مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٨) العدد(٣) ٢٠٢٤
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

تصميم نظام صوتي لتمييز أصوات المدخنين من غير المدخنين بالاعتماد على تقنيات التعلم العميق

د. م. رنيم كناج•

(تاريخ الإيداع ٢٠٢٣/١١/٦ . قُبِل للنشر في ٢٠٢٤/٣/١١)

□ ملخّص □

تعد تقنيات التعرف بكافة أشكالها من أهم التقنيات الحديثة التي دخلت بقوة في مجالات الحياة المختلفة. في هذا البحث تم تطوير نموذج للتعرف على حالة تدخين الشخص من خلال صوته، باستخدام تقنيات وخوارزميات التعلم العميق.

مبدأ العمل هو تحويل الإشارات الصوتية (Audio speech) إلى مخططات طيفية (spectrograms) ثم استخلاص السمات من صور المخططات باستخدام الشبكات العصبونية الالتفافية (CNN) Convolutional Neural Network فهي أكثر دقة في استخلاص سمات الصور[2] [1].، ثم الاعتماد على نموذج للذاكرة طويلة المدى ثنائية الاتجاه Bi-directional Long Short-Term Memory (Bi-LSTM)، وهي تقنية لتحسين تحليل الإشارة الصوتية عن طريق تحديد الأجزاء الأكثر أهمية في المخطط الطيفي والتركيز عليها. وتم أداء هذا البحث بالاعتماد على النموذجين VGG-16 & ResNet-50. يمكن تنزيل هذه النماذج مع ضبط معلماتها على القيم التي كانت لديها في نهاية التدريب. بهذه الطريقة، يمكننا بسهولة الحصول على أداء ممتاز دون تدريب نموذج جديد بأنفسنا، ثم تمت عملية مقارنة النتائج بينهما، حيث تم استخدام عدة مقاييس لتقييم أداء النموذج وهي: معدل الدقة (Accuracy)، الدقة (Precision)، معدل الاستحواذ (Recall)، قيمة(F1 Score)، معدل الخطأ المتساوي Equal Error Rate (ERR). أظهرت نتائج المحاكاة والتنفيذ أن النموذج المقترح الذي يعتمد على ResNet-50 حقق دقة أعلى تصل إلى ٩٤,٦ % بينما النموذج الذي يعتمد على VGG-16 حقق دقة تصل إلى٩٣,٦%. حيث تم استخدام مجموعة بيانات تحتوي على ١٥٠٠ صوتاً مقسمة إلى ذكر وأنثى، وتم تقسيمها إلى ٨٠% بيانات تدريب و ٢٠% بيانات اختبار.

كلمات مفتاحية: التعلم العميق، نماذج شبكات التعلم العميق، الشبكات العصبونية الالتفافية، ذاكرة طويلة المدى ثنائية الاتجاه، التعلم بالنقل، آلية الانتباه.

• مدرس، قسم هندسة النظم الحاسوبية والالكترونية، كلية هندسة تكنولوجيا المعلومات والاتصالات، جامعة طرطوس، طرطوس سوريا.raneem.knaj@gmail.com

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٨) العدد(٣) ٢٠٢٤
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (3) 2024

Designing an audio system to distinguish the voices of smokers from non-smokers based on deep learning techniques

•Dr. Eng. Raneem Knaj

(Received //2023 . Accepted //2024)

□ ABSTRACT □

Recognition technologies in all their forms are one of the most important modern technologies that have entered strongly in various areas of life. In this research, a model was developed to recognize a person's smoking status through his voice, using deep learning techniques and algorithms. The principle of action is to convert audio speech into spectrograms and then extract features from the diagram images using convolutional neural networks (CNN), which are more accurate in extracting image features [1][2]. This research was performed based on the VGG-16 and ResNet-50 models. Such models can be downloaded with their parameters set to the values they had at the end of training. In this way, we can easily get excellent performance without training a new model ourselves, and then the results were compared between them, where several measures were used to evaluate the model's performance: accuracy, precision, recall, F1 score, and equal error rate (ERR). The results of the simulation and implementation showed that the proposed model based on ResNet-50 achieved a higher accuracy of 94.6% while the model based on VGG-16 achieved an accuracy of 93.6%. A dataset of 1,500 votes divided into male and female was used, and it was divided into 80% training data and 20% test data.

Key words deep learning, deep learning network models, convolutional neural networks, bidirectional long-term memory, transfer learning, attention m

•lecture, Department of CESE, Faculty of Information and Communication Technology, Tartous University, Tartous, Syria, e-mail: raneem.knaj@gmail.com

١- المقدمة:

يعتبر التعلم العميق Deep Learning(DL) واحد من أحدث وأهم مجالات تعلم الآلة وأكثرها تطوراً واستخداماً من قبل الشركات العالمية ضمن مجالات متعددة من التطبيقات. تستخدم تطبيقات التعلم العميق في العديد من المجالات مثل التعرف على الصوت والصور والتحدث إلى الكمبيوتر وحل مشكلات معقدة بشكل أفضل، لهذا السبب أن التعلم العميق يلعب دوراً مهماً في تحويل العالم ويمكننا القول أن تقنيات التعلم العميق هي مفتاح ازدهار الذكاء الاصطناعي والتكنولوجيا الذكية في المستقبل.

هناك الكثير من التطبيقات التي كانت صعبة الإنجاز أو ربما مستحيلة بدون شبكات التعلم العميق. تعد تقنية التعرف على الكلام من أهم التقنيات الحديثة التي دخلت بقوة في مجالات الحياة المختلفة سواء الطبية أو الأمنية أو الصناعية، بناء عليه تم تطوير العديد من الأنظمة المعتمدة على طرق مختلفة في استخلاص السمات والتصنيف.

يعتمد البحث الحالي على تطوير خوارزمية هجينة تعتمد على تقنيات التعلم العميق الذي يعتمد بدوره على شبكات عصبونية اصطناعية لتحليل الأصوات وتمييزها من خلال تدريبها على مجموعة كبيرة من الأصوات المختلفة حيث يتم تحليل كل صوت وفصله إلى مكوناته الصوتية المختلفة ثم تدريب الشبكات العصبونية على تمييز هذه المكونات وتحديد إذا كانت تنتمي إلى صوت مدخن أو صوت غير مدخن. فقد تستخدم هذه الخوارزمية في العديد من التطبيقات المختلفة مثل التحليل الصوتي للأغاني والموسيقى وفي التحكم بالأجهزة الالكترونية التي تحتاج الى التعرف على صوت المستخدم وفي تطوير الروبوتات والاجهزة الذكية التي تحتاج الى التفاعل مع المستخدم عن طريق الصوت.

٢- الدراسات المرجعية ذات الصلة:

في السنوات الأخيرة شهدت تقنيات وخوارزميات التعلم العميق تطوراً كبيراً واستخدمت في مجالات عديدة أهمها مجالات التعرف (Recognition) والرؤية الحاسوبية (computer vision). تستخدم تقنيات التعلم العميق في مجال التعرف لتحسين أداء البرامج والأنظمة المستخدمة في هذا المجال، حيث تستخدم شبكات عصبونية عميقة لتحليل البيانات الصوتية والبصرية واستخلاص المعلومات المهمة المتعلقة بالمشاعر، وتعتبر تقنيات التعلم العميق في مجال الرؤية الحاسوبية مفيدة جداً في العديد من المجالات، مثل التعرف على الأشخاص والأشياء والأماكن، وكذلك في التحكم في المركبات الذاتية والروبوتات، وفي تطوير الألعاب الإلكترونية والواقع الافتراضي، وغيرها من المجالات التي تعتمد على تحليل وفهم الصور والفيديوهات. وقد أجريت العديد من الدراسات البحثية في هذا المجال، نذكر أهمها:

- عبر العقد الأخير من الزمن، تطور استخدام الشبكات العصبونية والتعلم العميق بشكل كبير في مجال معالجة اللغات الطبيعية Natural Language Processing (NLP) حيث بيّن Collobert وآخرون أن إطار التعلم العميق البسيط قد تفوق على كل الطرق التقليدية المستخدمة لمعالجة اللغات الطبيعية. كما أوضح أن العديد من التطبيقات مثل التعرف على المكونات وعنونة الصور المعتمدة على التسميات المعنونة وغيرها

تستخدم التعلم العميق كحل فعال لها وخصوصاً الشبكات العصبونية الالتفافية Convolutional Neural Network (CNN)والشبكات العصبونية التكرارية(RNN). [3]

- بينما الباحث Deshmukh وآخرون قدموا أبحاثاً حول الدراسة التفصيلية لتطبيقات التعلم العميق (DL) للتعرف على الكلام العاطفي(SER)، حيث استخدموا تقنيات التعلم العميق (DL) مثل الشبكة العصبونية التكرارية (RNN) والشبكة العصبونية الالتفافية (CNN)، علاوة على ذلك: تم وصف تقنيات (DL) والهيكل الطبقي بشكل شامل يعتمد على تصنيف الآراء الطبيعية المختلفة مثل الخوف السعيد والحزين والحيادي [4].
- بينما اعتمد الباحث Khalil في نتائجه على استخراج ثلاث ميزات من الإشارة الصوتية وهي: Voice Pitch و (MFCC) Mell Frequency cepstral coeffective و(STE) Short Term Energy . كما استخدم عينات البيانات الصوتية في كلام الفتيات والفتيان وحقق أقصى دقة للتصنيف.[5]
- أما بالنسبة للباحث RICHARD و زملائه حاولوا جعل النظام يتعرف على المتحدث من خلال صوته، فالفكرة هي تحديد الاختلافات الكامنة في الأعضاء المفصلية (بنية السبيل الصوتي، وحجم تجويف الأنف، وخصائص الحبل الصوتي) وطريقة التحدث. حيث تمكنوا من تصنيف أصوات الكلام إلى ثلاث فئات متميزة: الأصوات الصوتية، والأصوات الاحتكاكية أو غير الصوتية، والأصوات الانفجارية. فقد اعتمدوا على استراتيجيتين لتقنيات الكلام القوية وهي:
  - التخفيف من حدة المشاكل التي تنشأ بسبب تأثيرات القناة والضوضاء.
  - الاعتماد على مصنفات قوية للحصول على نتائج أكثر دقة. [6]
- [7] اقترح الباحثون نهجاً جديداً للتعرف التلقائي على اللغة language identification (LID) استناداً إلى الشبكات العصبونية التكرارية (RNNs) للذاكرة قصيرة المدى (LSTM). بدافع من النجاح الأخير للشبكات العصبية العميقة (DNNs) لـ LID، اكتشفنا LSTM RNNs باعتبارها بنية طبيعية لتضمين المعلومات السياقية الزمنية داخل نظام الشبكة العصبونية. تم مقارنة النظام المقترح بنظام قائم على i-vector والمكونات المختلفة لتغذية DNNs إلى الأمام. تظهر النتائج على NIST LRE 2009 (تم تحديد ٨ لغات وحالة ٣s) أن بنية LSTM RNN تحقق أداءً أفضل من أفضل نظام DNN ذي ٤ طبقات باستخدام معلمات أقل ٢٠ مرة (M١٠ مقابل M٢١٠). بالإضافة إلى ذلك، وجدنا أن درجات LSTM RNN تمت معايرتها بشكل أفضل من تلك الناتجة عن المتجه i أو أنظمة DNN. يوضح هذا العمل أيضاً أن كلاً من أنظمة LSTM RNN و DNN تفوق بشكل ملحوظ أداء نظام i-vector الفردي. علاوة على ذلك، يمكن الجمع بين كل من نهج الشبكة العصبية مما يؤدي إلى تحسين ٢٥٪.

تدرس الأعمال السابقة [8] [9] عملية استخراج الميزات من مجموعة البيانات باستخدام ميزات المستوى الصوتي والسمعي.

وفي الأعمال الحديثة أيضاً [10] [11]، تم اقتراح آليات تعلم عميق مختلفة بميزات صوتية مميزة بالإضافة إلى الاختناق وحققت أعلى دقة بنسبة ٩٧,١٠٪ و ٩٧,٣٠٪ على التوالي.

ولعل الدراسة [12] في عام 2021 بعنوان:

" Hybrid Deep-Learning and Machine-Learning Models for Predicting COVID-19"

للباحثين:

٣- أهمية البحث وأهدافه:

تأتي أهمية البحث من تسهيل التفاعل الطبيعي مع الآلات عن طريق التفاعل الصوتي المباشر بدلاً من استخدام الأجهزة التقليدية كمدخلات لفهم المحتوى اللفظي وتسهيل تفاعل المستمعين من البشر بالإضافة إلى تشكيل قاعدة بيانات صوتية باللغة العربية هذا الأمر سيغني برامج مخاطبة الآلة الناطقة باللغة العربية. كما أنه وسيلة لفرز حالة الأصوات البشرية بشكل آلي وأتوماتيكي دون التدخل البشري وهذا يساعد على بناء نظام أكثر موثوقية وفاعلية.....الخ.

يهدف هذا البحث إلى تطوير نموذج هجين يستخدم النموذجين المدربين مسبقاً (pre-trained model) وهما (VGG-16 و ResNet-50) مع شبكة (BI-LSTM+Attention) للتعرف على حالة الشخص من خلال صوته وذلك باستخدام تقنيات وخوارزميات التعلم العميق.

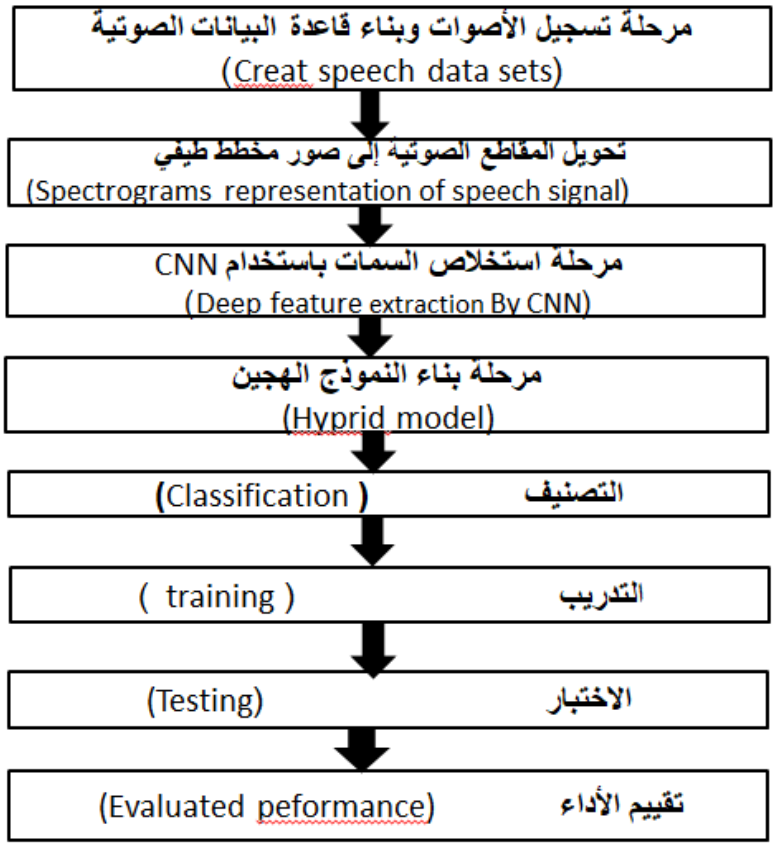
٤- طرائق البحث وموارده:

يعتمد هذا البحث على المنهج التحليلي والتجريبي في عملية المقارنة بين النموذجين المقترحين ومناقشة النتائج. تم الاعتماد على مجموعة من الأدوات البرمجية والمادية. حيث تم الاعتماد على نظام التشغيل windows 10 ولغة البرمجة Python ٣,١١ ومنصة الـ spyder كذلك تم استخدام مكتبتي (Keras و TensorFlow) من أجل بناء نماذج شبكات ال .DL . تم استخدام خوارزميات التعلم العميق مثل: RNN وَ CNN ونماذج شبكات التعلم العميق المدربة مسبقاً. (VGG-16 , ResNet-50) بالنسبة لمجموعة البيانات، تم استخدام 1500 عينة مقسمة ٨٠% بيانات تدريب و ٢٠% بيانات اختبار.

٥- النظام المقترح:

يتألف النظام المقترح من خطوات أساسية كما هو واضح في الشكل (١):

النظام المقترح



الشكل (١): وصف النظام المقترح

٥-١: مرحلة تسجيل الأصوات وبناء قاعدة البيانات الصوتية:

تشكل قواعد بيانات الكلام (Speech Database) الركن الأساسي لبناء النظم الحاسوبية المختلفة كنظم التعرف الآلي على الكلام والنطق الآلي والتعرف على المتحدث والتعرف على اللغات واللهجات. وتتكون قواعد بيانات الكلام عادة من ملفات صوتية (Wave Files) سبق أن سجلت لأشخاص باللغة المطلوبة، وكلما كانت قاعدة بيانات الكلام غنية وثرية في محتواها كلما ساعد ذلك على إخراج نظم حاسوبية ذات أداء متميز.

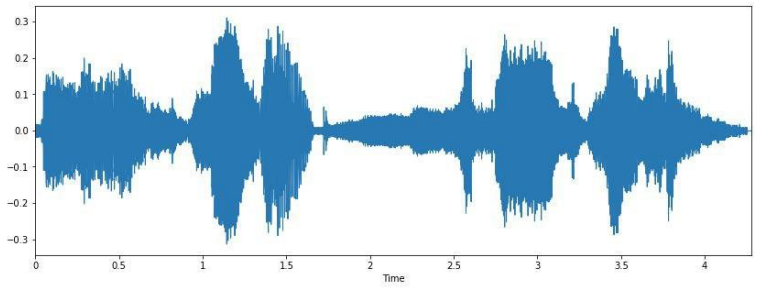
بالنسبة لمجموعة البيانات المستخدمة:

في هذا البحث تم تسجيل أصوات /٥٠/ شخص نصفهم مدخن والنصف الأخر غير مدخن باستخدام الميكرفون بواسطة الحاسب المحمول عبر برمجية Audacity. ثم تم تكرار التسجيل ثلاث مرات. كل شخص يلفظ الجملة ثلاث مرات ثم يتم تقطيع الجملة عشر مقاطع عبر برنامج Audacity (٥٠*١٠*٣=١٥٠٠مقطع تسجيل ناتج)، يتم اللفظ بشروط عادية من السرعة ومستوى رفع الصوت، وبعد إنهاء التسجيلات كانت مرحلة الترميز الصوتي للملفات الصوتية، ويتطلب الترميز الصوتي توفر القدرة على التركيز والمتابعة الدقيقة للأصوات اللغوية.

٥-٢: تحويل المقاطع الصوتية إلى صور مخطط طيفي:

يتم تمثيل الصوت في شكل إشارة صوتية تحتوي على بارامترات هامة متل التردد وعرض النطاق الترددي والديسبل وما إلى ذلك، يمكن التعبير عن الإشارة الصوتية النموذجية كدالة للسعة والوقت. تمتلك لغة بايثون بعض المكتبات لمعالجة الصوت مثل: (librosa و pyaudio).

يبين الشكل (٢) رسم الإشارة الصوتية قبل تحويلها إلى مخطط طيفي:



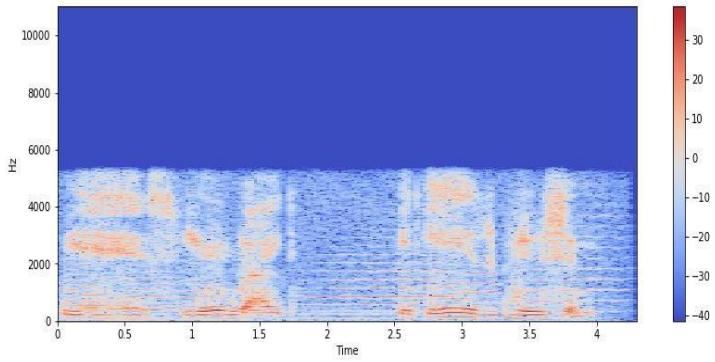
الشكل(٢): الإشارة الصوتية

ثم نقوم بتحويل الإشارة الصوتية إلى صورة مخطط طيفي، المخطط الطيفي هو وسيلة مرئية لتمثيل الإشارة بمرور الوقت عند ترددات مختلفة موجودة في شكل موجة معين. حيث توفر هذه الصور معلومات محددة ودقيقة لتصنيف الصوت [13] [14] [15].

تتمثل خطوات تحويل الإشارة الصوتية إلى صور مخطط طيفي كمايلي:

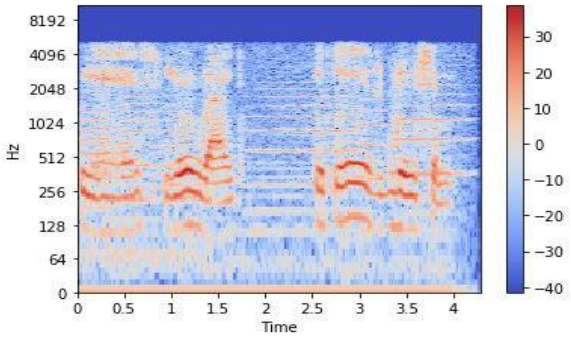
- تحويل فوريه السريع للحصول على مقدار طيف التردد لأي نقطة زمنية معينة.
- تحويل السعة الطيفية بواسطة بنك مرشح ميل إلى طيف ميل.
- حساب قيم الطيف باستخدام دالة اللوغاريتم لتحسين التمثيل الصوتي
- إظهار صورة المخطط الطيفي
- تطبيق تقنيات معالج الصور لتحسين جودة المخطط الطيفي.

عادةً ما يتم تصوير المخطط الطيفي كخريطة حرارية، أي كصورة ذات كثافة تظهر من خلال تغيير اللون أو السطوع. يمكننا عرض مخطط طيفي باستخدام الأمر البرمجي التالي: librosa.display.specshow.



الشكل(٣):مخطط طيفي لأحد عينات الصوتية التي تم تسجيلها

يُظهر المحور الرأسي لهذا المخطط الترددات (من ٠ إلى ١٠ كيلو هرتز)، ويُظهر المحور الأفقي وقت المقطع. وبما أننا نرى أن كل الأحداث تحدث في الجزء السفلي من الطيف، فيمكننا تحويل محور التردد إلى محور لوغاريتمي وتصبح الصورة كمايلي:



الشكل(٤):المخطط الطيفي باستخدام اللوغاريتم

٣-٥ : مرحلة استخلاص السمات باستخدام CNN:

السمات (features) هي معلومات مفيدة ومميزة تستخرج من البيانات، كالصور أو الصوت أو النص، تساعد في فهم وتصنيف وتحليل البيانات بشكل أفضل. الشبكات العصبونية الالتفافية (CNN) هي تقنية قوية لاستخلاص السمات من الأصوات، حيث تستخدم طبقات متعددة من المرشحات التي تكشف عن أنماط وتفاصيل مختلفة في الأصوات. ولكن ما هي بالضبط هذه "الميزات features"؟ بشكل عام في التعلم العميق، ندع الشبكة تكتشف الميزات من تلقاء نفسها أثناء التدريب. لا حاجة لتصميم وترميز خوارزمية استخراج ميزة معقدة، فنحن فقط نعطي الشبكة بُنية ذات مرونة كافية ونقوم بتدريبها بالأمثلة. وفي هذا البحث يتم استخراج السمات باستخدام إجرائية التعلم بالنقل (Transfer Learning) بالاعتماد على النماذج المدربة مسبقاً(pre-trained). الهدف من استخدام التعلم بالنقل هو استخدام الخبرة المكتسبة من تدريب نماذج سابقة على مجموعة بيانات كبيرة لتحسين دقة التصنيف على مجموعة بيانات جديدة وأصغر، فبدلاً من تدريب النموذج من الصفر يتم استخدام نموذج مدرب سابقاً كنقطة انطلاق لتحسين أداء النموذج الجديد، يتم ذلك عن طريق استخدام طبقات الإدخال والتحويل والتجميع من النموذج المدرب سابقاً وتدريب طبقات التمثيل والإخراج على مجموعة البيانات الجديدة وهذا يؤدي إلى تقليل وقت التدريب وتحسين دقة النتائج. والنماذج المدرب سابقاً المستخدمة في هذا البحث لإنجاز عملية الوصف الصوتي هي: (vgg-16 & resnet-50).[16,17]

يمكن وصف بنيته المعمارية كما يلي[18]:

- طبقة الدخل (Input Layer): عبارة عن عينة دخل بحجم ٢٢٤*٢٢٤ بكسل
- الطبقات التلافيفية (Convolution Layers): تتميز بحقل استقبال صغير حجمه ٣*٣ بكسل وتم تصميمها لاستخراج الميزات والملامح الأساسية من الصوت.
- طبقات التجميع (max-pooling layers): تستخدم لتقليل حجم الإشارة وتحسين الأداء مما يساعد على تقليل تعقيد الحساب في الشبكة.
- طبقات الاتصال:(fully connected layers): تكون في نهاية الشبكة وتستخدم للتصنيف وذلك بالاعتماد على السمات التي تم استخراجها سابقاً.

سيتم بناء النموذج الخاص بالبحث بحيث يتكيف مع الهدف، باستخدام مكتبة كيراس يتم استيراد فقط الجزء التلافيفي الخاص باستخلاص السمات، ونقوم بتجميد بقية الطبقات ((dens)prediction& flatten) عن طريق ضبط البارمتر include_top = false وهذه نتيجة التنفيذ:

```
block3_conv3 (Conv2D)        (None, None, None, 256)   590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, None, None, 256)   0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, None, None, 512)   1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, None, None, 512)   2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, None, None, 512)   2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, None, None, 512)   0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, None, None, 512)   2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, None, None, 512)   2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, None, None, 512)   2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, None, None, 512)   0
=================================================================
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
_________________________________________________________________
```

الشكل(٥):طبقات النموذج vgg-16

يوضح الشكل(٥) طبقات بنية النموذج vgg-16 المستخدمة لاستخلاص السمات من هذا النموذج.

ثم نجعل جميع المخططات الطيفية بمقاس واحد (٢٢٤*٢٢٤) كمايلي:

```
#Load the image with the target size of (224, 224)
image = load_img('wave1.jpg', target_size=(224, 224))
```

وبما أنه لدينا مجموعة إشارات للتدريب نريد معالجتها لتكن n يتوجب جعل جميع هذه الإشارات بمقاس واحد لذلك نخزن هذه الإشارات في مصفوفة numpy ذات الشكل (n,224,224,3) باستخدام السطر البرمجي التالي:

```
# Convert the image to a numpy array
image = img_to_array(image)
```

ثم نقوم بعملية (preprocess) لهذه الإشارات لتكون دخل للنموذج vgg-16 عن طريق استخدام الأمر البرمجي التالي:

```
image = preprocess_input(image)
```

وأخيرا عملية استخلاص السمات من الإشارات وتخزينها في المتغير features:

```
#Extract features using the VGG model
features = vgg_conv.predict(image)
```

النقطة الأكثر أهمية هو أنه تم تدريب VGG-16 على الإشارات المعالجة مسبقاً للمخطط الطيفي (نقلاً عن ورقة VGG) كمايلي:

المعالجة الوحيدة التي نقوم بها هي طرح متوسط RGB المحسوبة على مجموعة التدريب من كل بكسل. للوصول إلى أقصى أداء قمنا بتطبيق نفس المعالجة المسبقة للوصول بها للأداء المطلوب قبل تقييم الشبكة.

بشكل عام يتميز النموذج VGG ببساطته وفعاليته لذلك تم استخدامه على نطاق واسع في مختلف مهام الرؤية الحاسوبية مثل تصنيف الإشارات والكشف عن الكائنات.

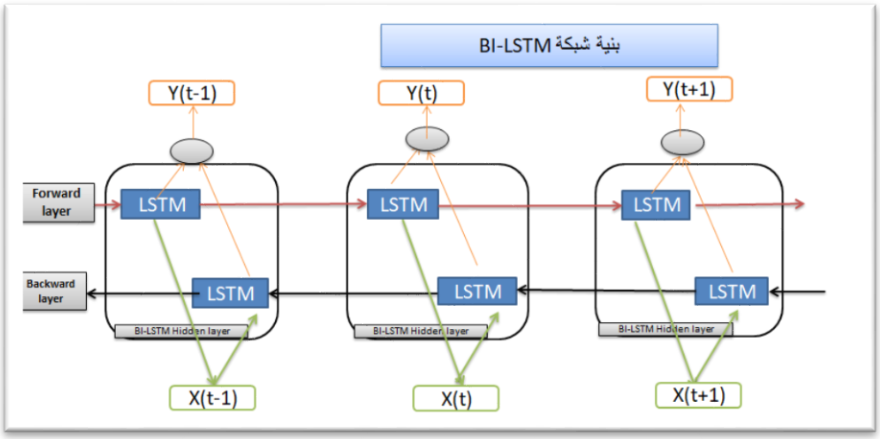
أما بالنسبة للنموذج ResNet-50 هو نموذج عميق مدرب مسبقاً من نماذج الشبكة العصبونية، تم تقديمه من قبل باحثين في مايكروسوفت عام ٢٠١٥. وتتمثل مزايا ResNets في أنه يمكن تدريب شبكات عدد كبير من طبقات العمق بسهولة، فهي لا تواجه تأثير التلاشي (مشكلة التدرج)، فهي تتعلم من رسم الخرائط المتبقية الأكثر توافقًا مع التحسين [19] ، وتتقارب بسهولة وتوفر تعميماً أفضل، كما يتميز هذا النموذج بطبقات (skip connection) أو (Residual connection) وهو عبارة عن اتصال مختصر بين طبقتين في الشبكة هذا الاتصال المختصر يسمح للشبكة بتخطي طبقات معينة والانتقال مباشرة إلى طبقات أعمق فهذا ساعد على تجنب مشكلة تدهور الأداء الذي يحدث في الشبكات العميقة [20]. يتكون نموذج Resnet50 من ٥٠ طبقة، وهو يعتبر من أحدث النماذج في مجال تصنيف النماذج. يتكون النموذج من عدة طبقات، بما في ذلك طبقات الإدخال والإخراج، وطبقات الانتقال العميق وطبقات التعلم العميق. تبدأ بنية النموذج بطبقة الإدخال (Input layer) التي تستقبل الإشارات المدخلة إلى النموذج. ثم تأتي طبقات الانتقال العميق (Convolutional layers) التي تساعد في استخلاص المعلومات المهمة من النماذج، وتحديد الملامح الأساسية لها. بعد ذلك، يأتي طبقة التعلم العميق (Deep learning layer) التي تستخدم تقنيات التعلم العميق لتحسين قدرة النموذج على التعرف على الأشياء وتصنيفها بدقة. وتستخدم هذه الطبقة تقنية الانتقال العميق لتحسين أداء التصنيف وتقليل الأخطاء. وتتضمن بنية النموذج أيضًا طبقات التجميع (Pooling layers) التي تستخدم لتقليل حجم النموذج وتسهيل عملية التصنيف. وتحتوي الطبقات الأخيرة من النموذج على طبقات الإخراج (Output layer) التي تقوم بإخراج نتائج التصنيف. ويتميز نموذج Resnet50 بأنه يحتوي على طبقات مكررة (Residual layers)، والتي تساعد في تحسين أداء التصنيف وتقليل الأخطاء، وذلك من خلال تجاوز بعض الطبقات المعقدة والتي قد تؤدي إلى فقدان المعلومات المهمة. ويتم استخراج السمات من هذا النموذج بشكل مشابه لخطوات استخراج السمات للنموذج السابق، ثم يتم إدخال السمات المستخلصة من النماذج المدربة مسبقاً إلى دخل شبكة BILSTM.

٥-٤:مرحلة بناء النموذج الهجين Bi-LSTM :

أولاً: Bi-LSTM

BiLSTM: (Bidirectional Long Short-Term Memory)، وهي نوع من الشبكات العصبية المتكررة (RNN) التي تتعلم الاعتمادية طويلة المدى بين خطوات الزمن من بيانات التسلسل أو السلاسل الزمنية. تستخدم (BiLSTM) شبكتين فرعيتين من (LSTM) واحدة تأخذ المدخلات في اتجاه متقدم (من اليسار إلى اليمين)، والأخرى تعمل في الاتجاه المعاكس (من اليمين إلى اليسار). وتحتوي كل من الطبقات الأمامية والخلفية على ٢٥٦ خلية لكل منهما، حيث يتم تسلسل المخرجات من الطبقات الأمامية والخلفية ل BiLSTM قبل تمريرها إلى الطبقة التالية. وهذا يزيد من كمية المعلومات المتاحة للشبكة، ويحسن من السياق

المتاح للخوارزمية (مثل معرفة ما الكلمات التي تسبق وتلي كلمة معينة في جملة). تستخدم BiLSTM في مجالات مختلفة مثل تحليل المشاعر، والتعرف على الكيانات المسماة، والإجابة على الأسئلة، وغيرها [21]. يتوجب إعادة تشكيل مخرجات (VGG-16 و ResNet50) بحيث تتناسب مع شكل مدخلات BiLSTM، يتم استخدام طبقة التجميع العالمية (global average pooling) أو طبقة التجميع الأقصى (global max pooling) بعد (VGG-16 و ResNet50)، وهذا سيقلل المخرجات إلى متجه ثنائي الأبعاد بشكل (batch_size, 512)، حيث يمثل كل إشارة متجهاً واحداً يحتوي على ٥١٢ سمة. ثم يتم استخدام طبقة تكرار المتجه (repeat vector) أو طبقة توزيع الزمن (time distributed) لتكرار هذا المتجه إلى تسلسل من خطوات زمنية، حتى تصبح المخرجات متجهًا ثلاثي الأبعاد بشكل (batch_size, timesteps, 512). ثم نقوم بإدخال هذا المتجه إلى دخل شبكة (BiLSTM). وهكذا تم استخدام شبكة (BiLSTM) كطبقة تعلّم من التسلسلات الزمنية، ونحصل على مخرجات بشكل (batch_size, 49, 64)، حيث يمثل كل إشارة تسلسلاً من ٤٩ متجهاً، كل منها يحتوي على ٦٤ سمة. هذه المخرجات تعبر عن التمثيل الزمني لكل إشارة. هذا الشكل يبين بنية BI-LSTM:



الشكل (٦):بنية BI-LSTM

ثانياً: Attention mechanism

هي تقنية تستخدم في نماذج التعلّم العميق، وتهدف إلى تحسين قدرة النماذج على التركيز على المعلومات الأكثر أهمية في البيانات المدخلة، وتجاهل المعلومات الغير مهمة. وتعتمد هذه التقنية على إضافة طبقة attention layer إلى النموذج، والتي تعمل على تحديد الأجزاء الأكثر أهمية في البيانات المدخلة. تعمل هذه التقنية عن طريق تعيين أوزان لأجزاء مختلفة من بيانات الإدخال، والتي تستخدم بعد ذلك لحساب المبلغ المرجح الذي يمثل المعلومات الأكثر صلة بالمهمة، يسمح هذا للنموذج بضبط تركيزه ديناميكياً أثناء معالجة المدخلات، ومنه تحسين دقة النماذج، وتقليل عدد الخطأ في التحليل، وذلك بتحديد الأجزاء الأكثر أهمية في البيانات والتركيز عليها. كما أن استخدام هذه التقنية يساعد في تحسين سرعة التحليل، حيث يتم تجاهل المعلومات الغير مهمة والتركيز على المعلومات الأكثر أهمية فقط. يتم الحصول على مدخلات آلية الانتباه من متجه الإخراج للطبقة العليا التي يتم تشغيلها بواسطة شبكة Bi-LSTM. وتلي طبقة الإدخال الطبقات مخفية (hidden layers) التي تعمل على استخراج المعلومات الأكثر أهمية من البيانات، ثم تأتي طبقة (attention layer) بعد الطبقات المخفية، وتعمل على تحديد الأجزاء الأكثر أهمية في البيانات المدخلة، وذلك بإعطاء

وزن أكبر لهذه الأجزاء في عملية التحليل، ثم طبقة الإخراج (output layer) التي تقوم بإخراج النتائج النهائية للنموذج.

٥-٥: التصنيف:

ثم نستخدم طبقة متصلة بالكامل (fully connected layer) لإضافة طبقة إخراج (output layer) بشكل (batch_size, 2)، حيث تمثل كل إشارة توزيعاً احتمالياً على الفئتين المستهدفتين المدخنين وغير المدخنين. تم استخدام دالة softmax لحساب هذا التوزيع الاحتمالي.

٥-٦: التدريب:

بعد الانتهاء من بناء النظام المقترح، سيتم تدريب كل نموذج من النماذج المستخدمة بالاعتماد على قاعدة البيانات التي تم إنشاؤها للمدخنين وغير المدخنين. سيتم حساب تابع الخسارة بالإضافة إلى الدقة ضمن مرحلة التدريب والتي ستستخدم لاحقاً في علمية تقييم أداء النظام. سيتم تكرار عملية التدريب ٣٠ مرة، وفي كل تكرار يتم تدريب النموذج وتحديث الوزن والمعاملات في الطبقات المخفية وطبقة الانتباه بشكل متكرر خلال عملية التدريب بشكل يتناسب مع النتائج.

٦- معايير التقييم:

تعد عملية التقييم مرحلة أساسية من مراحل تصميم وبناء نماذج التعلم العميق، فهو يساعد على فهم دقة توقع النموذج المستخدم للخرج ليكون مطابقاً للخرج الصحيح. إضافة إلى مهمته الأساسية في معرفة إذا كان بالإمكان أن نعتمد على النموذج ليستخدم في قطاع العمل المناسب أو يجب إضافة تحسينات أو تعديلات أو حتى الانتقال إلى نموذج أخر يحقق المطلوب. هناك عدة معايير تستخدم لتقييم أداء نموذج الشبكة العصبونية، لذلك يتوجب علينا اختيار المعايير المناسبة للنموذج المستخدم والذي يعطي نتائج واضحة.

معايير التقييم التي تم الاعتماد عليها هي:

- معدل الدقة (Accuracy): أثناء التدريب يتم حساب دقة التعرف على الخرج الصحيح من الخرج المتوقع لعينات التدريب (Dataset Training) وهذا ما يسمى دقة التدريب Accuracy (Training)
- الدقة (Precision): يقيس مدى قدرة النموذج على التنبؤ بشكل صحيح بالحالات الإيجابية، يمكن اعتبارها كمؤشر لدقة التصنيف في حالة وجود نتائج إيجابية.
- معدل الاستحواذ (Recall): يقيس مدى قدرة النموذج على التنبؤ بشكل صحيح بالحالات الإيجابية، ويتم حسابه عن طريق تقسيم عدد الحالات الإيجابية التي تم تصنيفها بشكل صحيح على عدد جميع الحالات الإيجابية.
- قيمة F1 (F1 Score): يقيس مدى قدرة النموذج على التنبؤ بشكل صحيح بين جميع الفئات المختلفة، ويتم حسابه عن طريق مزج مقاييس الدقة والاستحواذ.
- معدل الخطأ المتساوي Equal Error Rate (ERR): وهو معيار يستخدم لتقييم أداء نظام التعرف الصوتي. يتم حساب قيمته عند نقطة يكون فيها معدل القبول الخاطئ FAR مساوياً لمعدل

الرفض الخاطئ FRR. وهو مقياس مهم جداً لتقييم أداء النموذج في التصنيف بين فئتين مختلفتين بنفس الدرجة من الأهمية. وهو يقاس بالنسبة المئوية. على سبيل المثال، إذا كان EER يساوي ٥٪، فإن ذلك يعني أن النظام يرفض ٥٪ من المستخدمين الشرعيين ويقبل ٥٪ من المستخدمين غير المصرح بهم.

٧- النتائج والمناقشة:

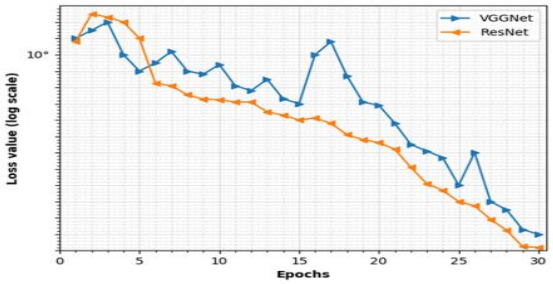
تظهر النتائج أن النموذج الهجين (ResNet-50 + BiLSTM-Attention) يوفر دقة أعلى من النموذج الهجين (VGGNet + BiLSTM-Attention).

الجدول التالي يبين تلخيص نتائج معايير التقييم المستخدمة للتصنيف بين أصوات المدخنين وغير المدخنين لكل نموذج حيث تم اعتماد ٨٠% من البيانات للتدريب و ٢٠% للاختبار:

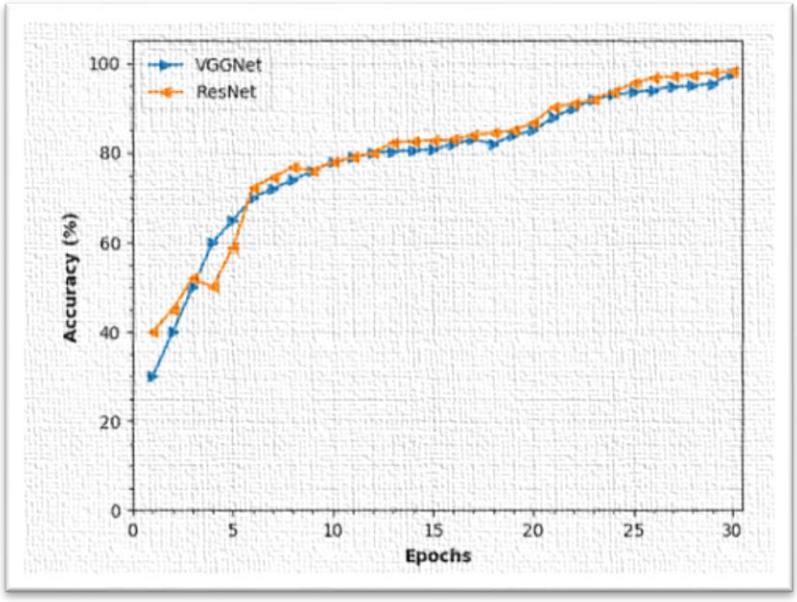
الجدول(١): مقارنة بين النموذجين المستخدمين وفق معايير التقييم المستخدمة

بارمتر التقييم (%) / model		معدل الدقة	الدقة	معدل الاستحواذ	F1 قيمة
VGG-16		%٩٣,٦	%٩٤,٣	%٩٣,٨	%٩٤
ResNet-50		%٩٤,٦	%٩٥,٦	%٩٥,٧	%٩٤,٨

ومن أجل فترات تدريبية مختلفة تتراوح من (٥ إلى ٣٠) تمت مقارنة النتائج بالاعتماد على تابع الخسارة وتابع الدقة:



الشكل(7):المقارنة بين النموذجين(vgg-16 و resnet-50) بالاعتماد على تابع الخسارة



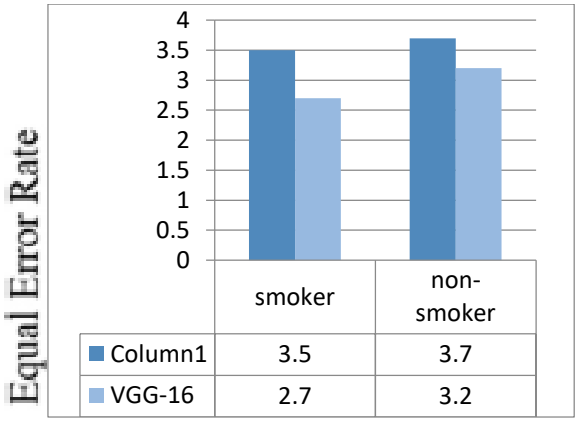
الشكل(٨):المقارنة بين النموذجين(vgg-16 و resnet-50) بالاعتماد على تابع الدقة

و لحساب ERR يجب علينا حساب عدد المفردات التي تم تعرفها بشكل صحيح (Nc)، وعدد المفردات التي لم يتم تعرفها بشكل صحيح (Nf). ثم نحسب ERR باستخدام المعادلة التالية[12]:

$$ERR = Nf / (Nc + Nf)$$

لحساب EER بشكل برمجي باستخدام Python، يمكن استخدام مكتبة Scikit-learn، وذلك عبر تحديد قيمة threshold التي تحقق معدل خطأ متساوٍ بين معدلات الخطأ النوعية والكمية، ومن ثم حساب معدل الخطأ المتساوي باستخدام دالة roc_curve ودالة interp1d من مكتبة Scipy.

تم تقييم مقياس EER للنموذجين المستخدمين. يوضح الشكل التالي مقارنة أداء النموذجين من ناحية (EER) لتصنيف أصوات المدخنين وغير المدخنين.



الشكل(٩): مقارنة النموذجين من ناحية معدل (EER)

٨- التوصيات:

هناك مجموعة من التوصيات المستقبلية تشير إلى الإجراءات والتحسينات التي يمكن اتخاذها لتطوير النموذج الهجين وتحسين أدائه في التمييز بين أصوات المدخنين وغير المدخنين. منها:

١. إضافة طبقات إضافية: يمكن استكشاف إضافة طبقات إضافية إلى النموذج الهجين لزيادة قدرته التمييزية.

٢. استخدام تقنيات تعلم عميق أخرى: يمكن استكشاف تقنيات تعلم عميق أخرى بجانب VGG16 و ResNet50 لتحسين أداء النموذج. على سبيل المثال، يمكنك استخدام شبكات التحويل الانتقالية (Transformers) التي تعتبر فعالة في معالجة المعلومات السمعية.

٣. استخدام مجموعات بيانات أكبر وأكثر تنوعاً: يمكن جمع مجموعات بيانات أكبر وأكثر تنوعاً لتحسين أداء النموذج. يمكن أن تساعد المجموعات البيانات الأكبر في تعزيز قدرة النموذج على التعرف على أصوات الجنس البشري بدقة أكبر.

٤. استخدام تقنيات التعديل والتحسين: يمكنك استخدام تقنيات التعديل والتحسين لتحسين أداء النموذج. على سبيل المثال، يمكن استخدام تقنيات التعديل مثل Dropout و Batch Normalization لتجنب الحالة الزائدة (Overfitting) وتحسين قدرة النموذج على التعامل مع بيانات جديدة.

٥. استخدام تقنيات تحسين الأداء: يمكنك استخدام تقنيات تحسين الأداء لتحسين أداء النموذج. على سبيل المثال، يمكن استخدام تقنيات التحسين مثل تعديل معدل التعلم (Learning Rate Scheduling) وتقنيات تحسين الأوزان (Weight Decay) لتحسين دقة التصنيف وتقليل الخطأ.

٦. تحليل الأخطاء وتحسينات النموذج: بتحليل أخطاء النموذج واستنتاج الأسباب المحتملة والتحسينات التي يمكن اتخاذها. يمكن أن تشمل هذه التحسينات تعديل البيانات، وتغيير معمارية النموذج، وتحسين تقنيات التدريب.

باستخدام هذه التوصيات، يمكن تحسين النموذج الهجين وزيادة قدرته على التمييز بين أصوات المدخنين وغير المدخنين بدقة أكبر.

المراجع المستخدمة:

- [1] Liu Z, Wu Z, Li T, Li J, Shen C. GMM and CNN hybrid method for short utterance speaker recognition. IEEE Trans Ind Inf 2018;14(7):3244–52.
- [2] Albadr MA, Tiun S, Ayob M, AL-Dhief FT. Spoken language identification based on optimised genetic algorithm–extreme learning machine approach. Int J Speech Technol 2019;22(3):711–27.
- [3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol.12, no. Aug, pp. 2493–2537, 2011.
- [4] Deshmukh, G., Gaonkar, A., Golwalkar, G., and Kulkarni, S."Speech based Emotion Recognition using Machine Learning." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 4 , no 4(2019).
- [5] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. "Speech emotion recognition using deep learning techniques: A review." IEEE Access 2, no. 7 (2019).
- [6] Salaba, A., & Chan, L. M. (2023). Cataloging and classification: an introduction. Rowman & Littlefield.
- [7] Zhu D, Li H, Ma B, Lee C-H. Optimizing the performance of spoken language recognition with discriminative training. IEEE Trans Audio Speech Lang Process 2008;16(8):1642–53.
- [8] Sun L, Chen J, Xie K, Gu T. Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition. Int Speech Technol 2018;21(4):931–40.
- [9] Zhang, Y., Wu, J., Qin, T., Xu, Z., Qu, S., Pan, L., ... & Xiao, Z. (2022). Comparison of the revised 4th (2016) and 5th (2022) editions of the World Health Organization classification of myelodysplastic neoplasms. Leukemia, 36(12), 2875-2882.
- [10] Roy P, Das PK. Comparison of VQ and GMM approach for identifying Indian languages. International Journal of Applied Pattern Recognition 2013;1 (1):99–107.
- [11] Jog AH, Jugade OA, Kadegaonkar AS, Birajdar GK (2018) Indian language identification using cochleagram based texture descriptors and ANN classifier. In: 2018 15th IEEE India Council International Conference (INDICON). IEEE, pp.1-6
- [12]] Chieregato, M., Frangiamore, F., Morassi, M., Baresi, C., Nici, S., Bassetti, C., ... & Galelli, M. (2022). A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. Scientific reports, 12(1), 4329.
- [13] Adeeba F, Hussain S. Native Language Identification in Very Short Utterances Using Bidirectional Long Short-Term Memory Network. IEEE Access2019;7:17098–110.
- [14] Sun L, Chen J, Xie K, Gu T. Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition. Int J Speech Technol 2018;21(4):931–40.

[15] Assari, Z., Mahloojifar, A., & Ahmadinejad, N. (2022). Discrimination of benign and malignant solid breast masses using deep residual learning-based bimodal computer-aided diagnosis system. Biomedical Signal Processing and Control, 73, 103453.

[16] arXiv:1512.03385v1 [cs.CV] 10 Dec 2015.Simonyan K, Zisserman A, " VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION", Conference Paper at ICLR, 2015.

[17] Xiao R, Tang H, Gu P, Xu X. Spike-based encoding and learning of spectrum features for robust sound recognition. Neurocomputing 2018;313:65–73.

[18] Sharan RV, Moir TJ. Acoustic event recognition using cochleagram image and convolutional neural networks. Appl Acoust 2019;148:62–6.

[19] Birajdar, G. K., & Raveendran, S. (2023). Indian language identification using time-frequency texture features and kernel ELM. Journal of Ambient Intelligence and Humanized Computing, 14(10), 13237-13250. https://doi.org/10.1007/978-981-33-6881-1_32.

[20] Romero-Rivas, C., & Costa, A. (2022). On the flexibility of the sound-to-meaning mapping when listening to native and foreign-accented speech. Cortex, 149, 1-15.

[21] Cieśla, K., Wolak, T., Lorens, A., Mentzel, M., Skarżyński, H., & Amedi, A. (2022). Effects of training and using an audio-tactile sensory substitution device on speech-in-noise understanding. Scientific Reports, 12(1), 3206.