeAVCeleb dataset. 0.9949748 for the Audial model using the MFCC function with 60 features per audio. 0.9698492 for multimodal applied on the FakeAVCeleb dataset. The combination of the audial and visual models in a single multimodal enhances the overall deepfake detection accuracy.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

نموذج كشف التزييف العميق للوجه بالاعتماد على التعلم الالي

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

ضحى عامر سلطان عبدالقادر

بإشراف
أ. د. لهيب محمد إبراهيم محمد

١٤٤٦ هـ ٢٠٢٤ م

الخلاصة

لقد أدى الظهور والتطور السريع لتكنولوجيا التزييف العميق إلى ظهور حقبة جديدة من الانبهار والقلق في عالم الوسائط الرقمية. لقد أظهرت تقنية Deepfakes ، التي تغذيها خوارزميات التعلم الآلي المتقدمة، قدرة غير مسبوقة على التلاعب بالمحتوى الواقعي وتوليفه، مما يؤدي إلى اخفاء الحدود بين الواقع والتزييف. وبما أن تقنيات الوسائط المزيفة هذه أصبحت أكثر تعقيدا، فقد تصاعدت احتمالات سوء الاستخدام والخداع، مما قد يشكل ضررا على الكثير من الناس. واستجابة للتهديد المتصاعد، تم إجراء العديد من الدراسات لفهم كيفية عمل التزييف العميق وكيفية اكتشاف هذه الوسائط المزيفة. في هذه الدراسة، تم اقتراح نموذج جديد أطلق عليه اسم VADD، لكشف الفيديو المزيف والذي يستغل الميزات المرئية والسمعية للفيديو. VADD يعمل وفقا لاستراتيجيتين رئيسيتين, Unimodal: تم تصميم نموذجين منفصلين لفحص مقاطع الفيديو بناءً على ميزاتها المرئية والسمعية بشكل منفصل. Multimodal: تم دمج النموذجين المنفصلتين معًا الوسائط المتعددة: لإنتاج نموذج متعدد يعتمد على الميزات المرئية والصوتية معا لاكتشاف التزييف العميق.

في البدء، تم تقسيم كل فيديو إلى ١٠ و٢٠ و٣٠ إطارًا مع فواصل زمنية ٠ و١ و٣ بين الإطارات. تم اكتشاف الوجوه باستخدام خوارزمية MTCNN. ميزات الوجه استخرجت باستخدام أحد النموذجين:FaceNet+PCA+HeadPose estimation،

VGGFace+PCA+HeadPose estimation

بينما تم استخراج ميزات الصوت باستخدام وظيفة التعلم الآلي لمعاملات التردد الرأسي (MFCC) تم تدريب النماذج على نوعين من مجموعات البيانات. مجموعات البيانات هذه هي: FaceForensics++(FF++)، التي تحتوي على ١٠٠٠ مقطع فيديو حقيقي و١٠٠٠ مقطع فيديو مزيف من الفيديوهات الصامتة، و FakeAVCeleb التي تحتوي على ٥٠٠ مقطع فيديو حقيقي و٥٠٠ مقطع فيديو مزيف من الفيديوهات الناطقة.

في جميع الحالات السابقة، استخدم LazyClassifier لتحديد ما إذا كان الفيديو مزيفا أم لا.

لتقييم النموذج استخدمت أربعة مقاييس تقييم، وهي: Accuracy, , Precision, F1_Score. Recallأفضل قيم الدقة (Accuracy) التي تم الحصول عليها هي: ٠,٩٣٠٩٤٦٢ للنموذج المرئي باستخدام VGGFace+PCA+HeadPose estimation المطبق على ٢٠ إطارًا لكل فيديو مع فاصل 1_interval مأخوذ بين الإطارات المستخرجة من مجموعة بيانات FF++ و٠,٨٦ على مجموعة بيانات FakeAVCeleb. أعلى قيمة للدقة هي ٠,٩٩٤٩٧٤٨ للنموذج الصوتي باستخدام دالة MFCC مع ٦٠ ميزة مستخرجة من كل ملف صوتي. أعلى قيمة للدقة لنموذج الوسائط المتعددة

هي ٠,٩٦٩٨٤٩٢ المطبقة على مجموعة بيانات FakeAVCeleb. أدى الجمع بين النماذج الصوتية والمرئية في وسائط متعددة واحدة إلى تعزيز الدقة الشاملة للكشف عن التزييف العميق.