Emotion Recognition Using MFCC, Spectrograms and Deep Neural Networks

Almotaz Bellah Adnan Fadel [*]
Dr. Dima Mufti Al-Chawafea**

(Received 2 / 7 / 2025. Accepted 23 / 11 / 2025)

□ ABSTRACT □

Speech recognition technologies are essential modern tools, with various systems differing in feature extraction and classification methods. This research examines multiple feature extraction approaches, highlighting the potential benefits of combining them to improve accuracy.

The proposed method was developed in three stages, each using a distinct method. The first stage employed MFCC (Mel-Frequency Cepstral Coefficients), achieving 78.62% accuracy. However, MFCC alone may lose important temporal and visual cues by segmenting signals into short frames, limiting emotion detection.

In the second stage, spectrograms were used، enhancing emotion recognition and achieving 93.20% accuracy by preserving energy distribution across frequencies. The third stage applied Feature-Level Fusion, combining MFCC and spectrogram outputs. This hybrid model, evaluated using a Random Forest classifier, reached a 97% accuracy rate, with F1-score, Precision, and Recall also at 97%.

The results show that fusing acoustic and visual representations significantly improves performance compared to individual models. Our proposed approach demonstrates superior effectiveness in emotion recognition from speech.

Keywords: Deep Learning, Emotion Recognition, Feature Extraction, Spectrograms, Neural Networks.

[*] **Master's student, Computer Engineering, Faculty of Electrical and Electronic Engineering, Aleppo University, Aleppo, Syria, Email**: motazfadel18@gmail.com
** **Assistant Professor, Computer Engineering, Faculty of Electrical and Electronic Engineering, Aleppo University, Aleppo, Syria, Email:** dima.mufti@gmail.com

التعرف على العواطف باستخدام MFCC والمخططات الطيفية والشبكات العصبونية العميقة

المعتز بالله عدنان فاضل*

د. ديما مفتي الشوافعة**

(تاريخ الإيداع 2 / 7 / 2025. قُبِل للنشر في 23 / 11 / 2025)

□ ملخّص □

تُعد تقنيات التعرف على الكلام من الأدوات الحديثة الأساسية، حيث تختلف الأنظمة المستخدمة فيها من حيث طرق استخراج السمات وأساليب التصنيف. يركّز هذا البحث على دراسة عدة منهجيات لاستخلاص السمات، مع التركيز على الفوائد المحتملة لدمج هذه الطرق بهدف تحسين الدقة.

تم تطوير الطريقة المقترحة على ثلاث مراحل، كل منها اعتمد على أسلوب مختلف عن الاخر. في المرحلة الأولى، تم استخدام خوارزمية MFCC (Mel-Frequency Cepstral Coefficients) وحققت دقة تحقق بلغت 78.62%. قد يؤدي الاعتماد على MFCC وحدها إلى فقدان مؤشرات زمنية وبصرية مهمة، نتيجة لتقسيم الإشارة الصوتية إلى إطارات زمنية قصيرة، مما يحد من فعالية الكشف عن العواطف.

في المرحلة الثانية، تم استخدام الصور الطيفية(Spectrograms) ، مما ساهم في تحسين التعرف على العواطف عبر الحفاظ على توزيع الطاقة عبر الترددات، وحققت دقة بلغت 93.20%. أما المرحلة الثالثة فقد اعتمدت على دمج السمات(Feature-Level Fusion) بين مخرجات MFCC والصور الطيفية، وتم تقييم النموذج المهجّن باستخدام مصنفRandom Forest ، وحقق دقة وصلت إلى 97%، بالإضافة إلى قيمF1-score، Recall ، بنسبة 97%.

تُظهر النتائج أن دمج التمثيلات الصوتية والبصرية يعزز الأداء بشكل كبير مقارنة باستخدام كل نموذج على حدة. وتبرهن المنهجية المقترحة على فعاليتها العالية في التعرف على العواطف من خلال الكلام.

الكلمات المفتاحية : التعلم العميق، التعرف على العاطفة، استخراج السمات، الصور الطيفية، الشبكات العصبونية.

* طالب ماجستير، هندسة الحواسيب، كلية الهندسة الكهربائية والإلكترونية، جامعة حلب، حلب، سوريا.
Email: motazfadel18@gmail.com

** أستاذ مساعد، هندسة الحواسيب، كلية الهندسة الكهربائية والإلكترونية، جامعة حلب، حلب، سوريا.
Email: dima.mufti@gmail.com

مقدمة:

تشهد تقنيات الحوسبة تطوراً متسارعاً جعلها عنصراً أساسياً في مختلف مجالات الحياة، وأصبح نجاح الأنظمة الذكية مرهوناً بقدرتها على التفاعل الطبيعي مع الإنسان .ويُعد الصوت من أهم وسائل هذا التفاعل لكونه أداة تواصل فعالة تحمل في طياته دلالات لغوية وعاطفية في آن واحد.

هذا الاهتمام دفع الشركات العالمية الكبرى مثل Google و Appleو Amazonإلى تطوير أنظمة تفاعلية صوتية واسعة الاستخدام، مثل Google Assistant و Siriو Alexa، والتي أثبتت فعاليتها في التطبيقات العملية. ومع ذلك، لا تزال الحاجة قائمة إلى رفع مستوى فهم الأنظمة للمشاعر الإنسانية، بما يضمن تفاعلاً أكثر دقة وواقعية[1] .

في هذا السياق، برز مجال التعرف على العواطف من الكلام كأحد أبرز الاتجاهات الحديثة في أبحاث الذكاء الاصطناعي. فقد أتاح التقدم في تقنيات التعلم العميق واستخلاص السمات الصوتية تطوير نماذج قادرة على تحليل الإشارات الصوتية بدقة متزايدة، مما مهد لاستخدامها في تطبيقات متنوعة تشمل المساعدات الذكية، التعليم، والرعاية الصحية.

أهمية البحث وأهدافه:

مشكلة البحث:

لا تزال العديد من النماذج في مجال التعرف على العواطف الصوتية تواجه قصور في الدقة عند تصنيف المشاعر.

ويظهر هذا الضعف بشكل خاص عند التعامل مع المشاعر المتقاربة في خصائصها الصوتية، مثل الحزن والخوف أو السعادة والمفاجأة، حيث يؤدي التشابه في بعض السمات إلى حدوث خلط أو التباس في التصنيف.

هذا القصور ينعكس سلباً على كفاءة أنظمة التعرف على العواطف ويحد من قدرتها على توفير تفاعل أكثر دقة وواقعية بين الإنسان والآلة، الأمر الذي يستلزم البحث عن نهج بديل يعالج هذه الثغرات.

هدف البحث:

الهدف الأساسي من هذا البحث هو رفع دقة أنظمة التعرف على العواطف الصوتية، مع التركيز على تحسين قدرة النماذج على التمييز بين المشاعر المتداخلة .ولتحقيق ذلك، يقترح البحث استخدام نهج تكاملي يقوم بدمج أكثر من طريقة لاستخراج السمات الصوتية، مما يتيح معالجة القيود المرتبطة بالاعتماد على طريقة واحدة فقط.

ويتمثل الإطار العام للبحث في:

1. دراسة تأثير دمج طرق متعددة مثل MFCC و Spectrogramعلى دقة التعرف على العواطف.
2. تصميم نموذج هجين يعتمد على دمج السمات الصوتية والبصرية.
3. تحسين قدرة النظام على الفصل بين المشاعر المتقاربة (مثل الخوف والحزن أو السعادة والمفاجأة) والحد من التداخل في عملية التصنيف.

أهمية البحث:

تتجلى أهمية هذا البحث في تحسين دقة التعرف على العواطف الصوتية ومعالجة صعوبة التمييز بين المشاعر المتقاربة.

طرائق البحث ومواده:

مجموعات المعطيات المستخدمة:

تم الاعتماد في هذه الدراسة على مجموعتين أساسيتين من مجموعات المعطيات الخاصة بتصنيف المشاعر الصوتية، وذلك لاختبار النماذج المقترحة على مجموعات معطيات مختلفة من حيث الحجم والنوع:

(أ) مجموعة معطياتTESS :

تم استخدام مجموعة معطيات TESS(Toronto Emotional Speech Set) ، التي تحتوي على 2800عينة صوتية تغطي 7مشاعر (الغضب، الحزن، السعادة، الخوف، المفاجأة، الاشمئزاز، الحياد). تم تسجيل العينات من قبل ممثلين (تتراوح أعمارهم بين 26 و64 عاماً)، مما يسمح باستخلاص سمات دقيقة[2] .

(ب) مجموعة معطياتIEMOCAP :

تم استخدام مجموعة معطيات IEMOCAP (Interactive Emotional Dyadic Motion Capture Database)، وهي واحدة من أكثر مجموعات المعطيات استخداماً في أبحاث التعرف على المشاعر الصوتية. تحتوي على حوالي 12ساعة من المحادثات التمثيلية والحوارية، مسجلة من قبل 10ممثلين محترفين (5 ذكور و5 إناث). تشمل المعطيات مقاطع scripted(مسجلة نصياً) و improvised(ارتجالية)، وتتضمن حوالي 10,000 مقطع صوتي تمت عنونتها (annotated) بواسطة مقيمين بشريين.

المشاعر الأساسية المعتمدة مجموعة المعطيات هي أربعة مشاعر: الغضب(Angry) ، السعادة(Happy) ، الحزن (Sad)، والحياد (Neutral) .

تتميز IEMOCAP بتنوعها الكبير مقارنةً بـTESS ، إذ تضم حوارات طبيعية ومقاطع ارتجالية بالإضافة إلى المسجلة نصياً، مما يجعلها أكثر تحدياً وواقعية في تقييم النماذج المقترحة[3].

الأدوات البرمجية:

تم تنفيذ النموذج المقترح باستخدام لغة Python 3.10 ضمن بيئة Jupyter Notebook على منصةGoogle Colab بيئة العتاد تضمنت معالج Intel Xeon، ذاكرة RAMبسعة GB12، وبطاقة رسومية NVIDIA Tesla T4 GPU مما مكّن من تسريع عملية التدريب.

الاعتماد على مكتبات متخصصة في معالجة الإشارات الصوتية وبناء الشبكات العميقة، مثل:

- Librosa :لاستخراج الميزات الصوتية(MFCC) .
- OpenCVو Matplotlib :لتوليد الصور الطيفية(Spectrograms) .
- Tensor Flow/Keras :لبناء وتدريب الشبكات العميقة.
- Scikit-learn :لمعالجة المعطيات وتقييم النماذج.

1-الدراسات المرجعية:

1-1-أهم الدراسات التي اعتمدت على خوارزمية ال MFCC:

بدأت المحاولات مع Iqbal et.alحيث تم استخدام معاملات MFCC مع مصنف Support Vector (SVM) Machineللتعرف على العواطف الصوتية.

وقد اختبر الباحثون النموذج على مجموعتي معطيات مختلفتين:

- في مجموعة معطيات TESS، التي تضمنت أربع فئات عاطفية (الغضب، السعادة، الاشمئزاز، والمفاجأة)، حقق النموذج دقة مرتفعة بلغت 97%. ويعود ذلك إلى أن هذه الفئات تتميز باختلافات صوتية واضحة نسبياً، مما سهّل عملية التصنيف.
- في المقابل، عند اختبار النموذج على مجموعة معطيات Interactive Emotional Dyadic Motion (IEMOCAP) Capture، التي شملت ثلاث فئات (الغضب، الحزن، والحياد)، انخفضت الدقة إلى 86%. ويُعزى هذا الانخفاض إلى تقارب بعض الخصائص الصوتية بين الفئات، خاصة بين الغضب والحزن، مما أدى إلى ظهور حالات من الالتباس في التصنيف.

تُظهر هذه النتائج أن دقة النماذج القائمة على MFCC و SVMقد تكون عالية في حالة المشاعر المتباعدة صوتياً، لكنها تواجه صعوبة في التمييز بين المشاعر المتقاربة، الأمر الذي يبرز الحاجة إلى تقنيات أكثر شمولية للتعامل مع هذه الحالات كما برزت مشكلة Overfittingنتيجة صغر حجم المعطيات المستخدمة[4].

في محاولة لتجاوز هذه المشكلة، قام Choudharyوزملاؤه بدمج معاملات MFCCمع نماذج عميقة Convolutional Neural Network (CNN) و Long Short-Term Memory (LSTM) و Gated Recurrent Unit (GRU)، بهدف تحسين دقة التعرف على العواطف الصوتية. تم اختبار النموذج على مجموعتي معطياتTESS :التي تضمنت سبع فئات عاطفية (الغضب، السعادة، الحزن، الحياد، الاشمئزاز، الخوف، المفاجأة)، و Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS)التي شملت ثماني فئات بإضافة فئة الهدوء. حقق النظام دقة بلغت 97.1%على TESS، بينما تراجعت الدقة إلى 72.2%على RAVDESS، حيث أظهر الباحثون وجود خلط ملحوظ بين المشاعر المتقاربة مثل الخوف والمفاجأة وضعفاً واضحاً في فئة الاشمئزاز، إضافة إلى وجود خلط بين مشاعر مثل الخوف والمفاجأة .كما أبرزت النتائج مشكلة ضعف التعميم وتراجع الأداء عند الانتقال من مجموعة معطيات صغيرة إلى أخرى أكبر وأكثر تنوعاً، مما يوضح أن الشبكات العميقة رغم فعاليتها ما تزال تعاني من تحديات مرتبطة بتمييز المشاعر المتقاربة وحجم المعطيات المتاح[5] .

لاحقاً، اقترح Sarala Padiوزملاؤه نهجاً يعتمد على توسيع المعطيات متعدد النوافذ (Multi-Window Data Augmentation)، حيث يتم تقسيم الإشارة الصوتية الواحدة إلى عدة نوافذ زمنية بأطوال مختلفة بدلاً من الاكتفاء بنافذة واحدة ثابتة لتوليد عينات إضافية بغرض تحسين تدريب النماذج العميقة في التعرف على العواطف الصوتية. جرى تقييم هذا النهج على ثلاث مجموعات معطيات معياريةIEMOCAP : (أربع فئات: الغضب، السعادة، الحياد، الحزن)، وSurrey Audio-Visual Expressed Emotion (SAVEE) (ست فئات: الغضب، السعادة، الحزن، الحياد، الخوف، الاشمئزاز)، وRAVDESS (ست فئات بعد دمج الحياد والهدوء). وأظهرت النتائج تحسناً نسبياً مقارنة بالنوافذ الفردية، حيث بلغت الدقة غير الموزونة (UA) نحو 70% علىSAVEE ، لكنها ظلت محدودة على مجموعات معطيات أكبر مثلRAVDESS ، كما أبرز الباحثون صعوبة التمييز بين بعض الفئات المتقاربة صوتياً، مثل الحزن والحياد أو السعادة والحياد مما أدى إلى التباس في التصنيف[6] .

وفي محاولة أخرى، اعتمد Ishaqوزملاؤه على شبكة التحويل الزمني Temporal Convolutional Network(TCN).تم اختبار النموذج على مجموعة معطياتIEMOCAP (أربع فئات: الغضب، السعادة، الحزن، الحياد) وEmotional Database (EMO-DB) (سبع فئات: الغضب، السعادة، الحزن، الحياد، الخوف، الاشمئزاز، الملل). وقد حقق النموذج دقة بلغت 80.84%على IEMOCAP و 92.31%علىEMO-DB ، وهو تحسن مقارنة

بالـ CNN وLSTM.ومع ذلك، أظهرت النتائج استمرار وجود التباس بين بعض المشاعر المتقاربة، حيث تداخلت فئة السعادة مع فئات أخرى فيIEMOCAP ، كما واجه النموذج صعوبة نسبية في التمييز بين الخوف والاشمئزاز في EMO-DB، مما يشير إلى أن مشكلة المشاعر المتقاربة ما تزال قائمة رغم التحسن الملحوظ في الأداء[7] .

أخيراً، اعتمد Ouyangعلى دمج معاملات MFCCمع نموذج CNN-LSTM، واختبر الأداء على مجموعة معطيات مدمجة من SAVEEو RAVDESSتضمنت سبع عواطف (الغضب، السعادة، الحزن، الحياد، الاشمئزاز، الخوف، المفاجأة). بلغت الدقة الكلية للنموذج 61.07%فقط، حيث تميزت فئتا الغضب (75.3%) والحياد (71.7%) بوضوح نسبي، بينما واجه النموذج صعوبة في تمييز العواطف السلبية المتقاربة مثل الحزن والخوف أو الاشمئزاز والغضب، إضافة إلى التباس المفاجأة مع السعادة أو الخوف. وأبرزت هذه النتائج محدودية قدرة النموذج على الفصل بين المشاعر المتقاربة، مما يدعم الحاجة إلى نماذج أكثر قوة مثل Attentionأو Transformers [8].

1-2-أهم الدراسات التي اعتمدت على الصور الطيفية:

اتجه Zhangوزملاؤه إلى استخدام الصور الطيفية بدلاً من معاملاتMFCC ، وقدموا شبكة Fully Convolutional Network (FCN)مدعومة بآلية Attentionلتصنيف العواطف الصوتية. تم اختبار النموذج على مجموعة معطيات IEMOCAP (أربع فئات: الغضب، السعادة، الحزن، الحياد)، وحقق دقة بلغت 70.4%، وهو ما يمثل تحسناً عن الطرق التقليدية. ومع ذلك، أظهرت النتائج استمرار الالتباس بين بعض المشاعر المتقاربة مثل السعادة والحياد أو الغضب والحزن، مما يشير إلى أن آلية Attention حسّنت الأداء لكنها لم تحل مشكلة التشابه الصوتي بشكل كامل[9] .

قدّم Mustaqeemوزملاؤه منهجية تقوم على تقسيم الملف الصوتي إلى مقاطع صغيرة، ثم اختيار الأهم منها باستخدام التجميع العنقودي، وتحويلها إلى مخططات طيفية (Spectrograms) قبل تمريرها إلى شبكة ResNet-101لاستخلاص السمات. تم تقييم النموذج على ثلاث مجموعات معطيات:

IEMOCAP (أربع فئات: الغضب، السعادة، الحزن، الحياد) ، بدقة بلغت 72.25%، مع خلط واضح بين السعادة والحياد.

EMO-DB(سبع فئات: الغضب، السعادة، الحزن، الحياد، الخوف، الاشمئزاز، الملل) ، بدقة بلغت 85.57%، حيث تميّزت فئات مثل الغضب والخوف بدقة عالية، بينما التبست السعادة مع بعض الفئات الأخرى.

RAVDESS(سبع فئات بعد دمج الحياد والهدوء: الغضب، السعادة، الحزن، الحياد، الخوف، الاشمئزاز، المفاجأة) ، بدقة بلغت 77.02%، مع استمرار الالتباس بين الحزن والحياد ، إضافة إلى صعوبة في تمييز السعادة.

أظهرت النتائج أن النموذج كان قوياً في تمييز بعض المشاعر البارزة مثل الغضب والخوف، لكنه ظل يعاني من صعوبة في الفصل بين المشاعر المتقاربة، خاصة السعادة–الحياد والحزن–الحياد، مما يبرز أن مشكلة التشابه العاطفي ما تزال قائمة رغم التحسن في الأداء العام[10] .

اعتمدPadiوزملاؤه على نهج نقل التعلم باستخدام شبكة ResNet34مع تطبيق تقنيات Spectrogram Augmentationعبر أقنعة زمنية وترددية عشوائية لتوليد معطيات تدريبية إضافية. أُجريت ثلاث تجارب على مجموعة معطيات :الأولى IEMOCAPشملت أربع فئات (الغضب، السعادة، الحياد، الحزن) وحققت دقة بلغت 66.02%، فيما استبدلت التجربة الثانية فئة السعادة بفئة الحماسة وبلغت الدقة 65.62%، أما التجربة الثالثة فقد دمجت السعادة والحماسة معاً لتصل الدقة إلى .63.61%ورغم التحسن الناتج عن استخدام أسلوب التوسيع، أظهرت

النتائج استمرار الالتباس بين المشاعر المتقاربة، خاصة السعادة والحياد، إضافة إلى بعض الخلط بين الغضب والسعادة، مما يؤكد أن الفصل الدقيق بين هذه الفئات ما يزال يمثل تحدياً قائماً[11].

اعتمد Akinpeluوزملاؤه على منهجية نقل التعلم العميق (Deep Transfer Learning) باستخدام شبكة VGGNet لاستخراج السمات من المخططات الطيفية الصوتية. ولتحسين أداء النموذج، استُخدم أسلوب اختيار السمات(Neighborhood Component Analysis – NCA) لاختيار السمات الأكثر ارتباطاً بالتصنيف، ثم جرى تمريرها إلى مصنفي SVMو(Multi-Layer Perceptron) MLP وقد تم تقييم المنهجية على مجموعة معطيات:

TESS : سبع فئات (الغضب، الاشمئزاز، الخوف، السعادة، الحزن، المفاجأة، الحياد) وحقق النموذج دقة بلغت97.2%.

EMO-DB :سبع فئات (الغضب، الاشمئزاز، الخوف، السعادة، الحزن، الحياد، الملل) وحقق دقة بلغت 94.3%.

مجموعة معطيات مدمجة :(TESS + EMO-DB) ست فئات بعد الدمج، مع أداء جيد أيضاً.

ورغم هذه النتائج المرتفعة، أظهرت الدراسة استمرار الالتباس بين بعض العواطف المتقاربة، مثل السعادة والحياد أو السعادة والمفاجأة، إضافة إلى ضعف نسبي في تمييز الخوف مقارنة بالعواطف الأخرى. وتشير هذه النتائج إلى أن النموذج كان قادراً على الفصل بدقة بين العواطف المتباعدة، لكنه لم يتغلب بشكل كامل على تحدي التمييز بين المشاعر المتقاربة[12] .

قام Penumajjiوزملاؤه بتطوير نظام للتعرف على العواطف من الكلام باستخدام Mel-Spectrogramsمع بنية CNN، وذلك على مجموعة معطيات RAVDESSشملت التجارب سبع فئات عاطفية بعد دمج الحياد مع الهدوء، وهي :الغضب، السعادة، الحزن، الخوف، المفاجأة، الاشمئزاز، الحياد/الهدوء .وقد بلغت الدقة الكلية 68.88%، حيث كان الأداء جيداً في المشاعر السلبية مثل الغضب، الحزن، والاشمئزاز، بينما ظهر ضعف ملحوظ في المشاعر الإيجابية مثل السعادة والمفاجأة .وأظهرت النتائج التباساً واضحاً بين السعادة والمفاجأة، إضافة إلى بعض الخلط بين الحزن والحياد/الهدوء، مما يؤكد أن الاعتماد على CNN مع Mel-Spectrograms يحقق أداء مقبولاً في العواطف الواضحة لكنه يظل محدوداً عند التمييز بين المشاعر المتقاربة[13] .

1-3-اهم الدراسات التي اعتمدت على الدمج بين الMFCC وSpectogram :

لتجاوز قصور استخدام MFCC أو Spectrogram منفردة، ظهرت دراسات دمج الميزات.

بدأ Wangوزملاؤه نموذجاً يعتمد على بنية (Dual-Stream LSTM (DS-LSTM، حيث جرى دمج ميزات MFCC مع طيفين مختلفين من Mel-Spectrogramsيركز أحدهما على الدقة الزمنية والآخر على الدقة الترددية. تمت المعالجة عبر شبكة CNN لاستخلاص السمات الطيفية، تلتها آلية DS-LSTM للتعامل مع التسلسلين بشكل متوازٍ، بينما خُضعت ميزات MFCC لمعالجة مستقلة بواسطةLSTM ، ثم جرى دمج النواتج في المرحلة النهائية. وقد اختُبرت المنهجية على مجموعة معطيات IEMOCAPبأربع فئات (الغضب، السعادة، الحزن، الحياد)، وحققت دقة بلغت (WA) 72.7%و(UA) 73.3%، متفوقة على النماذج الأساسية. ورغم هذا التحسن، أظهرت النتائج استمرار الالتباس بين بعض المشاعر المتقاربة مثل السعادة والحياد والحزن والحياد، بينما تميزت الفئات الأكثر وضوحاً كـ الغضب بدقة أعلى في التصنيف[14].

طور Yurun He et.alإطار Cross-Attention Transformerلدمج Raw Waveform+MFCC+ Spectrogramاختُبرت المنهجية على مجموعة معطيات IEMOCAPبأربع فئات (الغضب، السعادة، الحزن،

الحياد)، وحققت دقة بلغت (WA) 73.8%و(UA) 74.25%. ورغم هذا التحسن، أظهرت النتائج استمرار الالتباس بين بعض المشاعر المتقاربة مثل الحياد والسعادة أو الحياد والحزن، بينما كانت العواطف الأكثر بروزاً كـ الغضب أسهل في التمييز[15] .

قدّم Basha وزملاؤه إطاراً يعتمد على Ensemble Attribute Selection (EAS)لدمج السمات المستخرجة من MFCCو Spectrograms، ومن ثم تمريرها إلى نموذج CNN–LSTMمدعوم بآلية انتباه. تم اختبار النموذج على ثلاث مجموعات معطيات هي RAVDESSو TESS وSAVEE، شاملةً ثماني عواطف (الحياد، الهدوء، السعادة، الحزن، الغضب، الخوف، الاشمئزاز، المفاجأة). وقد حقق النظام دقة بلغت 87.08%، متفوقاً على النماذج الفردية (CNNوحده 81% و LSTMوحده 84%)ورغم هذا التفوق، أظهرت النتائج استمرار صعوبة التمييز بين بعض المشاعر المتقاربة، خصوصاً بين السعادة والحياد، في حين تميزت الفئات الأكثر وضوحاً مثل الغضب والمفاجأة بدقة أعلى في التصنيف[16] .

جدول 1: مقارنة بين أهم الدراسات المرجعية في مجال التعرف على العواطف من الكلام.

المؤلف	مجموعات المعطيات والفئات	المنهجية	الدقة	نقاط القوة	التحديات (الخلط/التشابه)
Iqbal et al. (2021)	TESS (4 فئات)؛ IEMOCAP (3 فئات)	MFCC + SVM	97% (TESS)، 86% (IEMOCAP)	أداء قوي مع الفئات المتباعدة	التباس بين الغضب والحزن؛ لصغر حجم Overfitting المعطيات
Choudhary et al. (2022)	TESS (7)؛ RAVDESS (8)	MFCC + CNN/LSTM/GRU	97.1% (TESS) ، 72.2% (RAVDESS)	أداء مرتفع على مجموعات معطيات صغيرة	خلط بين الخوف والمفاجأة؛ ضعف في الاشمئزاز؛ ضعف التعميم
Sarala Padi et al. (2022)	IEMOCAP (4)؛ SAVEE (6)؛ RAVDESS (6)	Multi-Window Data Augmentation	UA ≈70% (SAVEE)	تحسين نسبي مقابل النوافذ الفردية	التباس بين الحزن والحياد؛ وأحيانًا السعادة والحياد
Ishaq et al. (2023)	IEMOCAP (4)؛ EMO-DB (7)	Temporal Convolutional Network (TCN)	80.84% (IEMOCAP) ، 92.31% (EMO-DB)	CNN/LSTM أفضل من في الاعتمادات الزمنية	التباس بين السعادة والحياد؛ خلط بين الخوف والاشمئزاز
Ouyang (2023)	SAVEE + RAVDESS (7)	MFCC + CNN-LSTM	61.07%	وضوح نسبي للغضب (75.3%) والحياد (71.7%)	ضعف في الحزن والخوف؛ خلط بين السعادة والمفاجأة/الخوف
Zhang et al. (2019)	IEMOCAP (4)	FCN + Attention مع Spectrograms	70.4% (WA)، 63.9% (UA)	تحسن عن الطرق التقليدية	استمرار الالتباس بين السعادة والحياد أو الغضب والحزن
Mustaqeem et al. (2020)	IEMOCAP (4)؛ EMO-DB (7)؛ RAVDESS (7)	ResNet-101 + Spectrogram Clustering	72.25% ، 85.57% ، 77.02%	تمييز قوي للغضب والخوف	التباس بين السعادة والحياد؛ وأحيانًا الحزن والحياد
Padi et al. (2021)	IEMOCAP (4)	ResNet34 + Spectrogram Augmentation	66.02% ، 65.62% ، 63.61%	تحسين عبر Augmentation	التباس بين السعادة والحياد؛ خلط بين الغضب والسعادة/الحماسة
Akinpelu et al. (2022)	TESS (7)؛ EMO-DB (7) ؛ مجموعة (6) معطيات مدمجة	VGGNet (VGGish) + NCA + SVM/MLP	97.2% (TESS) ، 94.3% (EMO-DB)	دقة عالية جدًا للفئات المتباعدة	التباس بين السعادة والحياد أو السعادة والمفاجأة؛ ضعف في الخوف

المؤلف	مجموعات المعطيات والفئات	المنهجية	الدقة	نقاط القوة	التحديات (الخلط/التشابه)
Penumajji et al. (2025)	RAVDESS (7: دمج الحياد والهدوء)	Mel–Spectrogram + CNN	68.88%	أداء جيد مع الغضب والحزن والاشمئزاز	خلط بين السعادة والمفاجأة؛ وأحيانًا الحزن والحياد/الهدوء
Wang et al. (2020)	IEMOCAP (4)	DS–LSTM + MFCC + Spectrograms	73.3% (WA)، 72.7% (UA)	دمج زمني وترددي مع MFCC	التباس بين السعادة والحياد والحزن والحياد
Yurun He et al. (2023)	IEMOCAP (4)	Cross–Attention Transformer (Raw + MFCC + Spectrogram)	73.8% (WA) ، 74.25% (UA)	تحسن عبر دمج متعدد المستويات	استمرار الالتباس بين الحياد والسعادة/الحزن
Basha et al. (2025)	RAVDESS + TESS + SAVEE (8)	EAS + CNN–LSTM + Attention	87.08%	LSTMو CNN تفوق على منفردين	:ضعف في العواطف الإيجابية السعادة والحياد

يتضح من استعراض الدراسات السابقة أن التطور من الأساليب التقليدية المعتمدة على السمات اليدوية مثل MFCCمع SVMإلى النماذج العميقة مثل CNNو LSTM، ثم إلى الأطر الأكثر تقدماً التي توظف Attentionو Transformers أو أساليب الدمج المتعدد، قد أسهم في رفع دقة أنظمة التعرف على العواطف الصوتية. ومع ذلك، أظهرت النتائج أن هذا التحسن ظل محدوداً عند التعامل مع مجموعة معطيات واسعة أو متنوعة، حيث يستمر الخلط بين المشاعر المتقاربة مثل السعادة والحياد أو السعادة والمفاجأة، وكذلك الحزن والحياد .كما أن الأداء العالي الذي تحقق في مجموعة صغيرة نسبياً مثل TESS و EMO–DBغالباً ما يتراجع في مجموعة أكبر وأكثر تعقيداً مثل IEMOCAP وRAVDESS، مما يعكس مشكلة ضعف التعميم .ومن هنا، يمكن القول إن الفجوة البحثية ما تزال قائمة وتتمثل في الحاجة إلى منهجيات أكثر تكاملاً تدمج بين أنواع متعددة من السمات وتستفيد من قدرات النماذج الحديثة على التقاط الفروق الدقيقة بين العواطف المتقاربة، وذلك بهدف تحسين دقة التمييز بين المشاعر المتقاربة ومعالجة مشكلة ضعف التعميم التي واجهتها الدراسات السابقة.

2–مقاييس الأداء:

1–الدقة (Accuracy):

دقة النموذج على مجموعة معطيات التدريب، تُحسب وفق المعادلة(1):

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

TP :الحالات الإيجابية التي صُنفت بشكل صحيح.

TN :الحالات السلبية التي صُنفت بشكل صحيح.

FP :الحالات السلبية التي صُنفت بالخطأ كإيجابية.

FN : الحالات الإيجابية التي صُنفت بالخطأ كسلبية.

2–دقة التحقق(Validation Accuracy) :

دقة النموذج على مجموعة معطيات التحقق(الاختبار) باستخدام نفس معادلة الدقة.

3–معدل الخطأ (Error Rate):

تمثل نسبة العينات المصنفة بشكل خاطئ، تُحسب وفق المعادلة(2):

$$\frac{FP+FN}{TP+TN+FP+FN} \quad (2)$$

4–الدقة الإيجابية (Precision):

تمثل النسبة المئوية من العينات التي تنبأ بها النموذج كإيجابية وكانت فعلًا إيجابية، تُحسب وفق المعادلة(3):

$$\frac{TP}{TP+FP} \qquad (3)$$

5-الاستدعاء (Recall):

يعكس مدى قدرة النموذج على اكتشاف كل العينات الإيجابية، تُحسب وفق المعادلة(4):

$$\frac{TP}{TP+FN} \qquad (4)$$

6-مقياس (F1-Score):

هو متوسط توافقي بين الدقة والاستدعاء، ويُستخدم عند وجود تفاوت بين الفئتين، تُحسب وفق المعادلة(5):

$$2\frac{\mathbf{Recall}\times\mathbf{Precision}}{\mathbf{Recall}+\boldsymbol{Precision}} \qquad (5)$$

3-المنهجية المقترحة(Proposed Methodology) :

تهدف المنهجية المقترحة إلى معالجة القصور الذي أظهرته الدراسات السابقة في التعرف على المشاعر الصوتية، خصوصاً في التمييز بين المشاعر المتقاربة مثل (السعادة-الحياد، الحزن-الحياد، السعادة-المفاجأة). وللتغلب على هذه التحديات، تم تصميم نموذج دمج Feature-Level Fusion.

3-1-(Feature-Level Fusion) :

1-المدخلات :تسجيلات خام لعواطف متعددة من مجموعة معطياتTESS .

2-Data Augmentation: زيادة العينات من 2800 إلى 8400 عبر إضافة ضوضاء، تغيير النغمة، وتمديد الإشارة.

3-استخراج السمات:

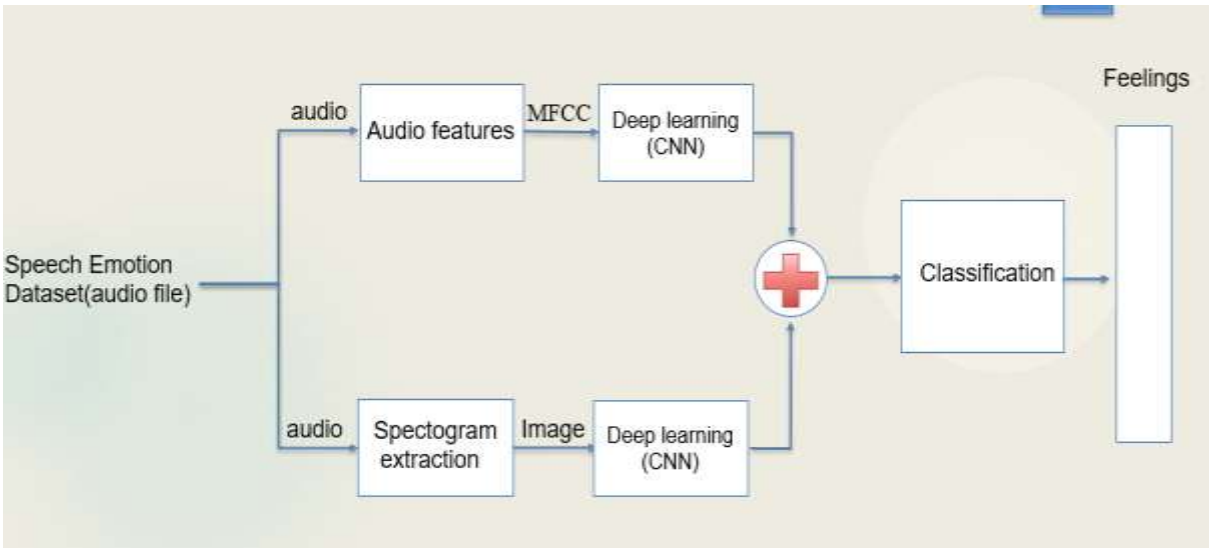
MFCCتُعالج بواسطة 1D-CNN لاستخلاص الأنماط الزمنية.

Spectrogram تُعالج بواسطة 2D-CNN لاستخلاص الأنماط الطيفية.

4-مرحلة الدمج :دمج السمات المستخرجة في متجه موحّد(Feature-Level Fusion) .

5-التصنيف النهائي :باستخدام Random Forest أوMLP .

6-الهدف :اختبار فعالية الدمج المبكر على مجموعة معطيات مختلفة الحجم.



الشكل (1): المخطط الصندوقي للمنهجية المقترحة

4-التنفيذ والنتائج (Implementation and results):

4-1-تصنيف المشاعر بالاعتماد على (Audio Features and MFCC):

تم استخراج سمات MFCC من ملفات مجموعة معطيات TESS (2800 عينة صوتية) باستخدام مكتبةLibrosa ، مع تحديد طول الإشارة بـ 2.5 ثانية. بعد تطبيقAugmentation ، تم رفع عدد العينات إلى 8400، وتم تقسيمها بنسبة 80% للتدريب و20% للاختبار.

تم بناء نموذج 1D-CNNلمعالجة مصفوفات MFCC (20 معاملات لكل عينة)، وتم التدريب على 30 دورة (Epochs)لضمان الاستقرار وتفادي فرط التعلم.

جدول 2: تقسيم مجموعة المعطيات إلى مجموعة تدريب (80%) ومجموعة اختبار (20%)

مجموعة المعطيات	X (السمات المستخرجة)	Y (التسميات Labels)
التدريب(80)%	(6300, 20, 1, 1)	(6300, 7)
الاختبار(20)%	(2100, 20, 1, 1)	(2100, 7)

X :يمثل السمات الصوتية (Features) المستخرجة من كل ملف صوتي باستخدام معاملاتMFCC .

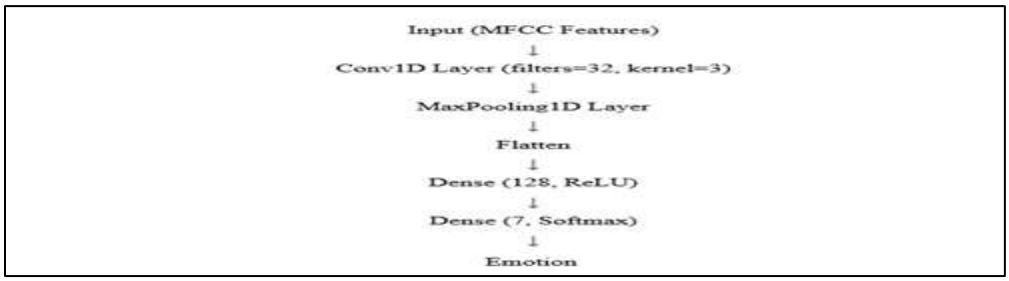
Y :يمثل التسميات (Labels) الخاصة بالعواطف المستهدفة (7 فئات).

(6300, 7) أو (2100, 7) تعني أن لدينا 6300 (أو 2100) عينة، وكل عينة مرتبطة بمتجه من 7 قيم يمثل الفئة العاطفية.

(6300, 20, 1, 1)أو(2100, 20, 1, 1)تعني أن كل عينة مُمثلة بـ 20معاملMFCC ، مع إضافة بُعدين إضافيين (1,1) لضمان توافق شكل المعطيات مع متطلبات شبكةCNN .

تم تحديد عدد معاملات MFCC بـ 20لأنه يحقق توازناً بين الاحتفاظ بالمعلومات الطيفية الأساسية وتقليل الأبعاد الحسابية. فاختيار عدد قليل جداً (مثل 12 أو 13) قد يؤدي إلى فقدان تفاصيل مهمة مرتبطة بالتعبير العاطفي، بينما زيادة العدد بشكل مبالغ فيه (مثل 40) قد تُدخل سمات غير ضرورية وتزيد من زمن التدريب. بناءً على ذلك، فإن 20 MFCCتُعتبر قيمة مثالية مستخدمة بكثرة في تطبيقات التعرف على الكلام والمشاعر الصوتية، خصوصاً مع مجموعات معطيات متوسطة الحجم مثلTESS ، حيث تساهم في تحسين الدقة وتفادي مشكلة الOverfitting .

تم بناء نموذج باستخدام الشبكة العصبونية الالتفافية(CNN) ، تم استخدام سمات MFCC المستخرجة من الإشارات الصوتية كمدخلاتٍ إلى نموذج CNN بتنسيق 1D كما هو موضح بالشكل (2):



الشكل (2): مخطط هيكلية الشبكة(1D-CNN) [17]

يعالج النموذج معطيات MFCC المُشكّلة على هيئة مصفوفات 1D، يسمح هذا التدرج في المعالجة باستخلاص تمثيلات عالية المستوى، والتي تُستخدم لاحقاً للتصنيف الدقيق للمشاعر، تم التدريب على EPOCH(30) .

تم اختيار (30) epoch بعد تجارب أولية مع قيم مختلفة (10، 20، 30، 50)، حيث لم يُتحسَّن معيار التحقق (Val_accuracy)بشكلٍ كبير بعد الـ 30 EPOCH.

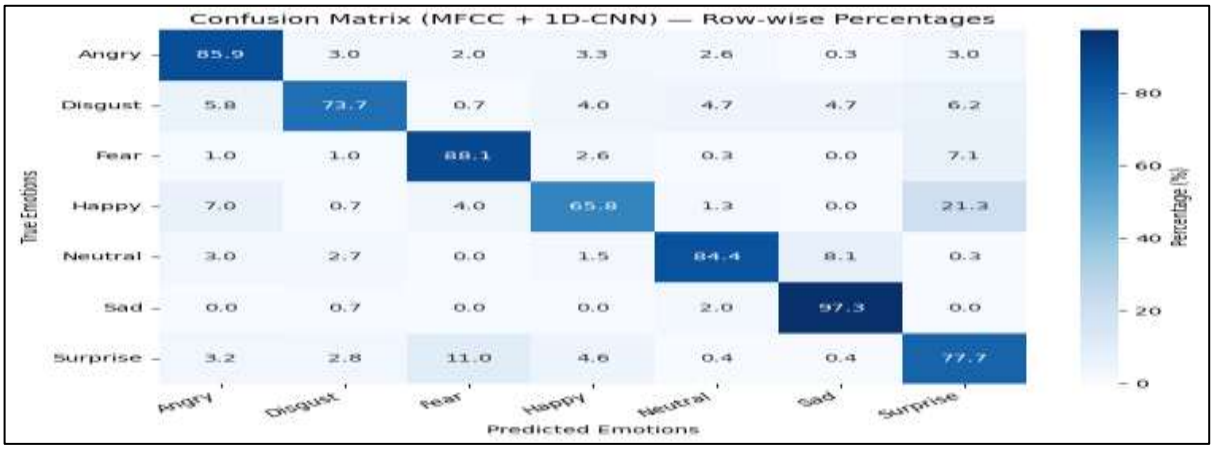
جدول 3: نتائج أداء نموذج(MFCC + 1D−CNN) بعد التدريب.

ACCURACY	VAL_ACCURACY	LOSS	VAL_LOSS
0.7929	0.7862	0.332	0.331

جدول 4: مقاييس أداء النموذج (MFCC + 1D−CNN) المدرب.

Error Rate	Precision	Recall	F1−Score
0.21	0.79	0.79	0.78

مصفوفة الالتباس(Confusion Matrix) :



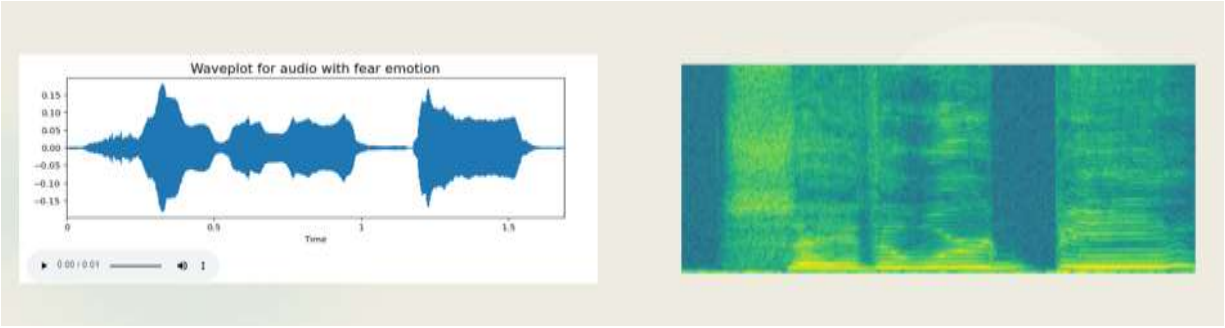
الشكل (3): مصفوفة الالتباس (Confusion Matrix) لنموذج (MFCC + 1D−CNN) باستخدام معطيات التحقق

تعكس مصفوفة الالتباس أداء النموذج في تصنيف المشاعر السبع المستهدفة وتُظهر عدد العينات التي تم تصنيفها بشكلٍ صحيحٍ مقابل العينات التي تم تصنيفها بشكلٍ خاطئ لكل فئة. تُبيّن القيم في القطر الرئيسي عدد التصنيفات الصحيحة، وكانت النسب عاليةً نسبياً عبر معظم الفئات، مما يشير إلى قدرة النموذج على تمييز السمات الصوتية الخاصة بكل شعورٍ وبدقةٍ معقولة.

ومع ذلك، أظهرت المصفوفة بعض التداخلات بين فئات محددة (مثلاً بين مشاعر مثل السعادة والمفاجأة، أو المفاجأة والخوف، او الحياد والحزن)، وهو أمرٌ متوقعٌ نظراً للتشابه في الخصائص الصوتية لبعض العواطف. يُعدّ هذا التداخل تحدياً معروفاً في معالجة اللغة المنطوقة وتصنيف المشاعر.

بالرغم من أن سمات MFCC تُعدّ من أكثر السمات شيوعاً واستخداماً في تحليل الإشارات الصوتية، إلا أن الاعتماد عليها فقط في تصنيف المشاعر الصوتية قد لا يكون كافياً، لأنها تفقد الكثير من المعلومات الزمنية، كما أن خصائصها الطيفية متقاربة بين بعض مثل (الحياد والحزن أو السعادة والمفاجأة). لذلك فهي تحتاج إلى دمج مع سمات أو تمثيلات أخرى لتحقيق دقة أعلى.

4-2-تصنيف المشاعر بالاعتماد على الصور الطيفية (Spectrogram):



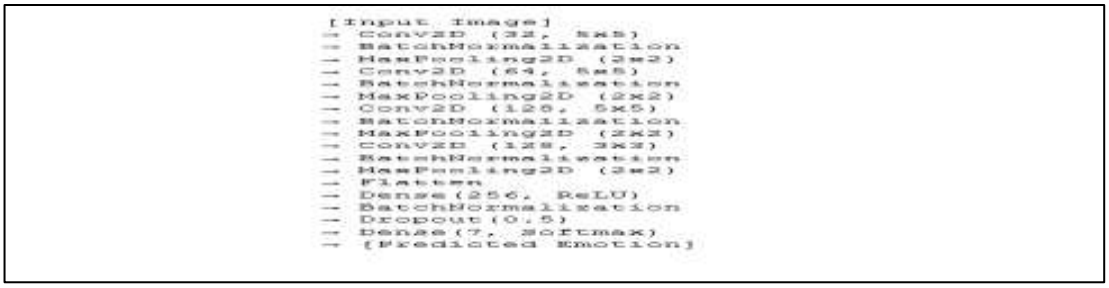
الشكل (4): تحويل ملف صوتي إلى صورة طيفية للعبارة الصوتية "Say the word bough" مأخوذة من مجموعة معطياتTESS

يُعدّ المخطط الطيفي (Spectrogram) تمثيلاً بصرياً يُظهر كيفية تغير ترددات الإشارة الصوتية عبر الزمن. أغلب مجموعات المعطيات الخاصة بالتعرّف على العواطف الصوتية توفّر ملفات صوتية فقط(مثل WAV أوMP3) ، ولا تتضمن صوراً طيفية جاهزة، لذلك كان من الضروري تحويل الإشارات الخام إلى تمثيل طيفي. تم استخدام 2800 عينة صوتية من مجموعة معطياتTESS ، حيث حُوِّل كل مقطع إلى صورة طيفية باستخدام تحويل فوريه قصير الأمد (STFT)مع نوافذ زمنية بطول 25ms. يمثّل المحور العمودي الترددات، والمحور الأفقي الزمن، بينما تُظهر شدة الألوان مقدار الطاقة عند كل تردد وزمن، كما هو موضح بالشكل(4).

تم بناء نموذج 2D-CNNمتعدد الطبقات للتعامل مع الصور الطيفية الملونة بحجم (256×256×3)، بالاعتماد على تسلسل من طبقات الالتفاف والتجميع.

بعد مرحلة التسطيح(Flatten) ، أُضيفت طبقة كثيفة (Dense) تتكون من 256 خلية عصبية تُنتج تمثيلات عالية المستوى (High-level Features) للصور المدخلة، وهي التي يعتمد عليها النموذج في اتخاذ قرار التصنيف النهائي. تم تدريب النموذج على معطيات مقسمة إلى تدريب (80%) وتحقق (20%) لمدة30 (EPOCH)، وذلك لأمرين رئيسيين:

أولاً، لضمان الاستقرار والوصول إلى تقارب تدريجي دون حدوث فرط تعلم(Overfitting) ، وثانياً لمواءمة عدد الدورات مع نموذج MFCC+1D-CNN بغرض تسهيل المقارنة العادلة بين النموذجين من حيث الأداء.



الشكل (5): نموذج 2D-CNN لاستخلاص السمات من الصور الطيفية

تم تقسيم جزء من المعطيات للتدريب 80% وجزء آخر للتحقق 20% من أداء النموذج.

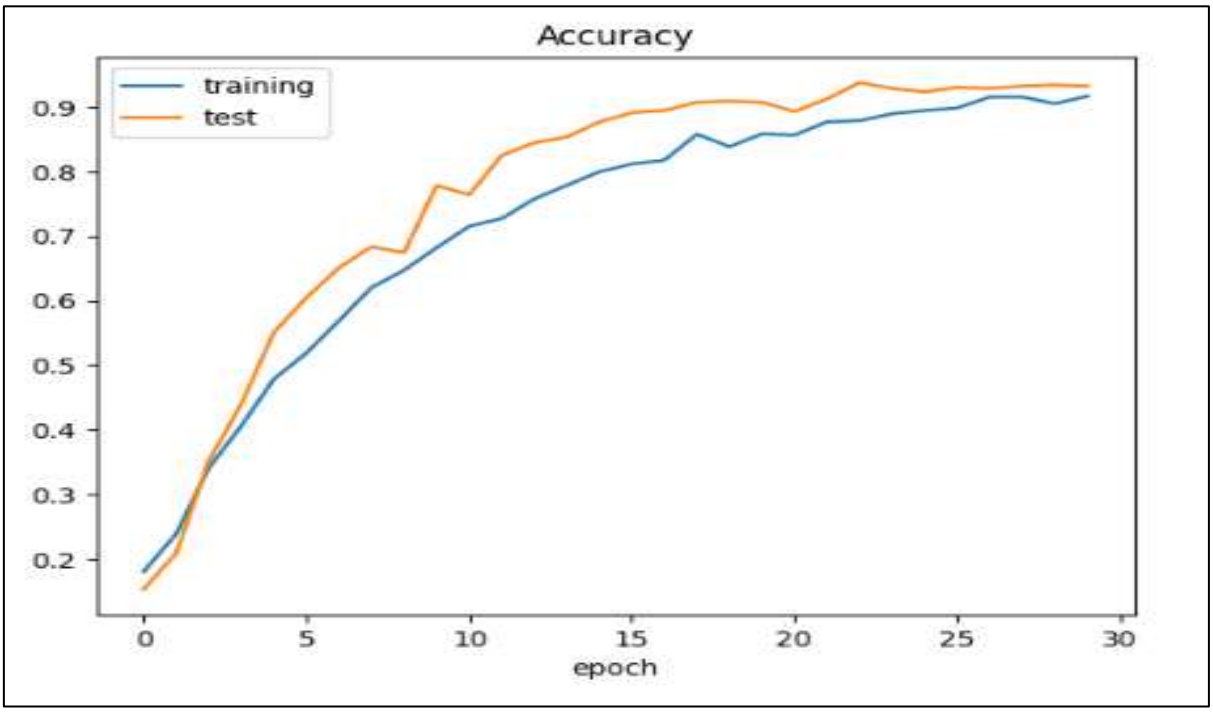
جدول 5: نتائج أداء نموذج(Spectrogram + 2D−CNN) بعد التدريب.

ACCURACY	VAL_ACCURACY	LOSS	VAL_LOSS
0.9165	0.9320	0.2489	0.2117

جدول6: مقاييس أداء النموذج (Spectrogram + 2D−CNN) المدرب.

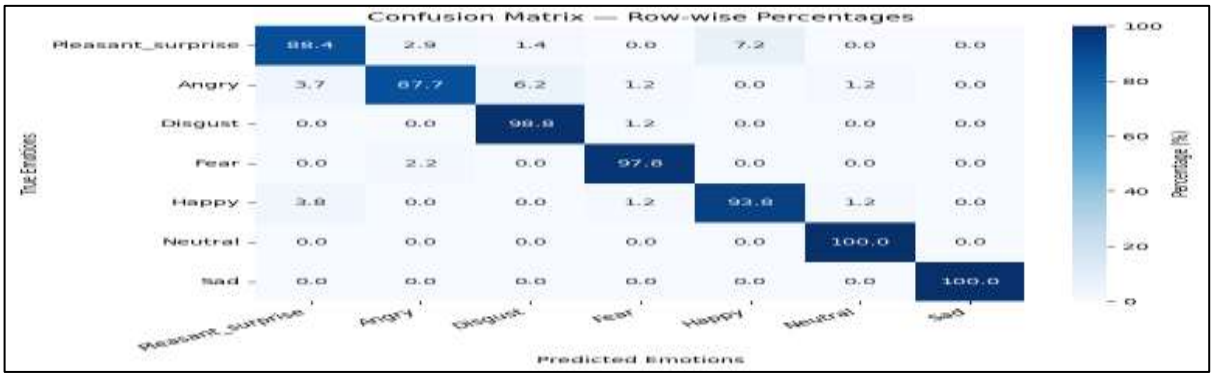
Error Rate	Precision	Recall	F1−Score
0.11	0.78	0.77	0.75

يمثل هذا الشكل (6) منحنيات الدقة (Accuracy) لكل من مجموعة التدريب ومجموعة الاختبار خلال 30 دورة تدريبية(epochs) ، يُلاحظ أن دقة النموذج على كل من التدريب والاختبار ترتفع بشكلٍ متسقٍ ومستقرٍ مع تقدم عدد الدورات.



الشكل (6): منحنيات الدقة (Accuracy) لنموذج (Spectrogram + 2D−CNN) خلال التدريب والتحقق

مصفوفة الالتباس(Confusion Matrix) :



الشكل (7): مصفوفة الالتباس (Confusion Matrix) لنموذج (Spectrogram + 2D−CNN) باستخدام معطيات التحقق

يُلاحظ أن النموذج تمكن من تصنيف غالبية العينات، حيث حققت بعض الفئات مثل Neutralو Sad نسبة تصنيف صحيحة بلغت 100%، بينما سجلت فئات أخرى مثل Disgustو Fearدقة 97%.

في المقابل، وُجد بعض الالتباس بين المشاعر المتقاربة، مثل الخلط بين Pleasant_surpriseو Happy، وكذلك بين Angryو Disgust، وهو أمر متوقع نظراً للتشابه في الخصائص الصوتية.

4-3-الدمج على مستوى السمات(Feature-Level Fusion) :

في هذه التجربة تم بناء نموذج مهجن يعتمد على دمج نوعين مختلفين من السمات الصوتية، بحيث يعمل كل نموذج عصبي كأداة استخلاص سمات (Feature Extractor) فقط، بينما مهمة التصنيف النهائية أوكلت إلى خوارزمية (Random Forest).

(أ) النماذج الأساسية(Pre-trained Models) :

سيتم استخدام كل من النموذجين السابقين الذي تم تدريبهم.

1. النموذج الأول(Spectrogram + 2D-CNN) :

- يدخل ملف الصوت كصورة طيفية(Spectrogram) .
- استخدام 2D-CNN مدرب مسبقاً لاستخراج 8192 سمة من طبقةFlatten .

2. النموذج الثاني(MFCC + 1D-CNN) :

- يدخل ملف الصوت.
- استخدام 1D-CNNمدرب مسبقاً لاستخراج 288 سمة من طبقةFlatten .

منهجية الدمج(Fusion Mechanism) :

تم اعتماد دمج السمات على مستوى Feature-Level Fusionلتجاوز القيود التي ظهرت عند استخدام MFCC أو Spectrogramبشكل منفرد، حيث تم دمج المتجهات التمثيلية الناتجة عن كل نموذج لتكوين متجه موحد يمثل كل عينة صوتية.

بدلاً من الاعتماد على طبقة التصنيف (Softmax) في كل نموذج، جرى استخراج المتجهات التمثيلية من طبقة Flatten (8192سمة منSpectrogram ، و288 سمة منMFCC) لأن طبقة Flatten تحتفظ بتمثيلات غنية (High-level Representations) تعلمتها الشبكة أثناء التدريب، بينما طبقة Softmax تؤدي دوراً نهائياً في اتخاذ القرار فقط، وبالتالي فهي غير مناسبة للدمج.

بعد ذلك، تم دمج المتجهين أفقياً (Concatenation) لتكوين متجه موحّد طوله:

$$F=Fspectrogram \oplus Fmfcc \Rightarrow 8192+288= 8480$$

هذا الدمج المبكر (Feature-Level) يتيح للنموذج الاستفادة من الارتباطات التبادلية (Correlations) بين الخصائص الطيفية والزمنية معاً، وهو ما لا يمكن تحقيقه في حال الاعتماد على Score-Level أو-Decision Level Fusionحيث يتم الدمج بعد اتخاذ القرار، مما يؤدي إلى فقدان جزء من العلاقات المفيدة بين السمات.

بدلاً من الاعتماد على Softmax داخل الشبكة، استخدم مصنف Random Forestلسببين:

1. كفاءته في التعامل مع فضاءات سمات عالية الأبعاد.
2. قدرته على تقليل خطر Overfitting مع إمكانية تفسير السمات

النتائج والمناقشة:

Accuracy: 0.9678571428571429

	precision	recall	f1-score	support
OAF_Fear	0.98	1.00	0.99	43
OAF_Pleasant_surprise	0.87	0.85	0.86	39
OAF_Sad	0.98	0.98	0.98	52
OAF_angry	1.00	1.00	1.00	42
OAF_disgust	0.98	0.98	0.98	41
OAF_happy	0.87	0.89	0.88	38
OAF_neutral	0.97	0.97	0.97	38
YAF_angry	1.00	0.94	0.97	36
YAF_disgust	0.95	1.00	0.97	36
YAF_fear	1.00	1.00	1.00	46
YAF_happy	0.97	1.00	0.99	37
YAF_neutral	1.00	0.97	0.98	31
YAF_pleasant_surprised	1.00	0.96	0.98	45
YAF_sad	0.97	1.00	0.99	36
accuracy			0.97	560
macro avg	0.97	0.97	0.97	560
weighted avg	0.97	0.97	0.97	560

الشكل (8): مصفوفة الالتباس للنموذج المهجن عند اختباره على معطياتTess

جدول 7: اهم مقاييس الأداة الناتجة عن اختبار النموذج المهجن

Accuracy	Precision	Recall	F1–Score
0.97	0.97	0.97	0.97

النتائج:

- على مجموعة معطيات TESS (7 فئات)، حقق النموذج دقة 97%.
- مصفوفة الالتباس أظهرت تقليل الالتباس بين الفئات المتقاربة مثل(Sad–Neutral) .
- الفئات الواضحة مثل Angryو Fearوصلت دقتها إلى 100%
- عند تطبيق النموذج على مجموعة معطيات IEMOCAP، التي تتسم بدرجة أكبر من التعقيد وتنوع المتحدثين مقارنةً بـ TESS، أظهر النموذج قدرة جيدة على التعميم، حيث وصلت الدقة 75%، مع تحسن في التمييز بين الفئات المتقاربة مثل Sad–Neutralو Happy–Neutralوتوضح مصفوفة الالتباس أن الأداء ظل مرتفعاً حتى مع المعطيات الأكثر تنوعاً، مما يؤكد كفاءة آلية الدمج المقترحة في التعامل مع بيئات أكثر تعقيداً.

الشكل (9): مصفوفة الالتباس للنموذج المهجن عند اختباره على معطياتIEMOCAP

	precision	recall	f1-score	support
angry	0.69	0.79	0.74	342
happy	0.80	0.83	0.82	341
neutral	0.56	0.44	0.50	342
sad	0.72	0.75	0.73	342
accuracy			0.75	1367
macro avg	0.69	0.70	0.69	1367
weighted avg	0.69	0.70	0.69	1367

جدول 8: مقارنة بين الدراسات المرجعية والنماذج المقترحة:

المرجع	مجموعة المعطيات	عدد الفئات	المنهجية	الدقة	أبرز نقاط الضعف
Wang et al. (2020) [14]	IEMOCAP	4	Dual-Stream LSTM (MFCC + 2 نوع Mel-Spectrogram)	72.7% (WA), 73.3% (UA)	الالتباس مستمر بين Happy–Neutral وSad–Neutral
Yurun He et al. (2021) [15]	IEMOCAP	4	Cross-Attention Transformer (Raw + MFCC + Spectrogram)	73.8% (WA), 74.25% (UA)	تحسن محدود الالتباس بين Neutral–Sad وNeutral–Happy مستمر
Basha et al. (2025) [16]	RAVDESS, TESS, SAVEE	8	Ensemble Attribute Selection + CNN-LSTM + Attention	87.08%	صعوبة التمييز بين Happy–Neutral
النموذج المقترح	TESS	7	Feature-Level Fusion (MFCC + Spectrogram) + Random Forest	97%	قلّل الالتباس بين الفئات المتقاربة مثل Sad–Neutral وHappy–Surprise وحقق توازناً ممتازاً بين جميع الفئات
النموذج المقترح	IEMOCAP	4	Feature-Level Fusion (MFCC + Spectrogram) + Random Forest	75%	تحسن في التمييز بين الفئات المتقاربة مثل Happy–Neutral وSad–Neutral مع أداء جيد عبر متحدثين متنوعين

يوضح الجدول (8) مقارنة شاملة بين نتائج الدراسات السابقة والنماذج المقترحة في هذا البحث، مع التركيز على نقاط القوة والقصور.

- النموذج المقترح Feature-Level Fusion على TESS حقق أعلى دقة (97%) مع المعطيات الصغيرة وقلّل الالتباس بين الفئات المتقاربة.
- عند استخدام معطيات أكبر وأكثر تنوعاً مثلIEMOCAP ، أظهر النموذج قدرة جيدة على التعميم، حيث حقق دقة 75%، مع تحسن ملحوظ في التمييز بين الفئات المتقاربة مثل Sad–Neutralو Happy–Neutral.

الاستنتاجات والتوصيات:

أظهرت نتائج البحث أن استخدام خوارزميات MFCCفقط يحقق دقة مقبولة في تصنيف المشاعر الصوتية، إلا أن محدوديتها في تمثيل الخصائص الزمنية والبصرية يقلل من قدرتها على التمييز بين المشاعر المتقاربة.

في المقابل، أثبتت الصور الطيفية (Spectrograms) فعاليتها كتمثيل بصري غني، إذ حققت دقة تصنيف بلغت 93.20%، مما يؤكد أهميتها في التعرف على الأنماط العاطفية.

عند تطبيق الدمج على مستوى السمات (Feature-Level Fusion) باستخدام نماذج التعلم العميق ومصنفات Random Forest، ارتفعت الدقة إلى 97%، وهو ما يبرهن على أن الدمج بين التمثيلات الصوتية والبصرية يعزز بشكل ملموس من أداء النماذج.

كما أثبتت نتائج الاختبار على مجموعة معطيات أكبر وأكثر تعقيداً مثل IEMOCAPأن النموذج يتمتع بقدرة جيدة على التعميم مع دقة 75%، ما يعكس كفاءته في التعامل مع بيئات أكثر تنوعاً وتعقيداً.

التوصيات المستقبلية:

- تطوير النموذج ليعمل في الزمن الحقيقي (Real-Time) لتوسيع فرص تطبيقه في الرعاية النفسية، التعليم، خدمة العملاء، والمساعدات الافتراضية.

- التوسّع في استخدام نماذج حديثة مثل Transformersوتقنيات Attentionو Autoencoderلتحسين التمثيلات وفهم السياق الزمني بشكل أعمق.
- الاعتماد على مجموعات معطيات متعددة اللغات والثقافات لتعزيز قدرة التعميم وتقليل الانحيازات.
- معالجة مشكلة عدم توازن المعطيات (Class Imbalance)عبر تقنيات متقدمة مثل SMOTEأو Data Augmentation.
- دمج سمات إضافية مثل تعابير الوجه والإيماءات الجسدية ضمن أنظمة Multimodal Emotion Recognition لتحقيق أداء أفضل في البيئات التفاعلية.

References:

[1] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," arXiv:1912.10458v1 [cs.SD], Dec. 22, 2019. [Online]. Available: https://arxiv.org/abs/1912.10458

[2] K. Dupuis and M. K. Pichora-Fuller, "Toronto Emotional Speech Set (TESS)," University of Toronto, 2010. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[4] M. Z. Iqbal and G. F. Siddiqui, "MFCC and Machine Learning Based Speech Emotion Recognition on TESS and IEMOCAP Datasets," *Foundation University Journal of Engineering and Applied Sciences (in Arabic)*, Pakistan, 2020.

[5] R. R. Choudhary, G. Meena, and K. K. Mohbey, "Speech Emotion Based Sentiment Recognition using Deep Neural Networks," *Journal of Physics: Conference Series*, vol. 2236, Article No. 012003, 2022. [Online]. Available: https://doi.org/10.1088/1742-6596/2236/1/012003

[6] S. Padi, D. Manocha, and R. D. Sriram, "Multi-window Data Augmentation Approach for Speech Emotion Recognition," arXiv:2010.09895v4 [cs.SD], Feb. 16, 2022. [Online]. Available:https://arxiv.org/abs/2010.09895

[7] M. Ishaq, M. Khan, and S. Kwon, "TC-Net: A Modest & Lightweight Emotion Recognition System Using Temporal Convolution Network," *Computers, Materials & Continua (CSSE)*, vol. 46, no. 3, pp. 3356–3357, 2023. [Online]. Available: https://doi.org/10.32604/csse.2023.037373

[8] Q. Ouyang, "Speech emotion detection based on MFCC and CNN-LSTM architecture," *Journal of Physics: Conference Series*, vol. 2137, Article No. 012012, 2022. [Online]. Available:https://doi.org/10.1088/1742-6596/2137/1/012012

[9] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," arXiv:1806.01506v2 [cs.SD], May 2, 2019. [Online].Available:https://arxiv.org/abs/1806.01506

[10] M. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2990405

[11] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proc. 2021 Int. Conf. on Multimodal Interaction (ICMI '21)*, Montréal, QC, Canada, 2021, pp. 8. [Online]. Available:https://doi.org/10.1145/3462244.3481003

[12] S. Akinpelu and S. Viriri, "Robust feature selection-based speech emotion classification using deep transfer learning," *Applied Sciences*, vol. 12, Art. no. 8265, 2022. [Online].Available:https://doi.org/10.3390/app12168265
[13] N. Penumajji, "Deep learning for speech emotion recognition: A CNN approach utilizing mel spectrograms," arXiv:2503.19677v1 [cs.SD], Mar. 25, 2025. [Online]. Available:https://arxiv.org/abs/2503.19677
[14] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," arXiv:1910.08874v4 [eess.AS], Feb. 13, 2020. [Online]. Available: https://arxiv.org/abs/1910.08874
[15] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *Proc. ICASSP 2023 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Rhodes, Greece, 2023. [Online]. Available: https://doi.org/10.1109/ICASSP49357.2023.10095777
[16] S. A. K. Basha, P. M. D. R. Vincent, S. I. Mohammad, A. Vasudevan, E. E. H. Soon, Q. Shambour, and M. T. Alshurideh, "Exploring deep learning methods for audio speech emotion detection: An ensemble MFCCs, CNNs and LSTM," *Applied Mathematics & Information Sciences*, vol. 19, no. 1, pp. 75–85, 2025. [Online]. Available: https://doi.org/10.18576/amis/190107
[17] M. Mustaqeem and S. Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021. [Online]. Available: https://doi.org/10.32604/cmc.2021.015070