SPEECH EMOTION ANALYSIS SYSTEM- Using Deep Learning and Natural Language Processing

Dr.D.Rasi ,Associate Professor
Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatore,India

Dr.T.LathaMaheswari ,Associate Professor
Computer Science and Engineeering
Sri Krishna College of Engineering and Technology
Coimbatore,India

Abinaya P
Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatore, India

Agalya T
Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatire,India

Harshitha C.S
Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatore, India

ABSTRACT Understanding human emotions through vocal expressions is a crucial aspect of artificial intelligence-driven speech analysis. This project explores deep learning techniques for Speech Emotion Recognition (SER) using voice-based cues such as tone, pitch, and rhythm. The proposed approach employs a 1D Convolutional Neural Network (CNN) designed for multiclass emotion classification.

The model is trained on the CREMA-D dataset, which contains 7,442 audio clips from 91 actors expressing six fundamental emotions: happiness, sadness, anger, fear, disgust, and neutrality. To enhance model performance, various data augmentation techniques, including noise injection, time stretching, shifting, and pitch modulation, are applied. Key speech features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and Mel Frequency Cepstral Coefficients (MFCC) are extracted for improved emotion classification. A segment duration of 2.5 seconds with a 0.6-second offset is used to maximize relevant emotional information.

The developed deep learning model achieves an accuracy of 95.52% on test data, demonstrating its efficiency in recognizing emotions from speech. This high-performance SER system has potential applications in mental health assessment, customer service enhancement, human-computer interaction, and AI-driven personalized experiences. By leveraging deep feature extraction and robust classification, the proposed model contributes to the advancement of emotion-aware AI systems in various industries.

INDEX TERMS Human-computer interaction, Deep learning, speech emotion recognition, convolutional neural networks, vision transformer, mel spectrogram.

INTRODUCTION

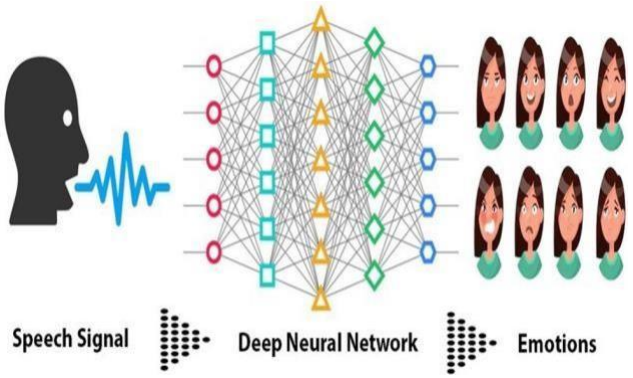

FIGURE 1. Basic speech emotion classification overview].

Traditionally, voice-based emotion recognition relied on rule-based and statistical models that used handcrafted features. Early approaches employed signal processing techniques such as spectral analysis, pitch contour mapping, and prosody-based classification. Classical machine learning methods, including Support Vector Machines (SVM), Hidden Markov Models (HMM), and Gaussian Mixture Models (GMM), were widely used to categorize emotions. These techniques required extensive feature engineering, where domain experts manually extracted relevant acoustic properties. While effective to some extent, traditional methods struggled with generalization across different speakers, languages, and recording conditions, limiting their scalability and robustness. The integration of deep learning in voice-to-emotion recognition has led to remarkable improvements in accuracy and efficiency. Modern neural networks can process vast amounts of speech data with high precision, enabling real-time emotion classification. This advancement has expanded the scope of applications, from AI-driven mental health monitoring and sentiment analysis in call centers to emotion-aware voice assistants and adaptive gaming experiences. Furthermore, combining deep learning with other AI technologies, such as natural language processing (NLP) and multimodal emotion recognition, has opened new possibilities for creating empathetic AI systems capable of understanding human emotions more holistically.

One of the earliest approaches to voice-based emotion detection involved rule-based systems, where specific speech characteristics, such as pitch, intensity, and duration, were analyzed manually. These systems depended on predefined thresholds to classify emotions, making them highly dependent on domain expertise. However, rule-based methods lacked flexibility and struggled with variations in speaker accents, background noise, and dynamic emotional expressions.

Another widely used method involved signal processing techniques that extracted acoustic features like prosody (intonation and rhythm), spectral properties, and formant frequencies. Researchers applied methods such as Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), and pitch contour analysis to represent speech signals in a structured manner. The following is how the rest of the work is organized: we present different deep transfer learning and 1DCNN architectures as they relate to speech emotion classification. As voice-to-emotion recognition continues to evolve, deep learning remains at the forefront of innovation, driving the development of intelligent systems that can perceive and respond to human emotions with unprecedented accuracy. By bridging the gap between human expression and machine understanding, this technology has the potential to revolutionize industries and redefine human-computer interaction in the years to come.

The main contributions of this review paper are outlined below:

- We present an unambiguous review of various cutting-edge deep learning (CNN) architectures for speech emotion classification, including their pros & cons.
- We also study the speech emotion corpora and their application in various types of research in speech emotion classification, through which the readers can be thoroughly furnished with the popular speech emotion dataset and why.
- Identification of salient and discriminating features from speech samples using a state-of-the-art deep learning model with a feature selection mechanism was also highlighted.
- We also provide a succinct summary of various deep learning methodologies and their effectiveness in state-of-the-art speech emotion classification nowadays. Emerging trends and future directions in speech emotion classification, open challenges and limitations of current approaches and the potential impact of advancements in deep learning model on Speech emotion classification performance are not left out.

II. DEEP LEARNING

Deep learning (DL) is a branch of machine learning that can handle nonlinear datasets. It is, in most cases, interchangeably used as Deep Neural Network(DNN). Deep Layers of stacked nodes make up a DNN, which is often trained through back-propagation and optimization techniques. Each layer has an activation function and corresponding weights. Deep learning has rapidly expanded over the last two decades and is now utilized in many facets of our everyday lives . Since 2011, Convolutional Neural Network (CNN) layers, for example, have enhanced deep learning models for computer vision and pattern recognition-related tasks, and as of now, the majority of DLs include CNN layers, which form the basis of Deep Convolutional Neural Network (DCNN).

III. DESIGN OF PROPOSED SYSTEM AND WORKFLOW

The proposed system leverages deep learning techniques to enhance the accuracy and efficiency of voice-to-emotion recognition. Unlike traditional approaches that rely on handcrafted feature extraction and conventional classifiers, this system utilizes a 1D Convolutional Neural Network (CNN) to automatically learn patterns from raw speech signals. The model is trained on the CREMA-D dataset, which contains 7,442 audio clips featuring emotional expressions from 91 actors across six distinct emotions: happiness, sadness, anger, fear, disgust, and neutrality.

To improve the robustness of emotion recognition, the system incorporates data augmentation techniques, including noise injection, time stretching, shifting, and pitch modification. These transformations create variations in the dataset, making the model more resilient to different recording conditions, speaker variations, and background noise. Additionally, key audio features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and Mel Frequency Cepstral Coefficients (MFCC) are extracted to enhance classification performance. A 2.5-second audio segment with a 0.6-second offset is used, ensuring that the most relevant emotional content is captured while excluding irrelevant portions.

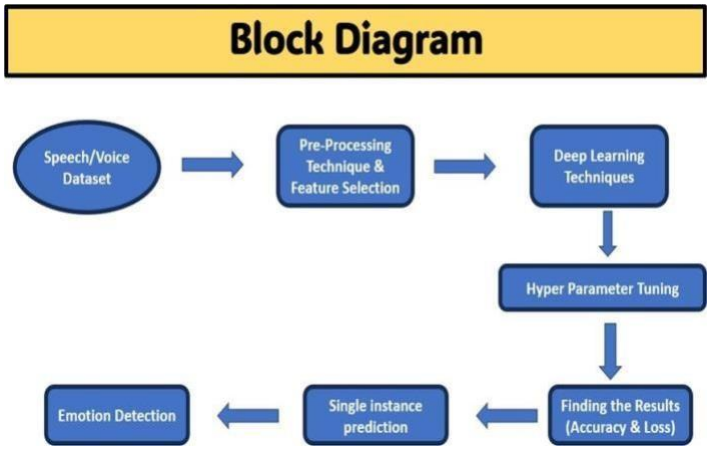

FIGURE 2. Process Flow of Emotion Analysis Using DCNN.

A. DATASET SELECTION AND PREPROCESSING

The project began with the selection of the CREMA-D dataset, which contains 7,442 audio clips featuring emotional expressions from 91 actors across six emotions: happiness, sadness, anger, fear, disgust, and neutrality. The dataset was analyzed to understand its structure, duration of audio samples, and speaker diversityng using a specific cost function. The training process is hindered by the problem of vanishing gradients, which reduces the effectiveness of error updates in the earlier layers. To enhance the quality and generalization of the dataset, preprocessing steps were implemented:

- Noise Reduction: Background noise was minimized to improve clarity.
- Segmentation: Audio clips were standardized to 2.5-second segments with a 0.6-second offset to focus on emotional content while discarding irrelevant sections.
- Feature Extraction: Key speech characteristics such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and Mel Frequency Cepstral Coefficients (MFCC) were extracted for better representation of vocal expressions.

	Emotion	Path
0	neutral	/content/drive/My Drive/dataset/speech emotion...
1	sad	/content/drive/My Drive/dataset/speech emotion...
2	angry	/content/drive/My Drive/dataset/speech emotion...
3	angry	/content/drive/My Drive/dataset/speech emotion...
4	happy	/content/drive/My Drive/dataset/speech emotion...

FIGURE 3. DataSet Selection and Processing.

B. DATA AUGMENTATION FOR ROBUSTNESS

To prevent overfitting and enhance the model's ability to recognize emotions across varied speakers and environments, data augmentation was applied. The following techniques were implemented:

- Noise Injection: Random noise was added to simulate different recording conditions.

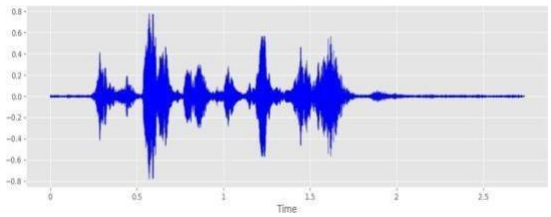

FIGURE 4. Noise Injection.

- Time Stretching: Speech was slightly sped up or slowed down without altering pitch.

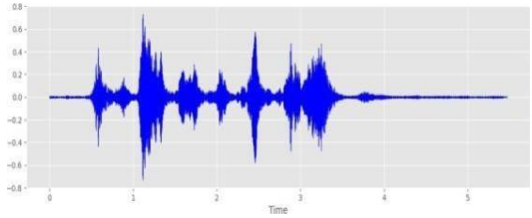

FIGURE 5. Time Stretching

- Pitch Shifting: The pitch of audio clips was modified to introduce variations in voice tone.
- Shifting: The audio signal was shifted slightly forward or backward to diversify the dataset.

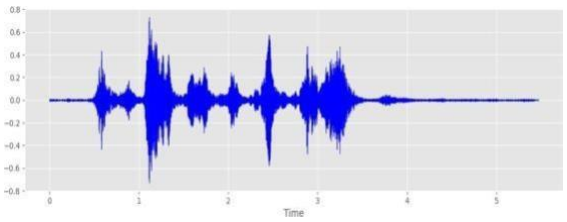

FIGURE 6. Pitch Shifting

These augmentations created new training samples, increasing dataset variability and improving model robustness.

C. FEATURE EXTRACTION FROM THE AUDIO/SPEECH

Extracting key acoustic features from speech is essential for distinguishing emotions. The model utilizes:

- Zero Crossing Rate (ZCR): Measures frequency of sign changes in the audio signal, indicating energy variations.
- Root Mean Square Energy (RMS): Captures the intensity of speech.
- Mel Frequency Cepstral Coefficients (MFCCs): Represents the spectral properties of the audio, crucial for emotion differentiation.

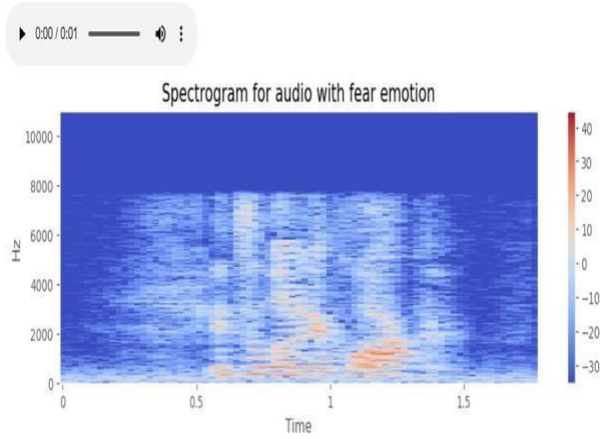

FIGURE 7. Spectrogram for Different Emotion(Fear)

The integration of acoustic features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and Mel Frequency Cepstral Coefficients (MFCCs) allows for efficient emotion differentiation, improving classification accuracy. The structured approach of data preprocessing, feature extraction, and deep learning-based classification ensures a high-performing and scalable SER system.

The proposed model has significant real-world applications, ranging from mental health assessment and AI-driven customer service to human-computer interaction and personalized virtual assistants. By bridging the gap between vocal expressions and emotional understanding, this work contributes to the advancement of emotion-aware AI systems. Future improvements could involve expanding the dataset, refining deep learning architectures, and integrating real-time speech recognition for more adaptive and responsive applications.

E. MODEL EVALUATION AND OPTIMIZATION

To Once trained, the model is tested on unseen data to assess its accuracy and effectiveness. Key evaluation metrics include:

- Accuracy (95.52%)
- Precision, Recall, and F1-Score
- Confusion Matrix Analysis to examine classification errors
- Evaluate the models using the Following formulas:

1. Accuracy

$$Accuracy = TP+TN \ / \ TP+TN+FP+FN$$

TP (True Positive): Correctly predicted positive cases

TN (True Negative): Correctly predicted negative cases

FP (False Positive): Incorrectly predicted positive cases (Type I error)

FN (False Negative): Incorrectly predicted negative cases (Type II error)

2. Precision (Positive Predictive Value)

$$Precision=TP \ / \ TP+FP$$

Measures how many of the predicted positive cases are actually positive.

3. Recall (Sensitivity or True Positive Rate)

$$Recall=TP \ / \ TP+FN$$

Measures how many actual positive cases were

correctly identified.

4. F1 Score (Harmonic mean of Precision and Recall)

$$F1Score = 2 \times Precision \times Recall / Precision + Recall$$

Balances Precision and Recall, especially useful when dealing with imbalanced datasets.

	precision	recall	f1-score	support
angry	0.95	0.96	0.96	1049
disgust	0.95	0.96	0.95	1007
fear	0.96	0.94	0.95	991
happy	0.94	0.94	0.94	1000
neutral	0.96	0.96	0.96	892
sad	0.97	0.97	0.97	1015
accuracy			0.96	5954
macro avg	0.96	0.96	0.96	5954
weighted avg	0.96	0.96	0.96	5954

FIGURE 8. Model Evaluation Using formulas

Further optimizations, such as hyperparameter tuning and regularization techniques, are applied to enhance performance.

IV. 1D CONVOLUTION NEURAL NETWORK(CNN) FOR MULTICLASS CLASSIFICATION

A 1D Convolutional Neural Network (CNN) was chosen for its efficiency in processing sequential speech data. The architecture was designed to:

- Extract hierarchical audio features automatically.
- Reduce the need for manual feature selection.
- Improve computational efficiency by processing raw waveforms directly.

A. CNN ARCHITECTURE AND ACTIVATION FUNCTION

Conv1D Layers – Extract deep hierarchical patterns from speech features.
Batch Normalization – Stabilizes activations and speeds up training.
MaxPooling1D – Reduces dimensionality while preserving key information.
Flatten Layer – Converts feature maps into a vector for classification.
Dense Layers (Fully Connected) – High-level feature learning before classification.
Softmax Activation – Outputs probability distribution for emotion categories.

Activation Formula:

f(x)=max(0,x)

A typical **CNN architecture** consists of the following :

1. **Input Layer**:
   - The input image is represented as a matrix of pixel values (e.g., 28×28×1 for grayscale or 224×224×3 for RGB images).
2. **Convolutional Layer**:
   - Applies filters (kernels) to extract features like edges, textures, and patterns.
   - The mathematical operation used is **convolution**.
   - Formula: Z=W∗X+bZ = W \ast X + bZ=W∗X+b where W is the filter (kernel), X is the input, b is the bias, and ∗ represents convolution.
3. **Activation Function (ReLU)**:
   - Introduces non-linearity by applying: f(x)=max(0,x)f(x) = \max(0, x)f(x)=max(0,x)
   - Helps the network learn complex patterns.
4. **Pooling Layer (Downsampling)**:
   - Reduces the spatial dimensions while retaining important features.
   - Types:
     - **Max Pooling** (takes the maximum value in a region).
     - **Average Pooling** (takes the average value in a region).
5. **Fully Connected (FC) Layer**:
   - Flattens the feature maps and connects to dense layers for classification.
6. **Softmax or Sigmoid Layer**:
   - **Softmax** is used for multi-class classification.
   - **Sigmoid** is used for binary classification.

B. KEY FEATURES OF 1D-CNN MODEL

➔ Deep 1D-CNN Architecture for Superior Feature Learning:

Multiple Conv1D Layers extract deep hierarchical patterns from the extracted speech features.
Batch Normalization stabilizes training and accelerates convergence.
MaxPooling Layers reduce dimensionality while retaining essential information.
Fully Connected Layers ensure high-level feature representation before classification.

➔ To prevent overfitting and ensure the model generalizes well, the following techniques are

implemented:

- Batch Normalization – Ensures stable activations and accelerates learning.
- Dropout Layers (if added) – Reduces overfitting by randomly disabling neurons during training.
- Learning Rate Scheduling – Adjusts the learning rate dynamically for better convergence.
- Early Stopping – Prevents unnecessary training and stops when optimal performance is reached.

The model was trained on the augmented dataset using optimized hyperparameters, including:

- Batch size: Controlled training stability.
- Learning rate adjustments: Fine-tuned for faster convergence.
- Dropout layers: Added to prevent overfitting.features from speech signals is quite challenging.

Traditional and primitive speech features (quality of voice and pitch) can be extracted in a handcrafted manner, but this may not be applicable in deep transfer learning. Majorly, the
popularly used feature extraction approach that exists in many
research works for the classification of speech emotion is known as acoustics features.

1. Temporal (Time-Domain) Features

Features extracted from the waveform directly.

(a) Zero Crossing Rate (ZCR)

ZCR=1N−1∑n=1N−11[s(n)s(n−1)<0]ZCR=
\frac{1}{N-1} \sum_{n=1}^{N-1} \mathbb{1}

[s(n) s(n-1) < 0]ZCR=N−11n=1∑N−11[s(n)s(n−1)<0]

- Measures how often the signal crosses zero.
- Used in **speech/music classification** (high for noisy signals, low for voiced sounds).

(b) Energy & RMS Energy

Energy=∑n=1Ns(n)2

- Represents the **loudness** of the signal.
- RMS (Root Mean Square) Energy is the square root of the average squared signal amplitude.

2. Frequency-Domain Features

C. ACOUSTIC FEATURES IN SPEECH AND AUDIO PROCESSING

Acoustic features can be categorized into spectral, prosodic and voice quality features. Spectral features are commonly used in speech emotion classification . Windowing the speech signal to disintegrate it into frames is the first step in computing MFCC, after which FFT is used on the frame to locate the power spectrum for every frame. The power spectrum is subsequently processed using a filter bank (mel-scale). After applying the Discrete Cosine Transform (DCT), the MFCC vector representation is produced.

Acoustic features are measurable properties of sound used in **speech recognition, music analysis, speaker identification, and emotion detection**. These features capture important information about **pitch, energy, frequency, and timbre**.

Moreover, in speech feature extraction, the presence of noise has been identified as the major factor that degrades the intelligibility of speech utterances. Therefore, the removal of these noises from raw speech signals while maintaining the emotional content is pertinent. The use of spectral subtraction and MEL filter approaches are the most common techniques for ensuring that the emotional quality of speech signal is retained, even though the background is not clear.

Extracted using **Fourier Transform (FFT)**.

(a) Spectral Centroid

Centroid=∑kfkS(k)∑kS(k)

Represents the "center of mass" of the spectrum (brightness of sound).

(b) Spectral Bandwidth

BW=∑k(fk−Centroid)2S(k)

Measures the spread of frequencies around the centroid.

(c) Spectral Roll-off

- The frequency below which **85%-95% of the total spectral energy** is contained.
- Differentiates **speech from music** and classifies genres.

3. Cepstral Features (MFCCs, LPCs)

Derived from frequency-domain features to

capture **human perception of sound**.

(a) Mel-Frequency Cepstral Coefficients (MFCCs)

- Mimics the human **auditory system** by using the Mel scale.
- Computed by:
  1. **Apply FFT** to get the spectrum.
  2. **Use Mel filter banks** to mimic the human ear.
  3. **Take logarithm and apply Discrete Cosine Transform (DCT)**.

(b) Linear Predictive Coding (LPCs)

- Models the human vocal tract as a filter.
- Used for **speech compression and** subjected to speech processing, especially, for the removal of noise, but when a non-acted speech dataset is to be used for emotion classification, the removal of background noise will have an effect on the result to be obtained.

D. ACCURACY CALCULATION IN SPEECH EMOTION ANALYSIS

As research Accuracy is a crucial metric in evaluating the performance of a Speech Emotion Recognition (SER) model, as it measures the proportion of correctly classified emotions out of the total predictions made. The accuracy of the 1D Convolutional Neural Network (CNN) used in this project is determined by comparing the model's predicted emotions to the actual labels in the test dataset.

- Accuracy is calculated using the following formula:

$$Accuracy = \frac{Correctly\ Predicted\ Samples}{Total\ Samples} \times 100$$

Where:

- Correctly Predicted Samples refers to the number of test instances where the predicted emotion matches the actual emotion.
- Total Samples is the total number of test instances used for evaluation.

E. CONFUSION MATRIX ANALYSIS

To gain deeper insights into model performance, a confusion matrix is used. It provides a breakdown of how many times each emotion was correctly or incorrectly classified. Metrics derived from the confusion matrix, such as Precision, Recall, and F1-score, further evaluate the model's reliability.

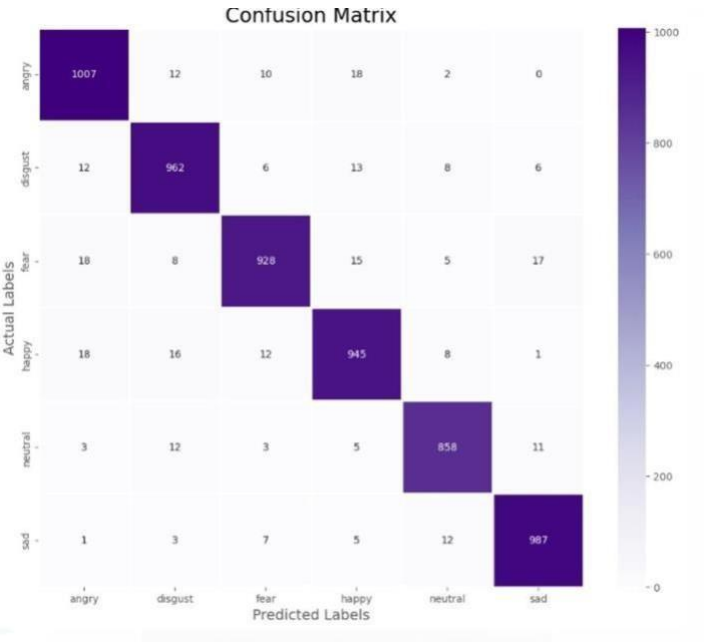

FIGURE 9. Confusion matrix analysis

In real life scenario, the feature selection phase takes input from the feature extraction phase under the deep learning model. Feature selection can be categorized into three major groups: filtered-based (e.g. correlation-based feature selection), wrapper-based and intrinsic feature selection. In redundant features were eliminated using the contribution analysis feature selection approach with Neural Network(NN).Their method proved the significance of feature selection in the classification of speech emotion. An accuracy of over 90% was achieved on the Berlin Emotional dataset.

V. OVERVIEW OF SPEECH EMOTION DATABASES

The success of deep learning in speech emotion classification rests heavily on the availability of speech samples (corpus) that can be used to train the deep learning model. Unlike other machine learning tasks and image processing (e.g. facial emotion), the speech utterance or training dataset requires labelling by hand through a human agent and the mode of perception differs from one person to the other. Therefore, it is necessary to have more than one person carrying out this task, to have an accurate label dataset . The quality of the dataset is proportional to the result of classification or prediction. In speech emotion classification, three major

The synthetic datasets (RAVDESS, EMO-DB, etc) are the acted speech samples collected from the professional speaker (actors) in a confined environment to extract emotion. They are artificially generated rather than a real world scene. The synthetic dataset is the most popular and widely available for speech emotion

classification tasks. Semi-natural is close to natural speech datasets, but they have elements of synthesis, e.g. NIMITEK. The last category is natural speech corpora, which capture the real-life scenario of human emotion. They can be obtained from TV shows, call centres, and online videos . However, there is limited availability of this category of speech datasets because of license issues. It is also prone to environmental noise, which must first be removed before it yields accurate results in emotion classification using a deep learning model. Also, it is obvious within the SEC community that there is a multiplicity of speech emotion datasets, but the question of standard measures for these datasets is still lingering to date. Below is a comprehensive description of some of the available speech emotion databases, and table 2 shows the summary of them all.

A. RAVDESS DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a speech corpus released under a Creative Commons license . It has open access for use in research. It captures eight different emotions (angry, sad, disgust, happy, fearful, calm, neutral and surprise) from the speech utterance, of twenty-four (24) professional actors. Twelve of these actors were male and twelve were female as well. Apart from additional neural emotional utterances, each utterance is recorded in two different categories of emotional strength high or low. It consists of 7356 files, evaluated ten times on emotional authenticity, reliability and validity by mature researchers from North America . Access to it is open for public use. Each file from 7356 total files has a distinct filename (e.g. 02-03-05-01-03-02-10.mp4), that represents the modality, vocal channel, actual emotion, intensity of the emotion and actor's gender (even number for female and odd number for male). RAVDESS falls

B. TESS DATASET

Toronto Emotional Speech Set is a publicly available speech emotion database that has been used by many researchers for the classification of emotion. The speech samples were recorded at No 6, Northwestern University Auditory in 2010 . Two actresses were requested to repeat a few hundred words throughout the spontaneous event, and their voices were captured. Seven distinct emotions were captured during the event (happy, angry, fear, disgust, pleasant, surprise, sad and neutral). A total of 2,800 files that depict human emotion were collected. It is audio-based only.

C. SAVEE DATASET

The Surrey Audio-Visual Expressed Emotion database is a unique emotional corpus recorded at CVSSP's 3D vision laboratory at the University of Surrey in 2010. Four actors, who were native English speakers and educationists (students and researchers) whose ages ranged from 27 to 31, were involved in the event. The English speakers were labelled as DC, JE, JK and KL. The recording took several months to establish the authenticity of the speech samples and gesticulation. Altogether, seven emotions were captured by this under a synthetic dataset where actors were producing audio speech of different emotional displays.

The interactive emotional dyadic motion capture database (IEMOCAP) is a speech corpus motivated by the fact that human emotion resides not only in speech but the combination of speech utterance and physiological ges ture. The speech samples were recorded by the Speech Analysis and Interpretation Laboratory (SAIL) located at Southern California University. Ten actors were recorded in interactional sessions during planned and unscripted verbal communication scenarios with labels on the strategic parts of their bodies (faces and hands), which provided extensive information about their body language and kinesics . The actors acted out emotional scripts for over twelve hours while also creating fictitious scenarios meant to elicit particular emotions (happy, furious, sad, frustrated, and neutral mood). This corpus has con tributed immensely to the development of multimodal design for emotion classification because of its size, interactivity and holistic elicitation of emotion.

D. EMO-DB DATASET

German emotional database (Emo-DB) is a recorded speech utterance (535) that contains 800 files spoken by ten professional actors where five were males and five were females respectively. It was initiated by the ICS, Technical University, Berlin, German .This corpus captures seven unique human emotions: bored, happy, sad, disgust, fear, angry and neutral. Prior to being resampled (down-sampled) to 16 kHz, the speech sample was originally recorded at a sampling rate of 48 kHz. Each file follows a unique naming pattern like:

database which are: anger,disgust, fear, happy, pleasant surprise, sad, and neutral. This corpus has a total size of 480 utterances, with a 44.1 kHz audio sampling rate. However, the database has restricted access, unlike others that are publicly available for research purposes.

E. CREMA-D DATASET

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA) is an audio-visual and large dataset (7,442 audio clips) suitable for multi-modal human emotion classification. The corpus captures six basic emotional conditions from the facial and spoken utterances of 91 actors, where 48 were males and 43 were females. The corpus is a cross-language comprised of actors from diverse continents (America, Africa, Asia and Europe) to avoid cultural or ethnic barriers that may lead to misclassification of emotion. A total number of 2443 people participated in rating 90 distinct audio clips. It is one of the largest publicly available corpus in use.

F. AESDD DATASET

Acted Emotional Speech Dynamic Database (AESDD) is a synthetic and publicly available corpus for the classification of emotion. It is a Greek-based emotional utterance, that captures five basic emotions

of anger, disgust, happy, sad and fear. The first version was released in 2018 with an accuracy rate of about 74% by listeners. The level of usage by researchers for this emotional corpus is very low due to solitary language, compared to English-based speech emotion corpus.

G. eNTERFACE DATASET

eNTERFACE is a German-based audio-visual speech database with open access. It is a multi-modal emotion database created for deep learning implementation in emotion classification and human-computer interaction. It was recorded in 2005 by Olivier Martin in collaboration with the eNTERFACE'05 workshop at TCTS Lab, Belgium. The project was fully funded by the Wallon region, Belgium, with contact number EPH3310300R0312/215286. The emotional corpus involved 42 actors (19% female and 81% male) from 14 different countries. A total of 1166 files comprised of six basic emotions were captured. Emotions were elicited by subjecting each actor to an atmosphere or condition where real human emotional outbursts could be captured.

One of the most popular databases in this sector has been used by scholars. Its applications include evaluating SVM methodologies both alone and in conjunction with hidden Markov models. Additionally, it has been used for cross corpus validation, unique deep learning techniques and gender-based emotion recognition . As a component of the VAESS project (Voice Attitudes and Emotions in Speech Synthesis), this database was created.

VI. MODEL TRAINING AND TESTING LOSS AND ACCURACY

1. Training and Validation: The model is trained using a subset of the dataset, while another subset is used for validation to fine-tune parameters and reduce overfitting.
2. Testing: After training, the model is tested on unseen audio samples to assess generalization.
3. Prediction Comparison: The model's predicted emotion labels are compared with the actual labels from the dataset.
4. Accuracy Computation: Using the accuracy formula, the final performance score is obtained.

Training vs. Testing (Validation)

Training Set

- The data used to train the model.
- The model learns patterns and adjusts weights to minimize loss.

Testing (Validation) Set

- Unseen data used to **evaluate** model performance.
- Helps check if the model is generalizing well.

Loss in Training & Testing

What is Loss?

- Loss is a **numerical value** that measures how far the model's predictions are from the actual values.
- Lower loss = **better** model performance.

Common Loss Functions

1. For Regression:
   - Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

   - Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

2. For Classification:
   - Cross-Entropy Loss (Log Loss)

$$Loss = -\sum_{i} y_i \log(\hat{y}_i)$$

Accuracy in Training & Testing

What is Accuracy?

- Accuracy measures the percentage of **correct predictions** over total predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \times 100\%$$

- **Higher accuracy = better performance**.

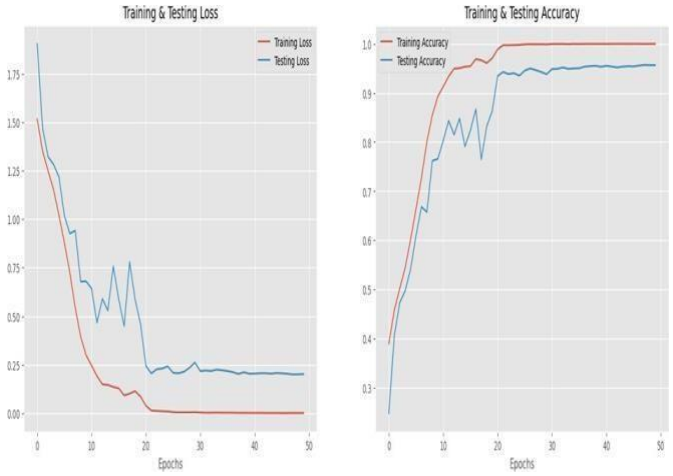

FIGURE 10. Training and Testing Loss&Accuracy

ACHEIVED ACCURACY

The In this project, the SER model achieved an accuracy of 95.52%, indicating its high effectiveness in recognizing emotions from speech signals. This high accuracy is attributed to the use of advanced feature extraction, data augmentation techniques, and deep learning-based classification. Further improvements can be explored by incorporating more diverse datasets, refining the model architecture, and optimizing hyperparameters.

FIGURE 11. Accuracy Graph

The accuracy of the 1D Convolutional Neural Network (CNN) used in this project is determined by comparing the model's predicted emotions to the actual labels in the test dataset.

VII. FINAL PREDICTION OF SPEECH EMOTION ANALYSIS SYSTEM

Output Format:

The system typically outputs either a probability distribution across a set of emotion classes (e.g., happy, sad, angry, neutral) or a direct label based on the highest probability.

- *Example:* Given an input audio clip, the system might assign 70% probability to "happy," 20% to "neutral," and 10% to "sad," with the final prediction being "happy."

- **Decision Thresholds:**

In some cases, confidence thresholds can be implemented to decide whether to trust a prediction or to mark it as uncertain for further review.

- **Real-Time vs. Batch Prediction:**

- **Real-Time:** For interactive systems, predictions are made on the fly to provide immediate feedback.
- **Batch Processing:** In research or analytics applications, multiple recordings might be processed together to analyze trends or aggregate emotional responses.

	Actual	Predicted
0	angry	angry
1	angry	angry
2	fear	fear
3	sad	sad
4	neutral	neutral
...	...	...
5949	neutral	neutral
5950	disgust	disgust
5951	angry	angry
5952	happy	happy
5953	happy	happy

5954 rows × 2 columns

FIGURE 12. Final Prediction of emotion

VIII. CONCLUSION

This project demonstrates the effectiveness of deep learning techniques in Speech Emotion Recognition (SER) by analyzing vocal expressions. Using a 1D Convolutional Neural Network (CNN) trained on the CREMA-D dataset, the model successfully classifies emotions with an impressive 95.52% accuracy. Through data augmentation techniques like noise injection, time stretching, shifting, and pitch modulation, the model's robustness is enhanced, ensuring it performs well across diverse audio samples.

The integration of acoustic features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and Mel Frequency Cepstral Coefficients (MFCCs) allows for efficient emotion differentiation, improving classification accuracy. The structured approach of data preprocessing, feature extraction, and deep learning-based classification ensures a high-performing and scalable SER system.

The proposed model has significant real-world applications, ranging from mental health assessment and AI-driven customer service to human-computer interaction and personalized virtual assistants. By bridging the gap between vocal expressions and emotional understanding, this work contributes to the advancement of emotion-aware AI systems. Future improvements could involve expanding the dataset, refining deep learning architectures, and integrating real-time speech recognition for more adaptive and responsive applications.

IX. FUTURE ENHANCEMENT AND APPLICATIONS

Data Augmentation and Diversity:

- **Increased Dataset Size:** Incorporate more diverse and representative speech samples to capture a wide range of accents, dialects, and speaking styles.
- **Augmentation Techniques:** Use noise injection, pitch alteration, and speed variations to improve the robustness of the model under different conditions.

Advanced Model Architectures:

- **Deep Learning Improvements:** Explore more complex architectures like CNN-LSTM hybrids or transformers, which can better capture temporal dynamics and contextual cues in speech.
- **Attention Mechanisms:** Integrate attention layers to focus on the most emotionally significant parts of the audio signal.

Multimodal Fusion:

- Combine audio with other modalities such as facial expressions (video) and textual content (transcripts) to enhance the accuracy and contextual understanding of emotions.

Robustness to Environmental Variability:

- **Noise Reduction:** Implement more sophisticated noise filtering and speech enhancement techniques to maintain performance in noisy or variable environments.
- **Domain Adaptation:** Use techniques that allow the model to adapt to new speakers, languages, or acoustic environments without extensive retraining.

Explainability and Interpretability:

- Develop methods to explain the model's predictions (e.g., which segments of the audio most contributed to a particular emotion) to improve trust and usability in sensitive applications.

Real-Time Implementation and Edge Computing:

- Optimize models for real-time performance on low-power devices, enabling deployment in mobile and embedded systems.

REAL-WORLD APPLICATIONS

Customer Service and Call Centers:

- **Emotion Monitoring:** Automatically detect customer frustration or satisfaction, allowing supervisors to intervene or tailor responses.
- **Agent Performance:** Provide feedback to agents on their interaction style and effectiveness based on detected emotions.

Healthcare and Mental Health Monitoring:

Remote Monitoring: Track emotional states over time for patients dealing with stress, depression,

or anxiety.

- **Therapeutic Tools:** Support emotion-aware therapies by providing objective feedback on patient mood changes.

Human–Computer Interaction (HCI):

- **Virtual Assistants:** Enhance virtual assistants by allowing them to respond empathetically to the user's emotional state.
- **Interactive Learning:** Adapt educational content dynamically based on the learner's engagement and emotional cues.

Security and Surveillance:

- **Anomaly Detection:** Identify stress or agitation in environments like airports or public spaces to flag potential security issues.
- **Behavior Analysis:** Assist law enforcement by analyzing communication patterns in suspicious calls or recordings.

Entertainment and Media:

- **Audience Analysis:** Assess audience reactions during live events or while watching media to fine-tune content delivery.
- **Adaptive Gaming:** Create games that adjust their difficulty or narrative based on the player's emotional responses.

REFERENCES

[1] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. Datta Barua, M. Murugappan, Y. Chakole, and U. R. Acharya, ''Automated emotion recognition: Current trends and future perspectives,'' *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106646.

[2] M. B. Akçay and K. Oğuz, ''Speech emotion recognition: Emo tional models, databases, features, preprocessing methods, supporting modalities, and classifiers,'' *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[3] J. Lerner, Y. Li, P. Valdesole, and K. Kassam, ''Emotion and decision making,'' *Annu. Rev. Psychol.*, vol. 66, p. 145, Sep. 2014.

[4] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, ''A review on speech emotion recognition using deep learning and attention mechanism,'' *Electronics*, vol. 10, no. 10, p. 1163, May 2021.

[5] G. Gosztolya, ''Posterior-thresholding feature extraction for paralinguis tic speech classification,'' *Knowledge-Based Syst.*, vol. 186, Dec. 2019, Art. no. 104943.

[6] S. Renjith and K. G. Manju, ''Speech based emotion recognition in Tamil and Telugu using LPCC and Hurst parameters—A comparitive study using KNN and ANN classifiers,'' in *Proc. Int. Conf. Circuit,Power Comput. Technol. (ICCPCT)*, Apr. 2017, pp. 1–6.

[7] I. Baabbad, T. Althubiti, A. Alharbi, K. Alfarsi, and S. Rasheed, ''A short review of classification algorithms accuracy for data prediction in data mining applications,'' *J. Data Anal. Inf. Process.*, vol. 9, no. 3, 2021, Art. no. 162174.

[8] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, ''Excitation features of speech for emotion recogni tion using neutral speech as reference,'' *Circuits, Syst., Signal Process.*, vol. 39, no. 9, pp. 4459–4481, Sep. 2020.

[9] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, ''Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence,'' *Inf. Sci.*, vol. 563, pp. 309–325, Jul. 2021.

[10] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, ''Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN,'' *Sensors*, vol. 17, no. 7, p. 1694, Jul. 2017.

[11] C. Bakir and M. Yuzkat, ''Speech emotion classification and recognition with different methods for Turkish language,'' *Balkan J. Electr. Comput. Eng.*, vol. 6, no. 2, pp. 122–128, Apr. 2018.

[12] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, ''The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning,'' *Sensors*, vol. 22, no. 7, p. 2461, Mar. 2022.

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, ''MELD: A multimodal multi-party dataset for emotion recognition in conversations,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, Art. no. 527536.

[14] S. Latif, A. Qayyum, M. Usman, and J. Qadir, ''Cross-lingual speech emotion recognition: Urdu vs. western languages,'' in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, p. 8893.

[15] R. Li, J. Zhao, and Q. Jin, ''Speech emotion recognition via multi-level cross-modal distillation,'' in *Proc. INTERSPEECH*, 2021, pp. 606–610. [16] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, ''Deep learning in computer vision: A critical review of emerging techniques and application scenarios,'' *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100134. [17] S. Oh and D.-K. Kim, ''Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning,'' *Appl. Sci.*, vol. 12, no. 3, p. 1286, Jan. 2022.

[18] S. Wu, S. Zhong, and Y. Liu, ''Deep residual learning for image steganalysis,'' *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10437–10453, May 2018.

[19] S. Byun and S. Lee, ''A study on speech emotion recognition system with effective acoustic features using deep learning algorithms,'' *Appl. Sci.*, vol. 11, no. 4, p. 115, 2021.

[20] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, ''Deep learning techniques for speech emotion recognition, from databases to models,'' *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021.

[21] M. Imani and G. A. Montazer, ''A survey of emotion recognition methods with emphasis on E-learning environments,'' *J. Netw. Comput. Appl.*, vol. 147, Dec. 2019, Art. no. 102423.

[22] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, ''Speech emotion recognition using deep learning techniques: A review,'' *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[23] S. Akinpelu and S. Viriri, ''Robust feature selection-based speech emotion classification using deep transfer learning,'' *Appl. Sci.*, vol. 12, no. 16, p. 8265, Aug. 2022, doi: 10.3390/app12168265.

[24] H. Iman, R. Arabnia, and R. Branchinst, ''Pathways to artificial general intelli- gence: A brief overview of developments and ethical issues via artificial intelligence, machine learning, deep learning, and data science,'' in *Proc. 22nd Int. Conf. Artif. Intell.*, Las Vegas, NV, USA, 2017, p. 80180.

[25] M. Abdolahnejad and P. X. Liu, ''Deep learning for face image synthesis and semantic manipulations: A review and future perspectives,'' *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5847–5880, Dec. 2020.

[26] K. Asifullah, S. Anabia, Z. Umme, and Q. A. Saeed, ''A survey of therecent architectures of deep convolutional neural networks,'' *Artif. Intell. Rev.*, vol. 53, 2020, Art. no. 54555516.

[27] Q. Zhang, N. An, K. Wang, F. Ren, and L. Li, ''Speech emotion

recognition using combination of features,'' in *Proc. 4th Int. Conf. Intell. Control Inf. Process. (ICICIP)*, Jun. 2013, pp. 523–528.

[28] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, ''DE-CapsNet: A diverse enhanced capsule network with disperse dynamic routing,'' *Sensors*, vol. 19, no. 12, p. 115, 2020.

[29] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Edward, ''Capsule networks a survey,'' *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 1295–1310, Jul. 2019.

[30] S. Sabour, N. Frosst, and G. Hinton, ''Dynamic routing between capsules,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp.

[31] Y. Li, H. Sun, S. Feng, Q. Zhang, S. Han, and W. Du, ''Capsule-LPI: A LncRNA–protein interaction predicting tool based on a capsule network,'' *BMC Bioinf.*, vol. 22, no. 1, p. 119, Dec. 2021.

[32] M. Ouyang, R. Das, J. Yang, and H. Li, ''Capsule network based end-to end system for detection of replay attacks,'' in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, 2021, pp. 1–12.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[34] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H.-N. Lee, ''Two-way feature extraction for speech emotion recognition using deep learning,'' *Sensors*, vol. 22, no. 6, p. 2378, Mar. 2022, doi: 10.3390/s22062378.

[35] S. Wu, ''Expression recognition method using improved VGG16 network model in robot interaction,'' *J. Robot.*, vol. 2021, pp. 1–9, Dec. 2021. [36] O. J. Agbo-Ajala and S. Viriri, ''Deep learning approach for facial age classification: A survey of the state-of-the-art,'' *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 179–213, Jun. 2021.

[37] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, ''A survey on deep transfer learning,'' 2018, *arXiv:1808.01974*.

[38] J. de Lope and M. Graña, ''An ongoing review of speech emotion recognition,'' *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.

[39] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. Schneider, ''Speech emotion recognition using machine learning—A systematic review,'' *Intell. Syst. Appl.*, vol. 20, 2023, Art. no. 200266, doi: 10.1016/j.iswa.2023.200266.

[40] H. Zhang, Z. Li, H. Zhao, Z. Li, and Y. Zhang, ''Attentive octave convolutional capsule network for medical image classification,'' *Appl. Sci.*, vol. 12, no. 5, p. 2634, Mar. 2022.

[41] K. Simonyan and Zisserman, ''Very deep convolutional networks for large–scale image recognition,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[42] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, ''Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,'' *J. Big Data*, vol. 8, no. 1, Mar. 2021, doi: 10.1186/s40537-021-00444-8.