International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 421–451

Review

Accent Classification Using Machine Learning Techniques: A Review

Sarah Jassim * and Husam Ali Abdulmohsin

Department of Computer Sciences, College of Science, University of Baghdad, Baghdad 10071, Iraq; Husam.a@sc.uobaghdad.edu.iq
* Corresponding author: sarah.ahmed1301@sc.uobaghdad.edu.iq

Received date: 15 December 2024; Accepted date: 15 May 2025; Published online: 29 May 2025

Abstract: Accent is a person's distinct manner of speaking a particular language. It dramatically influences communication by producing pronunciation variations, which makes it challenging for automatic speech recognition (ASR) systems to understand spoken language accurately. The growing need for more accurate speech recognition technology means that improving machines' capability to classify and recognize accents becomes an essential challenge in speech processing. In response to this problem, this paper reviews previous studies on accent classification models. It discusses the principal methodologies used in this research, including datasets, pre-processing techniques, feature extraction, evaluation metrics and classification methods based on traditional machine learning (TML) and deep learning (DL) techniques utilized for accent recognition. The review includes journal articles and conference proceedings published between 2015 and 2025, emphasizing recent years. Relevant articles were sourced from leading academic databases and platforms, including Scopus, IEEE, Springer, MDPI, Google Scholar, and ResearchGate. The study concludes by identifying key research gaps and proposing future directions to advance accent recognition systems, offering valuable guidance for addressing current challenges and exploring innovative methodologies. A comparative analysis shows that the k-NN is the most effective traditional machine learning (TML) classifier. Among DL models, the pre-trained xResNet18 model outperforms other deep learning (DL) models when applied to well-structured English accent datasets while CNN achieves higher accuracy for datasets with diverse English accents but relatively small dataset sizes. Additionally, the fine-tuned transformer Wav2Vec2 achieves higher overall accuracy using a balanced and diverse dataset of six English accents, demonstrating strong performance in raw audio-based accent classification.

Keywords: accent classification; automatic speech recognition; deep learning; traditional machine learning

1. Introduction

An accent is the variation in how people pronounce words within the same language [1]. It is a speech pattern observed in a speaker's language. The most noticeable characteristic of non-native speech is accentedness, which results from a mismatch between the properties of the native and non-native languages [2]. An accent is defined in phonetic terms as a specific type of pronunciation influenced by the phonetic characteristics of the speaker's native language, which are carried over into their use of a second language. Accents are shaped by the impact of the first language (mother tongue) on the second language [3].

Generally, when people hear a language, they can identify the region and accent within it. But the challenge is how we give this ability to the machine. With the rapid growth of Automatic Speech Recognition (ASR) technologies, which translates speech signals into comparable textual interpretation, ASR has primarily been used in data analysis applications that handle multimedia (audio/video) contents, like speaker detection, and in applications that use voice for human-machine interfaces, like interactive voice response (IVR) systems and intelligent personal assistants, accessibility, customer service, language translation, and Voice-controlled digital assistants such as Amazon’s Alexa, Google Home, and

Apple's Siri. Home assistant, etc. [4–6]. Due to variations in pronunciation based on accent, accented speech presents issues for speech recognition systems, resulting in inconsistent recognition, this matter is especially pertinent for non-native speakers, whose speech may include uncommon phonemes or sounds inadequately represented in the system's training data, accents significantly degrade the performance of ASR systems, as models trained primarily on standard accents struggle to generalize to diverse speech patterns, leading to increased word error rates (WER) and reduced transcription accuracy [7]. To address these challenges, speech recognition systems require improvements to raise overall performance, diminish recognition errors, and enhance accuracy under varied and demanding settings. So, the challenges of speech recognition are the main motivating aspects for analyzing the question of accent classification.

Accent classification, in particular, has received much attention. It plays a critical role in enhancing ASR systems and is considered a preprocessing step to speech recognition. Accent classification can only help fine-tune speech recognition systems to better detect the accent of the speech by identifying the speaker's ethnicity and adjusting settings accordingly. Accent classification, which provides identification of a speaker's origins, is crucial in applications like crime investigation [8,9].

Accent recognition is a complex problem due to the different characteristics that set accents possess. Accents differ by voice quality, prosody, and phoneme pronunciation, such as rhythm, stress, linking, intonation, and spectral features, which capture speech frequency. Such features include the Mel-Frequency Cepstral Coefficient (MFCC), Spectrogram, Chromagram, Spectral Centroid, and Spectral Roll-off [10]. Generally, an accent recognition system includes two main components: feature extraction and modeling [11].

The feature extraction stage converts the representative features into output feature vectors from the signal vectors of input speech produced from the samples of accented speech [12]. The quality of the feature vectors obtained from the feature extraction process significantly influences the accuracy and performance of the recognition system [13]. Many researchers evolved models by utilizing machine learning (ML) to create an effective tool for accent classification. ML is a field that emphasizes the learning aspect of AI by creating algorithms that best represent a set of data. With subsets of data, it creates an algorithm that might employ unique or unusual feature and weight combinations that can't be obtained using fundamental principles. In machine learning, four popular learning techniques can be used for tasks: unsupervised, semi-supervised, supervised, and reinforcement learning [14].

Our paper is interested in supervised machine learning techniques. We use labeled data to train the learning algorithm. The learning mechanism is described as a supervised one. Since the output of the target is known, the learning algorithm tries to estimate this output iteratively. It is corrected to lower the variation gap between the actual output and the estimated [15,16].

Since the early years, many supervised machine-learning techniques have been applied to speech applications. This is regarded as Traditional machine learning (TML), such as hidden Markov models (HMMs), Gaussian Mixture models (GMM), Support Vector Machines (SVM), and K-nearest neighbors (k-NN) [17–21]. TML techniques were limited in their ability to process natural data in their raw form [22]. These methods are used to extract features from audio recordings and then use the features to classify the audio into different classes [23]. The extraction and representation of voice features highly influence the performance of any ML method [24]. as well as factors like the specific task and dataset [25].

In recent years, deep learning (DL) technologies have significantly improved speech-based applications. It is considered a novel machine learning form and eliminates the need for a feature extraction stage [22]. DL models can automatically extract features from large-scale speech data, improving accent recognition ability [26]. DL is the method of feeding a variety of Machine Learning algorithms with multi-layered models as inputs. Typically, neural networks comprise these models. Varying degrees of non-linear procedures. Machine Learning algorithms try to learn from these deep neural networks by extracting information and specific features [27]. Before 2006, it was difficult to predictably and directly search deep architecture inputs; nevertheless, the creation of DL algorithms helped to address this problem and streamlined the procedure for looking through the parameter space of profound structures. It has been shown that using deep architectures rather than shallower structures is more effective at representing non-linear functions [27,28].

Due to the lack of review research on accent classification, this paper presents a new outline of the state-of-the-art speech accent classification research provided in this article, including details on classification methods, feature extraction, pre-processing, and the datasets used.

This review aims to bridge the gap by thoroughly analyzing recent studies on the classification of speech accents from various languages based on Traditional machine learning (TML) and Deep learning (DL) methods in multiple applications.

The rest of this paper is organized as follows: Section 2 Survey Methodology. Section 3 Literature

Survey. Section 4 demonstrates the Accent Classification model. Section 5 evaluation metrics. Section 6 Brief Comparison of Accent Classification Model Studies. Section 7 research gaps and future directions. Section 8 conclusion.

2. Survey Methodology

The literature survey is systematically examines contemporary research in speech-based accent classification. The main objective is to discover high-performing of accent classification models. The review comprises journal papers and conference proceedings published from 2015 to 2025, focusing on recent years. Relevant papers were obtained from prominent academic databases and platforms, including Scopus, IEEE, Springer, MDPI, Google Scholar, and ResearchGate. The search concentrated on English-language papers that specifically addressed accent classification through machine learning and deep learning methodologies for accents from different languages. The search utilized the following keyword combinations:

"Accent classification," "accent recognition," "accent identification," "native language identification," "non-native accent identification," and "foreign accent identification." Studies were considered if they focused on the classification of accent speech, particularly accent recognition by machine learning or deep learning techniques. Each chosen study was evaluated according to the subsequent criteria:

- The dataset includes size, speaker count, sample quantity, accent variation, limitations, and strengths.
- Preprocessing techniques.
- Feature extraction methods.
- The modeling methodologies field encompasses traditional machine learning algorithms (TML) and deep learning (DL) architectures.
- Performance evaluation.
- Future directions.
- Results and conclusions

The obtained information was examined and structured to illustrate the principal trends in speech accent classification research.

3. Literature Survey

Various recent methods and approaches have been developed to address accent classification in many languages and applications. This section summarizes the most important studies that have been conducted using many mechanisms of machine learning (ML) which depend on Traditional machine learning (TML) and Deep learning (DL) as classification models.

An improvement in the accent classification system was presented in [29]. It used a GMM-UBM (Gaussian Mixture Model—Universal Background Model) framework trained on normalized Perceptual Linear Predictive (PLP) features. Also used Principal Component Analysis (PCA) and Heteroscedastic Linear Discriminant Analysis (HLDA) to model vowel-specific information, utilizing the knowledge that vowel pronunciation carries significant accent cues. This method did much better than the baseline of 42.3% and got 53.7% accuracy in a 7-way classification challenge on the Foreign Accented English (FAE) corpus with only 20-second speech samples. The FAE corpus comprises 23 accents, although only 7 were chosen to represent the broader accent classifications. The dataset is unbalanced, with accents such as Brazilian Portuguese (459 utterances) being overrepresented relative to Arabic (112 utterances). The disparity skews the model since accents with more data have a higher impact during UBM adaptation, potentially leading to overfitting and reduced performance for underrepresented accents. This problem is fixed by eliminating silence, normalizing features, and giving vowel models weights based on frequency and discriminativeness. However, the skew in the dataset makes it hard to generalize fairly and generalize.

Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) were proposed for accent recognition in [30]. The method utilizes speech samples from 5,132 speakers from 11 native language backgrounds from the INTERSPEECH 2016 Native Language Sub-Challenge dataset. The authors used voice activity detection (VAD) to remove the silence from the audio samples and segment the audio into 4-second durations. DNNs evaluate long-term prosodic features, and RNNs investigate short-term acoustic variations to capture both. Fusion improves classification accuracy by integrating both models' results. This hybrid method exceeds SVM-based baseline systems.

An accent recognition subsystem for the Moodle e-learning platform was proposed, as described in [31]. This research uses MFCCs for feature extraction and a neural network (NN) classifier for accent classification. They experimented on the Wildcat Corpus of Native and Foreign-Accented English and

found that this method is 14% more reliable than traditional methods for recognizing speech. Integrating such a system into e-learning can enhance language learning by providing automated pronunciation feedback and adapting content based on the learner's accent.

A weighted accent classification approach was proposed, as introduced in [32]. This study employs Extreme Learning Machines (ELMs) to classify North American accents into seven groups using the TIMIT dataset. With MFCCs, normalized energy parameters, and first and second derivative features, the ELM method achieved an accuracy of 77.88%, showing its effectiveness in accent classification within this dataset.

Deep learning and machine learning techniques have been widely applied to classify accents in English speech. One study focuses on native English speakers from the United Kingdom and non-native speakers from Mexico. Long vowel sounds from the speakers are recorded. Then, MFCCs extract features from speech data and evaluate multiple models, including Random Forest, LSTM, neural networks, and Hidden Markov Models (HMM). The best performance is achieved through an ensemble approach, combining Random Forest and LSTM models with an accuracy of classification of 94.74%, which significantly outperforms the traditional HMM by a 5% margin [6].

The integration of spectral and prosodic features has also been explored in accent classification. Research on Classical Arabic accents, particularly Quranic recitations by Malay speakers, found that combining spectral features (MFCC) with prosodic characteristics such as spectral tilt, energy, and pitch improved recognition performance. Using a custom dataset of Surah Al-Fatihah recitations recorded by 14 certified reciters, the study achieved accuracy rates ranging from 81.7% to 89.6% with a GMM classifier [13].

Another approach, investigated in [33], applies k-NN to classify English accents spoken across different countries. A dataset of 330 speech samples from the U.S., Spain, France, Germany, Italy, and England was analyzed using MFCC features, 165 of which feature American accents and 165 of which feature non-American accents. With this balanced dataset, the k-NN classifier was effectively distinguished between American and non-American English accents, achieving an accuracy of 87.2%. Similarly, this research [34] explored k-NN and MFCC features to examine native accents in English speech with a dataset of 1023 audio samples from a speech accent archive. The selected audio of each accent is not balanced, and some accents have a minimal number of audio, so the authors select the highest number of audio covering Arabic, English, Korean, French, Spanish, and Mandarin speakers; this research has shown that the more training samples, then the higher of performance obtained. The best performance was obtained with K=3, yielding an accuracy of 57%.

The effectiveness of the k-NN classifier was investigated with the UK Ireland English Dialect Speech Dataset (UIED) [35]. This dataset comprises speakers from six regions: Ireland, Midland, Scotland, Southern England, Northern England, and Wales; the dataset comprises 17,877 audio samples, and the accents samples are imbalanced gender. Several ML models, including k-NN, SVM, and Random Forest, were evaluated, with k-NN achieving the highest performance at 98.48% accuracy.

A method for recognizing Turkish accents using the formant frequencies (F1, F2, and F3) taken from vowels was suggested [36]. This framework distinguished accents with up to 90% classification accuracy using k-NN, SVM, and Ensemble approaches; k-NN proved the most effective algorithm. The dataset included recordings from 112 monolingual speakers in various Turkish locations.

The k-NN as a classifier with harmonic pitch estimates and (MFCCs) as feature extractors were utilized for classifying British English preschoolers into native and non-native accent categories [37]. In the UK, 670 audio recordings were provided from five non-native and six native preschool children in various settings. At 94.5%, the system's categorization accuracy was relatively high. Since the employed datasets are not balanced perfectly, F-measure is also used to assess the performance of the proposed model, which achieved 94.8% for native children and 94.1% for non-native children.

The development of an automated classification system for identifying regional accents was proposed [12], focusing on non-native English speakers from South India, particularly Telugu, Kannada, and Tamil. The study used speech data from native and English utterances to investigate how the native language influences second-language speech patterns. Key acoustic features such as MFCC were utilized alongside classifiers, including i-vectors, GMM, and GMM-UBM. The study achieved an accuracy of 93.9% using an i-vector-based classifier, surpassing earlier methods.

Accent classification of ten English accents was suggested [38]. After feature extraction, the audio samples selected from the Speech Accent Archive dataset use various audio feature extraction techniques, including MFCCs, ZCR, and roll-off. The samples of each accent are unbalanced, ranging from 18 to 80 samples; the total number of data is 581, which is relatively small.Machine learning algorithms like Decision Trees, SVM, and Logistic Regression, along with deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), were used to classify accents. Decision Trees exceeded other models with an accuracy of 97.36%, while deep learning models like LSTM

achieved 68.26. The results show that classifying small datasets with TML is better than DL models.

Classification of Bengali accents from different regions in Bangladesh was proposed [39]. The researchers extracted MFCCs from speech samples and combined them with other spectral features (root mean square error, Chroma feature, spectral bandwidth, spectral centroid, zero crossing rate, roll-off), then applied several algorithms such as linear regression, decision trees, random forests, gradient boosting, and neural networks to classify the accents. The study achieved a maximum accuracy of 86% using random forest models on 9303 data samples collected from eight regions.

The challenge of (ASR) for Classical Arabic (CA) with a focus on Quranic accents, particularly as recited by Malay speakers, was investigated in [40]. The study demonstrates how Malay speakers' prosodic features (spectral tilt, energy, pitch, duration) influence the pronunciation of Quranic accents. A dataset of recitations of Surah Al-Fatihah by Malay speakers using seven different Quranic accents was created and used to develop an ASR system. The system combines prosodic features and spectral features for accent classification. GMM-UBM was utilized, presenting an improvement in recognition accuracy (86.148%) compared to other models like K-NN and GMM-i-vector.

A classification method to classify regional accents in Mandarin speech combining i-vectors and bidirectional Long Short-Term Memory (bLSTM) networks to capture temporal and speaker-specific features was utilized [41]. In addition to bLSTM, the authors evaluate Deep Neural Network (DNN) models with and without i-vectors. Feature extraction involves (MFCCs), and Fundamental Frequency Variation (FFV) features. The STM with i-vectors achieves the highest frame-level accuracy of 26.09% and speaker-level accuracy of 34.1%, outperforming traditional models like (SVM), which achieved only 15% speaker-level accuracy. The DNN with i-vectors also performed well, matching the speaker-level accuracy of the bLSTM but with slightly lower frame-level performance. The paper demonstrates that these methods can significantly improve accent classification and reduce ASR errors by dynamically selecting accent-specific models.

A novel method of accent classification was suggested in [42] by utilizing the VFNet (Variable Filter Net) convolutional neural network architecture. The model improves accent classification by capturing a hierarchy of information from speech signals through various filter sizes. This allowed for the detection of particular accent qualities that are frequency-dependent. The Speech Accent Archive dataset, which includes English accents from 109 speakers, was used to train the system. At an accuracy of 70.33%, VFNet's classification accuracy outperformed earlier models by a considerable margin. It performed 90% accurately for natural English, 71% for Arabic, and 50% for Mandarin accents.

A deep learning-based method for classifying and converting English accents was proposed in [8]. This research utilized a combination of CNNs, Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs). Using MFCC, short-term features, and long-term features as feature input. The Speech Accent Archive dataset was utilized to distinguish between American, Spanish, and Indian accents. The number of samples of each accent is not large, so classifying these data with the DL model yielded a relatively low accuracy of 68.67%.

A Convolutional Recurrent Neural Network (CRNN) as a classifier and Mel-spectrograms as input were utilized in [43], with 39,000 audio samples selected from the Common Voice dataset withbalanced ratio for all the accents. The CRNN achieved a higher accuracy (83.21%) than CNN alone (78.48%).

The CNN architecture was used to classify accents directly from raw audio waveforms instead of depending on more conventional handcrafted features like (MFCCs) and Mel-spectrograms [44]. The CNN's early layers are made to mimic filter-bank processes, where accent-specific properties are dynamically learned during training. The Common Voice dataset, 8 English accents, is used to assess the method. By obtaining a 10.94% improvement in Unweighted Average Recall (UAR) over baseline models employing MFCCs.

A 2-layer (CNN) model was utilized in [10] to study different spectral components for accent identification tasks. The study contrasts five features from spoken audio samples of five different accents: Arabic, English, French, Mandarin, and Spanish. These features include MFCC, Spectrogram, Chromagram, Spectral Centroid, and Spectral Roll-off. The Speech Accent Archive provided the dataset. According to the study, MFCC outperformed other features and produced the most excellent classification accuracy of 88%. Interestingly, the Chromagram suggested that accents might have melodic features, and the Spectral Roll-off feature, though uncommon in speech processing, showed promising results. The authors concluded that MFCC is still the most helpful feature for accent recognition, particularly in shallow CNN models.

The effectiveness of usingphonetic idiosyncrasies of French language with deep learning models'in accent recognitionwas investigated in [45]. This paper compares a 2-layer (CNN) with classic machine learning models such as Support Vector Machines (SVM) for multi-class accent classification using the Speech Accent Archive dataset, which contains five accents (Arabic, English, German, French, and Hindi). At 70.65% accuracy, CNN fared better than SVM. The research highlights the difficulties in

enhancing model performance with minimal data and the significance of comprehending phonetic characteristics unique to a language, such as the pronunciation of French vowels and consonants, to develop more effective accent identification algorithms.

A new accent recognition model was proposed in this research [46]. The authors integrated the phonetic characteristics of fixed and trainable acoustic models. The model extracts language-related phonetic features using an auxiliary ASR task to avoid overfitting to speaker-specific traits and enhance generalization. Embedding from a fixed acoustic model trained on standard speech data (Librispeech) and a trainable model optimized on accented speech (AESRC dataset) are fused in this approach. The ASR task supports the primary accent recognition task, enhancing the model even more. A Transformer-based baseline model is 8.02% inferior to this method's relative improvement.

A machine-learning approach to classifying accents in English speech, specifically focusing on distinguishing between Indian and American accents, was present in this research [9]. The study applies supervised learning techniques, including NN, logistic regression, KNN, and SVM. They utilize Sequential MFCCs as the primary feature; 20 coefficients of MFCCs are extracted from consecutive frames of the audio signal and concatenated to retain the temporal structure of the speech. This sequential representation allows the model to capture dynamic patterns in the speech signal, which are crucial for accent classification.The VCTK dataset was utilized; after addressing class imbalance through oversampling and tuning the models, the research achieved a maximum accuracy of 95%, with neural networks outperforming other methods. The paper highlights the importance of accent recognition for improving speech recognition systems, particularly for voice-enabled assistants who interact with speakers of different accents.

The formant-based characteristics with neural networks can significantly enhance accent classification,as demonstrated in [47]. The authors classified American English and Mandarin accents.Vowels, Formant frequencies (F1, F2, and F3), Voice Onset Region (VOR), and MFCCs were among the retrieved features using the Speech Accent Archive dataset. The study used linear neural networks and a more sophisticated neural network with two hidden layers (NN2HL) for classification. According to the results, the F3-F2 formants produced the highest accuracy rate (83.33%), and the accuracy increased to 86.67% when feature vectors were concatenated.

The classification of American, British, and Indian English accents was investigated [48]. This study uses deep learning methods, including a pre-trained CNN (MobileNetV2) model. This paper performed transfer learning to modify the model for accent classification and used Mel-spectrograms as features to represent the audio data utilized from the AccentDB dataset. The number of samples for each accent is 222. With an excellent classification accuracy of more than 90% for each of the three accents, this study demonstrates how reliable deep learning is for accent recognition.

DL techniques, including CNNs, have demonstrated remarkable results in accent classification. Research utilizing the Speech Accent Archive indicated that several acoustic features—such as spectrograms, spectral roll-off, spectral centroid, and MFCCs, improve model performance [49].CNN trained the Speech Accent Archive to address the problem of accent recognition. The study investigates the effects of several features, such as spectrograms, spectral roll-off, spectral centroid, and other time-frequency features, including chromatograms and MFCC. The authors show that linear-scale amplitude Mel-spectrograms enhance classification performance significantly. The model got an accuracy of between 0.964 and 0.987 for nine European accents, including Germanic, Romance, and Slavic languages. This was the best result from the accent speech archive dataset, which kept pauses and used these attributes. However, the dataset exhibited class imbalance, as some accents had significantly fewer recordings than others. To address this, the study applied data augmentation techniques to expand underrepresented accents artificially. Additionally, the study limited the number of samples per class to 80 recordings for larger groups to prevent model bias toward overrepresented accents. These preprocessing techniques helped mitigate dataset imbalance and contributed to higher classification accuracy, demonstrating the effectiveness of augmentation in improving model robustness. The study also shows that accent recognition could be enhanced by combining MFCC and other factors, especially in real-life situations, using a wide range of data collected by the public.

DL models were utilized to develop a multi-accent identification system that recognizes eight distinct English accents [50]. The system combines transfer learning via the ResNet50, ResNet18, and xResNet18 models and 1D Convolutional Neural Networks (1D-CNN). Features such as MFCC and Mel-Spectrogram were utilized. The xResNet18 model achieved 100% accuracy, precision, recall, and F1-score on Mel-Spectrogram features with zero padding. This work also presents a novel dataset called IndicAccentDB, which is a comprehensive resource for multi-accent recognition, containing audio samples from Native American and Australian English accents as well as six Indian accents: Gujarati, Hindi, Kannada, Malayalam, Tamil, and Telugu. The dataset addresses the issue of speaker mismatch and gender imbalance that were present in previous datasets. The research suggests future work

integrating advanced models like Graph Neural Networks and Transformer-based models for improved performance.

Researchers have suggested hybrid deep learning architectures, such as the CNN-LSTM model proposed in [51], to improve classification performance. This paper proposes a novel framework for accent classification using intra-native accent-shared features. They focus on classifying native and non-native English accents and evaluating their similarity using a deep learning-based Native Accent Identification (NAI) model, leveraging shared features among speakers with the same native accent. To fix the class imbalance problem, the study uses the Speech Accent Archive, a collection of evenly spread speech samples from different linguistic backgrounds, to ensure that each accent classification task has an equal number of speakers. Their method improves accuracy by 3.7% to 7.5% over baseline models. The paper also introduces a new method for accent similarity evaluation between native and non-native speakers. Results indicate that Mandarin native speakers' accents are the closest to native English accents, while Hindi accents are the most distant. The findings have implications for improving accent recognition and similarity evaluation in speech processing systems.

Researchers have explored various DL approaches to improve accent classification performance. One such method involves converting audio signals into spectrogram images, allowing CNNs to process two-dimensional data effectively [52]. Transfer learning was employed using the pre-trained AlexNet model to enhance the performance of utilizing the UK Ireland English Dialect Speech Dataset (UIED), which is unbalanced regarding gender. The system achieved 93.38% accuracy in gender-dependent assessments and 92.92% in gender-independent assessments. This research shows that spectrogram-based CNNs, integrated with transfer learning, produce effective solutions for regional accent classification.

Recent research has explored hybrid approaches to enhance accent classification, particularly in Arabic. One study investigated a Deep Neural Network (DNN) and Hidden Markov Model (HMM) hybrid approach for acoustic modeling and feature extraction [53]. The system classifies a speaker's accent by analyzing the predominant phoneme versions of five main Arabic accents (Gulf, Iraqi, Egyptian, Levantine, and Maghrebi). The model's accuracy ranges from 79% to 86%, depending on the dataset and conditions. This paper selects 88.6 hours of annotated speech from 2,900 speakers from the GALE (phase 3) dataset, focusing only on speakers where annotators unanimously agreed on the accent, ensuring reliable training and evaluation for Arabic accent classification.

A novel approach to feature extraction, presenting spectrograms and introducing a novel descriptor called Grad-Transfer was proposed [54]. The new descriptor was extracted using the Gradient-weighted Class Activation Mapping (Grad-CAM). This study significantly improves classification performance by utilizing Grad-CAM heatmaps to emphasize accent-specific regions in the spectrogram while utilizing the VCTK dataset. To avoid overfitting due to the limited number of speakers per accent and the uneven gender distribution in the dataset, different speakers are distributed among subsets to focus the model on accents rather than speakers. The method prevents speaker bias and advances state-of-the-art accent recognition algorithms.

Combining an attention mechanism with a 1D CNN, a Bi-directional Gated Recurrent Unit (CNN-BiGRU) architecture for English speech signal classification was suggested [55]. The VoxForge dataset, which includes a variety of English accents, was used for experiments. MFCC and filter bank (FBank), two important acoustic properties, are used. The proposed model, 1D CNN-BiGRU-Attention, performed best, averaging an F1 score of 85.52%, showing the method's quality for real-world language acquisition and security uses.

A novel approach to accent classification was suggested [4]. The goal of the project is to enhance accent classification models' performance by replacing traditional Fourier-based features, such as MFCCs and Mel-Spectrograms, with the Hilbert Mel-Spectrogram, which is more effective in capturing non-linear and non-stationary characteristics of speech. This work used a 4-layer- CNN model and a subset of accents from the Speech Accent Archive. The results show that the Hilbert Mel-Spectrogram performs better than conventional Fourier-based features, with accent classification accuracy reaching up to 88%. The paper concludes that HHT-based features provide a promising alternative for accent recognition tasks. However, despite this promising result, the dataset's quality limits the findings, which contain noisy, crowd-sourced samples, highlighting the need for cleaner, larger datasets in future research.

Three new deep learning models—Multi-task Pyramid Split Attention-DenseNet (MPSA-DenseNet), Pyramid Split Attention-DenseNet (PSA-DenseNet), and Multi-task DenseNet (Multi-DenseNet) for English accent classification were proposed [26]. These models integrate multi-task learning and attention mechanisms with DenseNet to enhance model performance. Multi-task learning allows the models to handle multiple related tasks simultaneously, improving generalization and reducing overfitting. Meanwhile, the attention mechanism is used to emphasize the most relevant features of the speech data, improving the model's ability to distinguish between accents. This study uses a common voice dataset and selected English accents from native speakers (England, USA) and non-native speakers

(Germany, India, Hong Kong), the number of samples of the accents ranging from (4159 to 7209). MPSA-DenseNet surpasses other models, achieving state-of-the-art accuracy by combining DenseNet's feature reuse capabilities with attention mechanisms and multi-task learning, leading to improved accent classification across different regions.

A new deep learning algorithm has been developed to analyze audio speech signals and identify Arabic accents in English of Saudi Arabia, Tunisia, Jordan, and Iraq speakers, was proposed in [1]. The authors designed a new dataset (ArL2Eng dataset to recognize Arabic accents from English speech) and utilized Mel spectrograms as input and an LSTM neural network as a classifier.The system's 79% success rate in identifying accents demonstrated the variations in English pronunciation among Arabic-speaking areas. The authors demonstrate that the proposed system can't achieve more than 79% accuracy cause of some factors. The first is the relatively large number of fluent speakers from all countries tested, which hinders the learning process, and the second factor is the presence of female speakers in both classes for all experiments. It is well known that the voices of female speakers are more similar to each other than males', providing more similarity between classes, which led to less accuracy.

A system for recognizing Maghrebian accents (Algerian, Tunisian, and Moroccan)was suggested [56]. This work uses a 2D (CNN) with MFCC as input. Silence in the speech signals, which can confuse accent recognition, is removed using the SVM classifier trained to detect silence based on MFCC features. The model achieved a maximum accuracy of 99.33% for a hearing duration of 1.5 seconds with the CNN model. The dataset consists of 150 audio files from 30 speakers. This system demonstrates that silence elimination and proper model design significantly improve the recognition of accents, making it an effective method for automatic speech recognition tasks.

The effectiveness of using the transfer learning approach with speech accent recognition (SAR) for low-resource languages, mainly Vietnamese, was investigated in [57]. Using pre-trained models from related speech tasks like ASR, Speaker Verification (SV), and SAR in other languages. This research uses various transfer learning techniques. A Vietnamese telephone dataset with three regional accents—Northern, Central, and Southern—tests the approaches. With an improvement of accuracy of about 46.7% over previous models, the results show that transfer learning, especially from ASR tasks, considerably improves the performance of SAR models. The study emphasizes the value of utilizing data from different languages and speech tasks and indicates that transfer learning is a feasible strategy for low-resource SAR tasks.

This paper investigates reducing data requirements without sacrificing accent identification accuracy using smaller datasets with difficult vowels [2]. The two professionally annotated datasets that the authors used for their studies were one with complete words and the other with single vowels. Utilizing a pre-trained Residual Networks (ResNet) classification model, they evaluated the model against a baseline SVM classifier. The findings demonstrated that segmental vowel-based classification still obtained a significant accuracy (78%) despite word-based classification yielding better accuracy (90%). This study proposes that accent identification systems could be improved using targeted smaller datasets focused on difficult non-native phonemes, perhaps reducing the data needed for languages with limited resources.

A novel Foreign Accent Identification (FAID) technique was proposed in [58]. This paper utilizes an MKELM with a pairwise weighted strategy for multi-class classification. The audio samples are selected from the speech accent archive dataset. The model surpasses conventional techniques like SVM, ANN, LSTM, and ELM, attaining an accuracy of 84.72% by using MFCCs and prosodic features (such as pitch and energy) as input. Although the suggested model increases computing efficiency and enhances accent differentiation, it recognizes that accent-sensitive features and speaker variability pose problems that could compromise classification accuracy. In the future, larger datasets for improved generalizability and word- or phoneme-level segmentation will be explored to improve performance. This method advances (ASR) systems, which may be used in speaker verification, voice-based systems, and e-learning.

A research study on English accent identification using advanced audio-based transformer models was presented in [59]. The authors construct a diverse dataset from YouTube videos illustrating six English accents: Indian, Arabic, Chinese, British, American, and African. The dataset comprises speech of 18,000 audio samples (3,000 per accent) from 365 speakers in several real-life circumstances (e.g., interviews, vlogs). They use fine-tuned, pre-trained versions to assess five transformer models: Wav2Vec2, UniSpeech, SEW, HuBERT, and AST. Wav2Vec2 achieves superior overall performance, with an accuracy of up to 99.74%, particularly excelling in the identification of Indian and British English accents. The authors show these models work well when looking at raw audio to classify accents.

4. Accent Classification Model

Most studies adopt a model that is used to accent classification, and it consists of several stages (Illustrated in Figure 1) below:

Figure 1. Accent Classification Model.

4.1. Datasets

The choice of dataset plays a crucial role in determining the effectiveness and generalization of accent recognition systems. This section provides an overview of the datasets used in the literature, highlighting their characteristics, strengths, and limitations. The datasets are organized based on language. Table 1 presents a list of these datasets along with their key attributes. Additionally, several datasets have undergone updates and refinements, improving their quality and applicability. The characteristics of each dataset play a vital role in shaping the design of features and methodologies tailored to accent recognition research [44].

4.1.1. English Accent Datasets

These datasets focus on English speakers with different regional or non-native accents.

- Foreign Accented English (FAE) corpus is a speech dataset. The Linguistic Data Consortium published it in 2007 (LDC2007S08). It has 4,925 English sentences from people with 23 different accents. Each recording is around 20 seconds long and phonetically rich, making it suitable for accent classification, speaker identification, and phonetic analysis. It has an unbalanced accent distribution [60]. This dataset was utilized by [29].
- TIMIT dataset is a speech corpus for speech processing research, especially in accent classification, speech recognition, and phoneme recognition. It was created by the Linguistic Data Consortium (LDC) and includes recordings of 630 people from eight major English dialect regions in North America. Each speaker reads ten phonetically varied sentences. Time-aligned phonetic, word, and sentence transcriptions are included in the collection, which offers comprehensive linguistic annotations [61]. TIMIT, which is recorded in 16 kHz, 16-bit WAV format, is very useful for training models that need reliable, high-quality speech data. It is widely employed in accent classification tasks because of its dialectal richness, which allows models to differentiate between regional differences in North American English speech patterns.It was utilized by [32].
- The Wildcat Corpus of Native and Foreign Accented English includes scripted and spontaneous speech from 24 Native American English speakers and 52 non-native English speakers. A fundamental element of the corpus is the spontaneous speech recordings obtained through a dialogue-based elicitation technique designed to encourage natural conversations. Speaker information includes native language, gender, and proficiency level. Although the speech is spontaneous, the recordings occur in a regulated environment, which may contrast with real-world applications. The dataset, comprising 76 speakers, may be inadequate for deep learning applications but is appropriate for small-scale machine-learning research [62]. It was used by [31].
- INTERSPEECH 2016 Native Language Sub-Challenge dataset is part of a challenge to recognize and classify speech set up at the INTERSPEECH 2016 conference. It comprises speech samples from 5,132 speakers, each providing a single 45-second recording, totaling 5,132 samples. These speakers represent 11 different native languages. It was made to address the task of identifying the speaker's native language based on their accents. The original Language Sub-Challenge aims to determine from accented speech what Native Language Sub-this dataset has two challenges. First, all the speech samples were taken with Babel background noise using low-quality head-mounted microphones, which makes the recordings less clear. Second, many speakers did not speak perfect English, and their accents made it hard to understand what they were saying. Consequently, there were several pauses throughout the speech. Even though it has some problems, the collection is still a good way to learn how to classify accents and identify languages [63]. This dataset was utilized by [30].
- Common Voice is a large-scale, crowd-sourced corpus that facilitates speech recognition research. It is one of the largest speech-processing resources open to the public. It has 32,585 hours of recorded speech and 21,594 hours of validated speech in 131 languages. Because of its large size and diversity, it works well for deep learning tasks because it gives us enough data to build strong models to learn from different speakers and accents. Some of the demographic metadata in the collection includes age, sex, and accent. This makes it useful for studying how speakers differ and

how bias affects speech recognition systems. This dataset is used for various speaker classification tasks, including speech recognition, speaker recognition, and language identification. But, even though the collection has some good points, it also has some drawbacks. Notably, some demographic metadata is missing or not given consistently, which could lead to biases in the model's training. Also, since the dataset comes from the public, the quality of the recordings can be different, and some examples may have background noise or incorrect speaker labels. Despite these problems, Common Voice is still essential for making speech recognition systems that can understand multiple languages and accents [64]. It has been utilized in various studies, including [25,38,44].

- VoxForge dataset is a multilingual, open-source dataset that contains speech recordings made by individuals who speak different languages, including English, with different accents. It is intended to support developing and assessing voice recognition engines, automatic speaker recognition systems, and other speech analysis applications. The website allows users to sign up and submit voice recordings, making the dataset an invaluable tool for researching multilingual speech variants and English accents. Its diversity is its greatest asset, especially regarding accent adaptability in ASR systems. However, because recordings are user-submitted and gathered under various circumstances, the inconsistent audio quality is a significant drawback [65]. It was used by [55].
- AccentDB dataset is a parallel corpus designed for accent classification, containing 16,984 recordings from 23 speakers across nine English accents [66]. It is structured to include labeled speech pairs, making it highly valuable for training models that require accent-specific labeled data. However, a key limitation is its small size and limited number of speakers, which may reduce the generalization capability for deep learning models. Despite this, it has been successfully applied to MobileNetV2-based accent classification, achieving an accuracy of more than 90% [48].
- Speech Accent Archive is a crowd-sourced pool of speech recordings [67]. This dataset includes 3037 voice samples. Each is from a different individual reading the same sentence, and they all show a variety of speech samples from various languages.Speakers originate from 177 countries and speak 214 different native languages. Non-native and native English speakers read the same paragraph. This structured approach ensures speaker diversity, linguistic consistency, and phonetic balance, making it valuable for accent classification and speech research. Its efficacy is limited, though, for large-scale deep learning models, which might need to be augmented with more speech data for efficient training. This dataset was used by [4,8,10,34,38,42,45,47,49,51,58].
- Voice Cloning Toolkit (VCTK) is a speech dataset containing data from 110 English speakers with different accents. Each speaker reads aloud about 400 sentences selected from a newspaper, the Rainbow Passage, and an elicitation piece for the Speech Accent Archive [68]. VCTK is best for speaker verification and text-to-speech models due to high-quality recordings from different speakers. Despite its strengths, the dataset has limitations for accent classification. Due to the relatively small number of speakers per accent, deep learning models trained on VCTK may learn speaker-specific features instead of accent-related features, potentially leading to overfitting. Also, this dataset is imbalanced in terms of gender distribution. These challenges make VCTK a challenging dataset for robust accent classification. This dataset was used by [9,54].
- UK Ireland English Dialect Speech Dataset (UIED) is an open-source licensed dataset. It contains high-quality audio recordings from volunteers who speak different regional accents of British English, including Southern England, Northern England, Scotland, the Midlands, Ireland, and Wales. The dataset comprises 17,877 recordings from 120 speakers (49 females and 71 males). The recordings cover a range of utterances with varying transcription lengths, from 9 to 169 characters, and the dataset includes over 31 hours of audio. The corpus was developed for speech technologies and language analysis [69]. The unique features of this dataset make it an excellent option. High phoneme coverage, a high sampling rate for intelligibility, and the selection of scripts that improve accent and idiolect elicitation are all necessary for analysis. It is particularly valuable for dialect classification and accent-aware speech recognition. However, its limited speaker number may not fully represent all regional dialect variations, and some accents may be underrepresented, leading to potential classification bias. This dataset was used by [35,52].
- Accented English Speech Recognition Challenge (AESRC) dataset consists of speech data from eight accents: Chinese, Indian, Japanese, Korean, American, British, Portuguese, and Russian. Each accent has approximately 20 hours of recorded speech from about 60 speakers, provided in Microsoft WAV format at 16 kHz, 16-bit mono. The dataset was collected in controlled, quiet environments using mobile devices, ensuring high-quality audio and particularly useful for accented ASR model adaptation and evaluation [70]. Its diverse accent representation allows researchers to analyze accented speech recognition performance across English-speaking populations. However, a key limitation is the limited speaker diversity (~60 speakers per accent), which may not fully capture intra-accent variability and could lead to overfitting in deep learning models. It was used

by [46].

- Librispeechdataset consists of approximately 1,000 hours of 16 kHz of read English speech but does not include accent labels [71]. Because of its quality transcriptions, constant audio recordings, and large size, it is beneficial for training ASR models. Its shortcomings, like the absence of accent labels, make it less appropriate for accent classification. Despite these drawbacks, LibriSpeech is often used alongside accented datasets like AESRC to enhance model robustness [46].
- Speech Recognition Dataset - England and Mexico involves recordings of seven phonetic sounds spoken by native British English speakers and fluent Mexican English speakers for accent classification and phonetic analysis. Four accent classes ensure fair machine learning application classification. It is beneficial for studying pronunciation differences and accent classification due to its apparent phonetic structure and balanced design. However, it has a small speaker number and restricted speech content (just phonetic sounds, not whole sentences). The dataset has been processed into a static dataset of statistical descriptions using MFCC features, with a sample window length of 0.02 seconds. It is publicly available on Kaggle for phonetic analysis and accent recognition research. It was used by [6].
- ArL2Eng dataset to recognize Arabic accents from English speech **c**onsists of 391 recordings of English speech from speakers in Jordan, Iraq, Saudi Arabia, and Tunisia, each lasting 21 to 52 seconds. The recordings capture the reading of a standardized English paragraph and have been converted to MP3 format with a 44,100 Hz sample rate [72]. This dataset was specifically created to address the lack of Arabic-accented English speech corpora and is used to train a deep learning model for classifying Arabic accents. This dataset has fluent speakers and is not gender balanced, both of which reduced accuracy [1].
- English accent dataset consists of 330 English speech sound samples, 165 of which feature American accents and 165 of which feature non-American accents, including accents from Spain, France, Germany, Italy, and England. The samples were collected from 22 speakers, equally divided between male and female participants (11 males and 11 females). It was used by [33].
- A dataset of six English accents was collected from YouTube videos covering Indian, Arabic, Chinese, British, American, and African accents, which was designed by [59]. It consists of 18,000 audio samples (3,000 per accent) from 365 speakers (175 female, 200 male), reflecting diverse and natural speaking contexts such as interviews, vlogs, and podcasts. All recordings were standardized to 16 kHz WAV format and trimmed or padded to 2.5 seconds. The data underwent preprocessing for noise removal and voice isolation to ensure quality. This dataset was created specifically for the research.

4.1.2. European Accent Datasets

- Two datasets of French-English vowels were collected by [2]. The first dataset consists of 2738 recordings of full words from the French language containing the vowels /u/ and /y/, spoken by 39 native English speakers and 20 native French speakers. All components were recorded in a soundproof booth in a controlled acoustic setting at 44.1 kHz sampling rate and 16-bit mono. Each audio file is then included with annotation text. The second dataset (Dataset II) was derived from the first and consisted of isolated recordings of the vowels /u/ and /y/ extracted from the whole words. The two datasets are helpful for phonetic analysis research. It is less appropriate for general accent classification since it concentrates exclusively on vowel sounds.
- Turkish accent/dialect dataset recorded by 112 monolingual university students from various Turkish areas. It comprises 103 syllables containing vowels and short words. The controlled environment of the recordings guaranteed excellent phonetic data. However, its use in speech recognition systems is limited by its emphasis on discrete syllables rather than continuous speech. It was used by [36].

4.1.3. Arabic Accent Datasets

- GALE (phase 3) Arabic Broadcast News and Conversations dataset: Contains speech recordings from five major Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and Maghrebian. This dataset comprises two primary parts: the first includes approximately 132 hours of Arabic broadcast news speech collected from 13 Arabic channels, and the second includes approximately129 hours of Arabic broadcast conversation speech (BC) from 17 channels, providing a diverse, real-world speech sample and including phonetic transcriptions to capture dialectal features. However, its accessibility is limited as it is not publicly available. There are 2552 speakers in this dataset, but just 2900 speakers were given dialectal labels [73]. This dataset was used by [53].
- Maghrebian dataset was collected from YouTube videos featuring 30 speakers (15 male, 15 female)

from Algeria, Tunisia, and Morocco, all speaking in Modern Standard Arabic (MSA). Each speaker contributed five audio files, resulting in 150 audio samples. These recordings varied from 30 seconds to five minutes and were sourced from different situational conversations involving journalists, politicians, and public figures. The total dataset duration was 1 hour and 34 minutes for Algerians, 1 hour and 20 minutes for Tunisians, and 1 hour and 26 minutes for Moroccan speakers.It is relatively small in size, and audio quality may vary. It has been applied in dialect classification models. It was used by [56].

4.1.4. South Asian Accent Datasets

- The Bengali accent dataset was collected manually from different people, including some Google forms and YouTube videos. It consists of 9303 voice samples. These samples were collected from speakers across eight regions of Bangladesh: Khulna, Dhaka, Rajshahi, Barisal, Chittagong, Sylhet, Noakhali, and Mymensingh. The dataset includes male and female voices, with the speakers aged 20 to 50 years. The duration of each recording is between (4 to 7) seconds. All audio files were normalized to a sample rate of 16,000 kHz to ensure consistency in the dataset. This dataset was used by [39].
- IndicAccentDB is a gender-balanced and labeled accent dataset with voice recordings in six non-native English accents (Gujarati, Hindi, Kannada, Telugu, Tamil, and Malayalam). Six non-native accents were gathered from volunteers with prominent non-native English accents who had at least a good deal of experience speaking one language spoken in India. Each speaker was requested to recite the Harvard phrases. The Harvard Sentences dataset has 72 sets containing ten phonetically balanced sentences. This dataset is relatively small, which could influence the generalization of deep learning models. This dataset was used by [50].
- South Indian language speech corpus comprises samples collected from native speakers of Tamil, Kannada, and Telugu. These languages, which represent a significant portion of the South Indian population, form the foundation of the dataset. None of the chosen speakers are well-versed in English, even though they are educated. The training set had a total of 5 hours of native speech data, with 20 speakers from each linguistic group, while the test set included 25 speakers, with speech samples of around 60 seconds each for evaluating non-native English accents. This dataset provides useful linguistic variation but has few speakers, which may impact classification accuracy. It was used by [12].

4.1.5. Asian Accent Datasets

- A Dataset of Mandarin speech consists of 135,000 utterances of Mandarin speech collected from 466 speakers across 15 geographical regions in China, representing various regional accents. All speakers are native to their respective dialect regions but speak Standard Mandarin as a second language. The recordings were made in cars during scripted in-car human-machine interactions, with equal distribution across male and female speakers. The dataset was manually transcribed and balanced across accents, ensuring the speech differences are acoustic-phonetic rather than linguistic, making it suitable for accent classification.This dataset was used by [41].
- A dataset of Vietnamese speech was collected from the telephone call. It consists of speech recordings provided by Viettel, the largest telecommunication provider in Southeast. The dataset comprises about 26 h of 8 kHz speech data uttered by 5009 people and is categorized into three major accents: Northern, Central, and Southern. It is particularly valuable for low-resource languages like Vietnamese. It is beneficial for spontaneous speech analysis, as the data comes from real-world phone conversations. However, background noise and recording variability may affect performance. it was used by [57].

4.1.6. Other Specialized Datasets

- Children's Speech Dataset involves 670 samples of speech collected from children in the UK, recorded in various environments, including those with background noise such as robot fan sounds, children's voices, and door closing. The participants include six native children (three males and three females) and five non-native children (three males and two females). This dataset was created to support the classification of native and non-native British English accents in preschool children. The dataset's ability to identify children's nativity from text-dependent and text-independent speech utterances led to their adoption in the suggested framework. It is suitable for commercial use in speech recognition systems for children. However, it has a small speaker pool, limiting its generalization to larger ASR models. It was utilized by [37].

- Surah Al-Fatihah Recitation Dataset was recorded from 14 certified Malay reciters who mastered seven Quranic accents: Khalad, Hafs, Bazzi, Khallaf, Ruwais, and Qunbul. Each reciter provided two Surah Al-Fatihah recitations in all seven accents, resulting in 5,684 words. The dataset is gender-balanced and recorded in a controlled environment, ensuring high-quality samples. However, it only applies to Quranic recitation and does not extend to general speech recognition. It was used by [40].
- The Quranic accents dataset consists of Quranic recitations from 14 certified reciters (Huffaz), with seven male and seven female speakers, all of whom are Malay speakers trained in Quranic accents (Qiraat). The recitations were recorded in a controlled environment, focusing on Sūrah Al-Fātiḥah, the first chapter of the Quran, which is mandatory in Muslim prayers. Each speaker recited the chapter twice, using seven different accents of the Quran (Qiraat). The audio recordings were segmented into sentences and words, transformed into a .wav format, and then downsampled to a 16 kHz sampling rate for analysis. It has only relevance for Quranic recitation and does not address general speech recognition.It was used by [13].

Table 1. Datasets utilized in the literature review.

Reference	Dataset	Type	Size/sample(s)	kHz	Speakers
[60]	FAE	Private	4925 s	8 kHz	-
[61]	TIMIT	Private	6,300 s	16 kHz	630
[62]	Wildcat	Public	-	22.05 kHz	76
[63]	INTERSPEECH 16 Native Language Sub-Challenge	Private	5132 s	-	5132
[64]	Common Voice	Public	32,585 h	22.05 kHz	-
[65]	VoxForge	Public	-	16 kHz	-
[66]	AccentDB	Public	19 h and 49 m	-	23
[6]	Speech Recognition Dataset - England and Mexico	Public	-	-	-
[67]	Speech accent archive	Public	3037 s	44.1 kHz	3037
[68]	VCTK	Public	-	48 kHz	110
[69]	UIED	Public	31 h	48 kHz	120
[70]	AESRC	Public	20 h	16 kHz	60
[71]	Librispeech	Public	1,000 h	16 kHz	-
[72]	ArL2Eng	Public	391 s	44,1kHz	391
[59]	dataset of six English accents	private	18,000 s	16 kHz	365
[2]	Two datasets of French-English vowels	Private	-	44.1 kHz	59
[39]	Bengali accentdataset	Private	-	16 kHz	418
[36]	Turkish accent/dialect dataset	Private	-	-	112
[37]	Children's Speech Dataset	Private	670 s	-	11
[40]	Surah Al-FatihahRecitation Dataset	Private	-	-	14
[41]	Dataset of Mandarin speech	Private	84.7 h	-	466
[50]	IndicAccentDB	Private	8210 s	-	19
[57]	dataset of Vietnamese speech	Private	26 h	8 kHz	-
[33]	English accentdataset	Private	-	-	22
[13]	Quranic accents dataset	Private		16 kHz	14
[12]	South Indian languages speech corpus	Private	5 h – for training	-	135
[73]	GALE (phase 3)	Private	261 h	-	2664
[56]	Maghrebian dataset	Private	1 h and 34 m for Algerians, 1 h and 20 m for	-	30

	Tunisians, and 1 h and 26 m for Moroccan

4.2. Pre-Processing

Pre-processing is a crucial step in speech processing, where data with flaws left in causes misleading and perhaps inaccurate processing results. Furthermore, the datasets obtained from diverse sources, including TV shows, YouTube videos, and broadcast recordings, frequently contain noise and require refinement to guarantee better audio quality by minimizing background noise and ensuring consistency before inserting it into the model to achieve the desired or expected results. Some of the main steps:

- Standardizing file formats: Audio files in different formats, such as wav, mp3, and mp4, are transformed into a standard format that works with the tools that were utilized [1,26,34,38,51,58].
- Speech segmentation: Speech is a nonlinear and non-stationary signal [54]. It is a dynamically fluctuating signal whose temporal properties alter with time. It is challenging to extract features from non-stationary signals. So, segmenting the speech signal into a sequence of frames is necessary before features are extracted. A sliding window is used to eliminate the speech signal's discontinuities [12,13,37].
- Silence removal: was employed to remove silence or pause from audio samples to enhance classification's performance [8,57,29,59]. Silence was removed from the speech signal using a pre-trained SVM classifier to get a pure speech signal and identify the silence segments, as described in[56]. However, Pauses can be a powerful sign of a foreign accent [29], showing that speech pauses positively impact the ability to recognize accents.
- Voice activity detection (VAD): This step is considered a fundamental pre-processing step in speech-processing systems. It is necessary to distinguish between signals with speech activity and those without. (VAD) contributes to the time savings needed to process speech data and increases the accuracy of the ultimate system by concentrating on the voiced portion of the speech [12,36,74,75].
- Data augmentation: This technique can prevent overfitting and enhance DL algorithms' ability for generalization. Multiple data augmentation methods existed for audio classification. These strategies include conventional techniques on raw audio signals and the augmentation of linear interpolation and nonlinear mixing on the spectrum [76]. There are many approaches to data augmentation, including data augmentation by SpecAug in [46], Dataset augmented to(word, 3 words, paragraph) in [10], and data augmentation also utilized in [49].

4.3. Feature Extraction

Following the completion of the preprocessing stage, the feature extraction procedure begins. It is critical in accent classification models and can be used with machine learning techniques to improve their accuracy and robustness. This subsection demonstrates the most common speech features used in the literature. Table 2 shows the Overview of speech features utilized in the literature of accent classification.

According to the literature MFCC is the most frequently utilized features. It offer a compact representation of an audio signal's spectral features. They capture essential frequency information while eliminating irrelevant details. The benefit of MFCCs is that they restrict unwanted spectral variations in the higher frequency bands. MFCC was used alone without additional features by [6,8,9,10,26,31,33,34,35,39].

The MFCC feature is relatively simple and does not need a deep model. It provides effective performance when compared with various spectral features, such as a Spectrogram, Chromagram, Spectral Roll-off, and Spectral Centroid [10].

MFCC and Filter Bank (FBank) features are designed to mimic how the human ear processes audio closely. In [44], the authors showed that learning directly from the raw waveform improved classification performance by 10.94% compared to traditional hand-engineered features like FBank and MFCCs. Similarly. In [55], the author demonstrated that FBank outperformed MFCC, with the 70FBank feature achieving the highest F1 value.

Several studies integrate MFCCs with other features instead of commonly used MFCCs to increase performance. In [37], harmonic pitch, which estimates how many components of frequency could be represented in the power spectrum, was applied to extract the essential signals of speech of children alongside MFCCs to distinguish between non-native and native children speaking British English. The study demonstrates that even in loud environments, the classifier with harmonic pitch, and MFCCs can

distinguish between non-native and native accents with the highest discriminating performance of 94.5% against noisy surroundings, producing good precision and recall.

In [40], prosodic features including (energy, pitch, spectral tilt, duration) were combined with MFCCs for accent classification. These features were extracted from entire waveforms or longer speech segments to capture speech variations influenced by dialect and accent. The model achieved an accuracy of 86.148%. Similarly, in [13], combining MFCCs with prosodic features resulted in a 5.5% to 7.3% improvement in accuracy compared to using MFCCs alone.

Also, in [41] the author combined fundamental frequency variation (FFV) features with MFCCs, where FFV is added to capture accent tonal fluctuations. In [38], Spectral Roll-off (SR) and Zero Crossing Rate (ZCR) were combined with MFCCs for accent classification. ZCR helps distinguish voiced from unvoiced signals by tracking how the signal crosses the zero axis, while SR measures spectral asymmetry, which is useful for identifying vocalized sounds. This feature combination achieved an accuracy of 98.05%, demonstrating its effectiveness for accent classification.

In [30], long-term prosodic features with 6373 features processed using DNNs, such as fundamental frequency (F0) statistics (range, max, min), sub-band energies and peaks, and mean, standard deviation, and kurtosis of MFCC, These features capture global speech patterns over an extended duration. The openSMILE toolkit was utilized to extract these features. The RNN also processed short-term features, including 39th-order Mel-scale filterbank features with logarithmic compression. These features were extracted from 25 ms frames with a 10 ms overlap. Using both sets of features together improves performance, proving that it is much more accurate at classifying accents than using just one set of features. This shows that combining global and local speech characteristics makes classification.

Another combination of MFCCs and other features includes Linear Predictive Cepstral Coefficients (LPCCs) and Perceptual Linear Predictive Coefficients (PLPs) used in [12]. The authors utilized spectral features, including LPCCs, to capture additional speech prosody-related information that distinguishes between languages, with the advantage of computationally being less expensive than MFCCs, which perform better when combined with other features. PLPs, based on psychoacoustic principles, are resistant to noise and perform better than MFCCs in hybrid environments. This paper used this combination to provide a comprehensive representation of features. Perceptual Linear Predictive Coefficients (PLPs) are also used in [29].

Mel-spectrograms are also the most common feature used for accent recognition. The Mel-Spectrogram visually represents time-varying energy in speech across different frequencies by converting the frequencies to the Mel scale. Mel bands in the spectrogram are evenly spaced, effectively simulating how the human ear perceives sound, which makes it easier to analyze speech patterns that vary by accent [77]. It was utilized by [1,2,43,48,50]. In [50], the Mel-Spectrogram outperformed MFCCs, achieving 100% accuracy.

Formants (particularly formants 2 and 3) indicate the physical movement of one's vocal cords while pronouncing different words. According to the book "AI 2003: Advances in Artificial Intelligence" [78], the frequency trajectory of formants 2 and 3, as well as the difference between them, can best indicate the feature of different accents. In [47], These features, combined with the Voice Onset Region (VOR), yield improved classification results. VOR is a crucial temporal feature, as different accents exhibit spectral differences in the Voice Onset Region, typically associated with the consonant part of a word, [79]; by concatenating the feature vectors, including formants and VOR, with classifier, the accuracy can be enhanced, achieving a rate of 86.67%.

Formant frequencies (F1, F2, and F3) features utilized by [36]. These formants are extracted from vowels. The formants are particularly useful for distinguishing accents because they capture acoustic variations caused by different regional pronunciations. These formant features are processed using Linear Predictive Coding (LPC) and used in statistical analysis and machine learning classification models, such as SVM and k-NN, to classify accents.

Phonetic features, including vowels and a small set of consonants, are used in [53] by training a speech recognizer to detect accent-specific variations in these phonemes. This enables the classification of accents based on how different speakers pronounce them.

Recently, some research has proposed new approaches to feature extraction. In [49], the primary feature that achieved high accuracy for accent classification is the amplitude Mel-Spectrograms on a linear scale, in contrast to the logarithmic scale commonly used in other studies. Using amplitude Mel-Spectrograms on a linear scale allows the model to establish broader boundaries between classes, proving more effective for accent classification and enabling the accuracy of state-of-the-art that ranges from 0.964 to 0.987. In [54] The author propose a novel approach to feature extraction by presenting spectrograms and introducing a novel descriptor called Grad-Transfer, extracted using the Gradient-weighted Class Activation Mapping (Grad-CAM). The study significantly improves classification performance by utilizing Grad-CAM heatmaps to emphasize accent-specific regions in the spectrogram,

advancing the state-of-the-art in accent recognition systems. In [4], a novel input feature is proposed , the Hilbert Mel-Spectrogram, a logarithmic frequency representation of the signal derived using the Hilbert-Huang Transform (HHT). By utilizing HHT, this feature overcomes the traditional limitations of Fourier analysis, such as limited frequency resolution, spectral energy leakage, and harmonic artifacts. While preliminary findings indicate that this feature might perform better than its Fourier-based features, such as MFCCs and Mel-Spectrograms, a more thorough investigation is needed to verify this efficiency. In [51], the key feature introduced is intra-native accent shared features. These features are derived by combining voice data from multiple speakers of the same native language rather than extracting features from individual speakers. The technique focuses on generating a single spectrogram from mixed voices of intra-native speakers. The intra-native shared feature extraction method improves the performance of accent classification tasks, especially in non-native and native English accent recognition.

Table 2. Overview of speech features by usage and performance in accent classification.

Feature	Feature Usage	Key Findings/Performance	used by
MFCC	Standalone	Widely used; compact spectral representation; robust performance	[6,8,9,10,26,31,33, 34,35,39]
MFCC + FBank	Combined	Raw waveform outperforms hand-crafted features; FBank outperformed the MFCC.	[44,55]
MFCC + LPCC + PLP	Combined	PLP robust to noise; LPCC computationally efficient	[12]
PLP	Standalone	Utilized with GMM-UBM, achieved an accuracy of 53.7%	[29]
MFCC (statistics) + Long-term Prosodic Features	Combined	Combining local and global features improves the classification	[30]
Mel-Spectrogram	Standalone	Simulates human hearing; [50] reported 100% accuracy	[1,2,43,48,50]
Amplitude Mel-Spectrogram (linear scale)	Modified Feature	Higher inter-class separability, accuracy up to 0.987	[49]
Grad-CAM-enhanced Spectrogram	Novel Descriptor	Attention-based accent-focused feature improved classification	[54]
Hilbert Mel-Spectrogram (HHT-based)	Novel Feature	Overcomes FFT limitations; promising alternative	[4]
Intra-native Accent Shared Spectrogram	Shared Feature	Improves performance for native vs. non-native classification	[51]
Spectral Rolloff + ZCR	Combined with MFCC	ZCR distinguishes voiced/unvoiced; SR detects vocalized energy	[38]
Harmonic Pitch	Combined with MFCC	Harmonic Pitch estimates how many components of frequency could be represented in the power spectrum, this combination provide 94.5% accuracy, even in noisy environments	[37]
Formants (F1,F2, and F3)	Standalone	These formants are extracted from vowels. The formants are particularly useful for distinguishing accents because they capture acoustic variations caused by different regional pronunciations.	[36]
Formants (F2,F3) +VOR	Combined	(F2,F3) Captures articulation differences across accents,	[47]

		VOR associated with the consonant part of a word	
FFV	Combined with MFCC	FFV captures tonal differences significant for accent cues	[41]
Phonemes (vowels/consonants)	Standalone	Accent-specific pronunciation used for classification	[53]

4.4. Classification Methods

4.4.1. Traditional Machine Learning (TML) methods

TML models have demonstrated considerable success in the task of accent classification. k-Nearest Neighbor (k-NN) is a machine learning classification method that employs a supervised algorithm. It determines the class of a new object by calculating the distance between the object and its nearest neighbors in the training data. k-NN uses the neighborhood classification approach, where the closest neighbors are utilized to predict the class of a new test sample [34]. k-NN is the most TML classifier used in the last year. It was used to highlight the challenges of different pronunciation styles across regions and indicate the possibility for further improvement with additional features and larger datasets [33]. The authors find that adding training data can enhance the accuracy of the k-NN model and outline future uses for better ASR systems [34]. The k-NN algorithm was utilized to classify speakers from six regions of the UK by using 17,877 audio samples. k-NN yielded higher performance than other TML algorithms in our literature review, with an accuracy of 98.48% [35]. In [36] k-NN classifier outperforms the other models. .The effectiveness of using a k-NN classifier with noisy data is demonstrated in [37].

The Gaussian Mixture Model (GMM) is another TML model used for accent classification. (GMM) is a parametric probabilistic model. It assumes that every data point originates from a mixture of a finite number of Gaussian distributions. The model comprises a weighted sum of Gaussian components as these distributions fully describe it [80]. In [13], it has been highlighted the importance of combining MFCC alongside prosodic features for better recognition of Quranic recitations of Surah Al-Fatihah with GMM, particularly for accents influenced by the speakers' native dialects. Due to the capacity of the GMM model to represent a broad range of sample distributions, it is frequently used to enable the ASR systems of the classical accent of Quranic accents. However, because singularities exist in GMM, overfitting occurs when the complexity of the model is high. To address this issue, a modification was introduced by using the Universal Background Model (UBM) as described in [40]. This makes training the models much simpler and more efficient, subsequently enabling a method of scoring quickly during testing.This paper demonstrates that GMM-UBM is an effective method for accent recognition in this task. A comparable GMM-UBM structure is utilized in [29]. However, the imbalance in the dataset introduces skew, making it difficult for the model to perform fairly and generalize well across all accent categories.

An i-vector-based classifier is frequently used for accent classification. It considers speaker and channel variability and generates GMM super vectors compactly [81]. It outperforms other models like GMM and GMM-UBM and highlights the importance of accent and regional language identification in multilingual environments utilizing an i-vector-based classifier

Decision Trees (DT) are another machine learning classifier. In DT-based classifiers, the nodes represent the point where an attribute is chosen. The values of attributes serve as the foundation for creating child nodes. This tree represents a competitive advantage in a chosen feature's value. The leaves indicate the actual output or class label. It suppressed other models like LSTM, as described in [38].

The Random Forests model is a TML classifier that integrates decision trees to enhance the outcomes of a specific task [62,63]. It was used to classify the Bengali language spoken in Bangladesh and outperformed other models [39]. This work improves the accuracy of previous research on Bengali language classification by using this model with more data.

Extreme Learning Machine (ELM) is a learning technique for single-layer feed-forward neural networks. It can also be used to train multilayer perceptrons using hierarchical frameworks. ELMs can be used for both regression and multiclass classification problems. ELMs aim to minimize training errors and the norm of output weights. It does not require any adjustments to the input weights of neurons.It was used to classify North American accents into seven groups, as described in [32]. This paper utilizes MFCCs, the normalized energy parameter, and their first and second derivatives as raw features for training ELMs and SVMs. Both ELMs and SVMs converge to a single global optimum solution. ELMs optimize the sum of squared errors, while SVMs construct a hyper-plane that maximizes the separation between the data classes [82–86]. ELMs provide a higher accuracy of 77.88%.

A new Foreign Accent Identification (FAID) approach was proposed in [58]. This approach utilizes a Multi-Kernel Extreme Learning Machine (MKELM) and a pairwise weighted scheme to address the

problem of multiple classifications. (MKELM) model is a hybrid model that integrates the benefits of the Bayesian "sum of kernels" model and the Extreme Learning Machine (ELM) [87]. Multi-kernel functions are advantageous because they combine several features with kernel functions to improve mapping performance. Different kernel functions produce different effects on the performance of the constructed MKELM model. The proposed approach suppressed the state-of-the-art models, including SVM, ANN, LSTM, ELM, MLELM, and KELM-tested models regarding computational complexity and accuracy.

4.4.2. Deep Learning (DL) methods

Many DL models were utilized with accent classification. A neural network (NN) is an algorithm of machine learning inspired by the biological nervous systems[88]. NN is used for classification and machine learning [89]. Every single NN contains nodes (analogous to cell bodies), and that node communicates with other nodes through connections (analogous to axons and dendrites). Connections between nodes in an NN are weighted based on their ability to provide a desired outcome [27,90]. Multiple studies employed the NN method for accent classification. A comparison of accent recognition models showed that using MFCC along with the NN algorithm was better than other methods, increasing recognition accuracy by 14%. The LPCC with NN and the GMM with MFCC were evaluated; nevertheless, MFCC-NN yielded the most successful results [31]. Furthermore, the NN algorithm was utilized to classify American and Mandarin accents using 50 speakers for each accent.Two NNs, one with linear classification (LNN) and the other with nonlinear classification and two hidden layers (NN2HL). NN2HL was provided the best accuracy with 86.67% [47]. Another study utilized the NN algorithm for binary classification to classify American and Indian accents. NN compared to other TML approaches like logistic regression, k-NN, and SVM, from results, the NN yielded good performance, but the main drawback of employing NN is their long computation times when used with considerable data input where logistic regression provides results closer to it but with less time [9].

A CNN is a deep feed-forward neural network that utilizes convolutional computations and a layered structure [48]. It is highly effective in feature extraction and performs exceptionally well with image and audio signal inputs [91,92]. CNN consists of various layers, such as the input, convolutional, pooling, and fully connected layers [93]. CNNs are commonly used for speech recognition and audio classification. The input for CNNs consists of audio and speech signals; however, the raw one-dimensional (1D) signals are typically not used for CNN-based schemes. 1D audio/speech signals are converted as a preprocessing step switch from a 1D to a 2D signal. The two-dimensional representation of the audio signal is then fed into a CNN model. This 1D to 2D conversion is often carried out to create spectrograms or other time-frequency representations, which serve as input features for CNN-based models [94–96]. CNN has many architectures, including AlexNet, ResNet, DenseNet, MobileNetV2, and others. Numerous studies utilize CNN classifiers for accent classification. One study utilized the VFNet architecture. This work demonstrates how varying filter widths can enhance accent identification across various accents and outperform other models like AlexNet, ResNet, and a combination of RNN-DNN models [42].

Some studies used speech accent archive datasets with CNN models, and [44] Used CNN architecture that learns to classify accents directly from raw audio waveforms. The study also emphasizes the importance of appropriate initialization of CNN filters to accelerate training and improve performance. In [10], the authors concluded that MFCC is still the most helpful feature for accent recognition, particularly in shallow CNN models. In [45], CNN with minimal data is used to focus on the significance of comprehending phonetic characteristics unique to a language, such as the pronunciation of French vowels and consonants, to develop more effective accent identification algorithms. In [49], it is shown that linear scale amplitude Mel-spectrograms enhance classification performance significantly when used with CNN other than other features like MFCC and Mel spectrogram. In [4], 4-layer- CNN with accent classification accuracy reached up to 88% utilizing Hilbert Mel-Spectrogram rather than conventional Fourier-based. However, despite this promising result, the dataset's quality limits the findings, which contain noisy, crowd-sourced samples, highlighting the need for cleaner, larger datasets in future research.

The author in [56] utilized (CNN) to classify Maghrebian accents using an SVM classifier for eliminating silence. The study evaluates multiple CNN architectures and shows that eliminating silence improves classification accuracy by approximately 2%, with the most miniature CNN models.

In recent years, many studies have applied Pre-trained CNN Models for accent classification utilizing a Transfer learning technique, a branch of machine learning that involves reusing knowledge from solving one problem to address a similar problem. In essence, it leverages previously trained models to tackle new tasks by transferring learned information to improve performance on the new challenge. Because the performance of machine learning algorithms is negatively impacted by the heterogeneity of the data, the pre-trained AlexNet model used in [52], this research shows that spectrogram-based CNNs, integrated with transfer learning, which overcome this problem, produce effective solutions for regional

accent classification. In [48] utilized pre-trained CNN (MobileNetV2). In [50] xResNet18 model was used. In [2], the authors proposed a cross-domain transfer learning method to identify phoneme-based accents using the architectural strengths of a pre-trained Residual Networks (ResNet) model. This framework explores how segmental accents, specific phoneme deviations rather than global accents, and the overall perception of an accent can be used to identify accents in ASR systems through focused datasets.

Long-Short-Term Memory (LSTM) is another DL method that can handle sequences of time-related data [6]. LSTM is often successfully experimented with in terms of accent recognition. In [1], LSTM was utilized to classify the English accents of speakers from four Arabic countries. In [41], i-vectors and bidirectional Long Short-Term Memory (bLSTM) networks are used to classify Mandarin accents, achieving higher accuracy than DNN with i-vectors and SVM.

Also, the hybrid CNN-LSTM approach was used in [51] to enhance the classification performance. The fusion of the CNN-LSTM model is proficient in reducing the overfitting issues. LSTM is replaced on the fully connected layer of the CNN model.

In [43] Fusion of CNN and RNN Convolutional Recurrent Neural Networks (CRNN) combines the strengths of CNNs and RNNs for foreign accent classification. The CNN component was used to capture spatial features, while the RNN component, specifically a Gated Recurrent Unit (GRU), was utilized to handle temporal dependencies in audio signals. The GRU mechanism addressed the vanishing gradient problem, making the model more computationally efficient than LSTMs.

In [30], a fusion of DNNs and RNNs was suggested to identify 11 native languages from accented speech. DNNs use long-term features, and RNNs use short-term acoustic. The proposed model provides an accuracy of 52.48%, UAR 52.5%, and improved from the SVM baseline (44.66%).

In [55], another combination was used by combining an attention mechanism with a 1D CNN, a Bi-directional Gated Recurrent Unit (CNN-BiGRU) architecture for English speech signal classification. A two-layer 1D CNN is used to extract local features, where the first layer's output is passed to the second layer for deeper, more abstract feature learning. The BiGRU is utilized to capture global features, and an attention module is added to modify feature weights, improving the model's focus on essential features.

In [8], the accent classification system of three English accents was enhanced by combining CNNs, DNNs, and RNNs. MFCC is fed as is feature input into the CNN, short-term features are fed into the RNN, and long-term features are fed into the DNN.

Multi-task learning has appeared as a powerful technique for accent recognition, particularly when integrated with advanced deep-learning architectures. Through exploiting multi-task learning, models can simultaneously learn several related tasks, such as accent classification and ASR, which reduces overfitting to speaker-specific features and improves generalization as introduced in [46], Using a CNN-based acoustic frontend (Jasper) and a Transformer-based aggregation module together in a hybrid model has that using an auxiliary ASR task can significantly improve accent recognition by using robust phonetic features. In [26] the authors applied multi-task learning alongside DenseNet and attention mechanisms, resulting in an improved accent classification.

Transformer-based models have demonstrated significant efficacy in accent classification. The research in [59] highlights the real-world application of Wav2Vec2, illustrating its importance in applications like voice assistants, speech analytics, and multilingual automated speech recognition systems due to its ability to derive comprehensive, self-supervised representations from raw audio.

4.4.3. Hybrid of Deep Learning (DL) and Traditional Machine Learning (TML)

Some research combines DL and TML methods for accent classification to utilize both methods' strengths. The research work in [6] highlights the potential applications of accent classification for improving speech recognition systems, especially for non-native speakers using an ensemble approach of Random Forest and LSTM models through a vote of average probabilities. In [53], a hybrid approach combining Deep Neural Networks and Hidden Markov Models (HMMs). The system first utilizes HMM-GMM models for phonetic transcription alignment, followed by DNNs to provide emission probabilities, enhancing the model's performance. They present a novel approach to identifying Arabic accents using a speech recognizer trained to distinguish between phonetic variations of key phonemes, specifically vowels and a small set of consonants.

5. Performance Evaluation Metrics for Accent classification

System performance metrics are essential in the comprehensive evaluation of classification. Consequently, various performance measures were employed for classifier evaluation. The confusion matrix is the most widely utilized statistic for assessing classifier performance in numerous applications. It is used to create performance evaluation measures that compare the estimations of the predicted

attribute with the actual values. Table 3 displays the frequencies of the evaluation metrics in the evaluated papers.

True positive (TP) refers to cases where the actual data value is positive and the predicted value is positive. True negative (TN) refers to cases where the actual data value is negative and the predicted value is negative. False positive (FP) refers to cases where the actual data value is negative. The predicted value is positive, and false negative (FN) refers to cases where the actual data value is positive and the predicted value is negative [97].

Several evaluation metrics were employed to assess the classifier's performance. Accuracy, recall, precision, and F-measure were the predominant metrics employed for accent recognition. The performance metrics are explained below.

5.1. Accuracy

This metric is primarily used to assess performance, reflecting the number of cases accurately classified by a specific method. It is calculated by the number of correct predictions to the total number of predictions. The formula for classification accuracy is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Relying solely on classification accuracy as a measure of success can be misleading, especially in imbalanced datasets. Consequently, additional success criteria must be assessed to evaluate the proposed algorithm's performance[98].

5.2. Recall

Recall is the ratio of positive cases in which both the predicted and actual labels are positive or correct, that is, True positive to the total number of positive cases. It is also known as the true positive rate. Recall can be calculated as follows [97]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

5.3. Precision

Precision is the metric that measures the ratio of correctly classified positive cases to the number of positive cases by the system. Precision can be calculated as follows [98]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

5.4. F-Measure

The F-measure is the weighted harmonic mean of recall and precision when FP and FN are in perfect balance. The typical F-measure is F1. The formula for the F-measure is presented below [98]:

$$\text{F-measure} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+ \text{Recall}} \tag{4}$$

5.5. Equal Error Rate

Equal Error Rate (EER) is used to calculate the average of the false acceptance rate (FAR) and false rejection rate (FRR).A lower EER indicates better performance. The EER is determined as follows:

$$\text{EER} = \frac{\text{FAR}+\text{FRR}}{2} \tag{5}$$

5.6. Receiver Operating Characteristic

The ROC curve illustrates the FAR as a function of the FRR for different values. The ROC curve is a graphic representation of the performance of the classification algorithm.

Table 3. Frequency of evaluation metrics for accent classification. ×: metric not used; ✓: metric used.

Reference	Accuracy	Precision	Recall	F-measure	EER	ROC
[33],	✓	✓	✓	✓	✓	×
[34],[36],[38],[47],[42] [44],[10],[4],[57],[1], [41],[6],[53],[54],[32], [31],[29]	✓	×	×	×	×	×
[37],[39],[58],[49],[48] [50],[52],[51],[43],[8] [56]	✓	✓	✓	✓	×	×
[99]	✓	×	×	×	✓	✓
[12]	✓	×	✓	×	✓	✓
[13],[40]	✓	×	×	×	✓	×
[45],[26]	✓	×	×	✓	×	×
[2],[9],[59]	✓	✓	✓	✓	×	✓
[55]	×	✓	✓	✓	×	×
[46]	✓	×	×	×	×	✓
[30]	✓	×	✓	×	×	×

6. Brief Comparison of Accent Classification Model Studies

This section provides an overview and brief comparison of the studies reviewed in this research, focusing on datasets, accuracy, feature extraction methods, classification techniques, and model performance. The comparisons are summarized in Table 4, Table 5, and Table 6, which correspond to Traditional Machine Learning (TML), Deep Learning (DL), and hybrid TML-DL models, respectively. Fields with the highest performance values are highlighted in bold. Table 4 presents TML-based studies, indicating that k-Nearest Neighbor (k-NN) is frequently used and often performs well, particularly with small to medium-sized datasets. MFCC features are commonly used alongside TML models. For instance, integrating k-NN with MFCC features to classify six UK dialects achieved the highest TML performance with 98.48% accuracy. Table 5 covers DL-based approaches, showing that models like CNNs, xResNet18 with transfer learning, and Transformer-based architectures such as Wav2Vec2 achieve high accuracy. For example, training a CNN on the Speech Accent Archive dataset using amplitude Mel-spectrograms (on a linear scale) achieved 96.4% to 98.7% accuracy. The xResNet18 model, when well-structured, achieved 100% accuracy, while Wav2Vec2 using raw audio input reached 99.74% accuracy on large datasets. Table 6 highlights hybrid approaches combining TML and DL techniques. Models like CNN-LSTM and DNN-HMM hybrids demonstrate robust performance across various datasets and languages.

Table 4. Comparison of accent classification using traditional machine learning (TML) studies.

Ref	Dataset	Year	Accent	Feature	Classification method	Pre-processing	Performance
[36]	Turkish accent/dialect dataset	2022	Turkish Accent/Dialect Classification of three regions	Formant Frequencies (F1, F2, F3)	k-NN	VAD	90%
[37]	Children's Speech Dataset	2022	(Native and Non-Native Children) with British English	Harmonic Pitch and MFCCs	k-NN	Speech Segmentation	94.8%
[40]	Surah Al-Fatihah Recitation Dataset	2023	Quranic Accents Recognition	Prosodic features (duration, energy, pitch,spectral-tilt), (MFCC)	GMM-UBM	Audio downsampled to 16 kHz	86.148%
[58]	Speech Accent Archive	2024	six different accents: English, Arabic, Chinese, Korean, French, and Spanish	MFCCs, Prosodic Features (Pitch, Energy)	MKELM	-Standardizing file formats - Down sampled audio to 16 kHz	84.72%

[38]	Speech Accent Archive	2022	10 different accents: (Dutch,Arabic, English, Hindi, French, Portuguese, Mandarin, Korean, Spanish, and Russian,)	-MFCC -ZCR -Spectral Roll-off	Decision Tree	- converted to mono standardizing file formats -feature scaling - PCA for feature reduction - Outlier Detection & Removal	98.05%
[33]	English accentdatas et	2020	six English accents: (Spain, France, Germany, Italy, England and America)	MFCC	K-NN	____	87.3%
[13]	Quranic accents dataset	2019	Quranic Accents Recognition	-MFCC - Prosodic features (energy, pitch, spectral tilt)	GMM	-speech segmentation -downsampled from 44.1kHz to 16kHz	81.7% - 89.6%,
[12]	South Indian languages speech corpus	2022	Native language identification using an English accent	MFCC, LPCC , PLPs	i-vector based classifier	-speech segmentation -voice activity detection (VAD)	93.9%
[34]	Speech Accent Archive	2020	6 accents (Arabic, English, French, Korean, Mandarin, Spanish)	MFCC	k-NN	Standardizing file formats	57%
[35]	UIED	2020	UK speakers- 6 accents (Ireland, Midland, Northern England, Scotland, Southern England, Wales)	MFCC	k-NN	___	98.48%
[39]	Bengali accent dataset	2021	Classifying Bengali accents from different regions of Bangladesh	MFCCs	Random forest	___	86%
[32]	TIMIT	2018	Seven North American accents	MFCCs, normalized energy parameter	ELM	___	77.88%
[31]	Wildcat Corpus	2017	Classifying Native- and Foreign-Accented English	MFCC	NN	___	91,43%
[29]	FAE	2015	7 non-native English accents	PLP	GMM-UBM	-Silence removal - feature normalizatio n	53.7%

Table 5. Comparison of accent classification using deep learning (DL) studies.

Ref	Dataset	Year	Accents	Feature	Classification method	Pre-processing	Performance
[59]	dataset of six English accents	2024	Six English accents (Indian, Arabic, Chinese, British, American, African)	Raw audio	Wav2Vec2	-Noise removal - silence trimming -resampling (16kHz) -segmentation to 2.5s	99.74% (Indian accent)
[26]	Common Voice datasets	2025	English accents from five regions: (England, the United States, Germany, Hong Kong, and India)	MFCCs	MPSA-denseNet models	-Standardizing file formats - Standardizing voice files	92.5%

[1]	ArL2Eng	2024	Arabic accents from English spoken speech, accents from (Jordan, Iraq, Saudi Arabia, and Tunisia)	Mel-spectrograms	LSTM	-Standardizing file formats -Sampling rate Format conversion	79%
[55]	VoxForge dataset	2023	Classification of English accents	f Bank	1D CNN-BiGRU-Attention model	_____	85.52%
[2]	Twodatasets of French-English vowels	2024	Automatic identification of the accent of French language spoken by non-native English speakers.	Mel-spectrograms	Transfer learning - ResNet	-convert files to 2Dmel-spectrogram -Z-normalization and value scaling.	90.5 For words 78.3 for vowels
[48]	AccentDB	2022	English Accent :(American, British, and Indian)	Mel Spectrogram	MobileNetV2	-convert files to mel-spectrogram	95%
[47]	The speech accent archive	2021	2 English accents: American English (AE) Mandarin-accented English	VOR and Vowel Formants (f3-f2)	NN	___	86.67%
[9]	VCTK	2021	2 English accents (Indian and American English)	MFCC	NN	-Oversampling underrepresented data (Indian accent)	95%
[42]	Speech Accent Archive	2019	Three English accents: Native English, Arabic-accented English, and Mandarin-accented English.	Spectrogram	VFNet	-Downsampling the audio to 16 kHz	70.33
[8]	The Speech Accent Archive	2020	three English accents: (American, Spanish, and Indian)	MFCC	Fusion of DNN. RNN, and CNN	-pause removal	68.67%.
[54]	VCTK	2023	five accents: British, American, Scottish, Irish, and Canadian	Grad-CAM-generated descriptors	DenseNet	Spectrogram generation	macro average accuracy (MAA) improvement is 23.00% (with Grad-Transfer)
[43]	Common Voice Dataset	2020	five English accents: American (US), British (England), Indian, Australian, and Canadian	Mel-spectrograms	CRNN	-convert all samples to 3.62 second	83.21%
[41]	Dataset of Mandarin speech	2019	regional accents in Mandarin speech	MFCCs, 7 (FFV) features Amplitude	bLSTMwith i-vectormodel	____	68.4%
[49]	Speech Accent Archive	2022	nine English accents from nine European languages	Mel-Spectrogram on a linear scale	CNN	-z-normalization (using z-score) -data augmentation	96.4% to 98.7%
[56]	Maghrebian dataset	2024	Maghrebian accent(Algerian, Tunisian, Moroccan)	MFCC	CNN	-Speech normalized between -1 and 1 - silence removal via SVM	99.33%
[52]	UIED	2023	Regional British English Accent	Spectrogram	CNN	-convert files to spectrogram	accuracy of 93.38% and 92.92for gender-independent

							and gender-dependent
[50]	IndicAccentDB	2022	Six Indian English accents + two native English accents (American and Australian).	Mel-spectrogram	Transfer Learning: xResNet18	-Zero-padding of audio files -convert files to spectrogram	100%
[57]	dataset of Vietnamese speech	2024	Vietnamese Accent Recognition	Log-Mel Spectrogram	Conformer-based model, Transfer Learningfrom ASR, and Speaker Verification tasks	Data normalization - silence removal	88.3%
[44]	Common Voice dataset	2020	eight English accents :(US, Australian, Canadian, English, Indian, New Zealand, Scottish, South Atlantic)	Raw Waveform	CNN	_	81.09%
[10]	Speech Accent Archive	2020	5 accents : (English, Spanish, French, Arabic, Mandarin)	MFCC	CNN	Data augmentation (word, 3 words, paragraph)	70.38%
[46]	AESRC+ Librispeech	2021	English accents	FBank Features + Phonetic	Hybrid model combining CNN (Jasper) and Transformer with Multitask Learning (MTL)	Data Augmentation using SpecAug	8.02% improvement over Transformer baseline
[4]	Speech Accent Archive	2023	non-native English speakers	Hilbert Mel-Spectrogram	CNN	___	88%
[45]	Speech Accent Archive	2021	five accents: Arabic, English, German, French, and Hindi	Spectrograms	CNNs	Convert audio to spectrograms	70.65%
[51]	Speech Accent Archive	2022	native and non-native English accents	Intra-native shared accent features extracted using spectrograms	CNN-LSTM	-Standardizing file formats - noise reduction	Accuracy improved by 3.7%
[30]	INTERSPEECH 2016 Native Language Sub-Challenge dataset	2016	11-native language from accented speech using	Long-term prosodic and short-termfeatures	Fusion of (DNN-RNN)	-VAD -segmenting speech into 4-second samples	52.48%

Table 6. Comparison of hybrid traditional machine learning (TML) and deep learning studies.

Ref	Dataset	Year	accents	Feature	Classification method	Pre-processing	Performance
[6]	Speech Recognition Dataset - England and Mexico	2019	4 accents (UK: West Midlands, London; MX: Mexico City, Chihuahua)	MFCC	Random Forest, LSTM	-	94.74%
[53]	GALE (phase 3)	2023	5 Arabic accents (Gulf, Iraqi, Egyptian,	Phonetic Transcriptions, Short/Long	HMM, DNN	-specific phonetic transcription	79 to 86%

	Levantine, and Maghrebian)	Vowels, Consonants

7. Research Gaps and Future Directions

Accent classification remains a compelling challenge in artificial intelligence (AI) , playing a key role in enhancing speech technologies such as automatic speech recognition (ASR) systems of various applications such as voice assistants, transcription services, language learning tools, and customer service automation. Despite significant progress, the field faces persistent challenges that demand ongoing research and development.

One of the primary issues is the lack of diverse and balanced datasets, particularly for underrepresented languages and regional accents. While numerous English-language datasets are publicly available, many contain limited samples for specific accent groups and often lack speaker diversity. In some cases, gender imbalance within these datasets further affects the fairness and generalizability of models. Addressing these gaps requires expanding existing datasets to include more speakers from different regions, ensuring better representation of less common accents, and achieving a balanced distribution across genders. Beyond English, it is also critical to develop datasets for underrepresented languages such as Arabic, Vietnamese, and Turkish to support more robust cross-lingual accent classification.

While many studies evaluate accent recognition models in controlled environments, recent research has begun incorporating real-world datasets featuring background noise, speaker variability, and recording inconsistencies. Datasets like VoxForge, YouTube Accents, the Vietnamese Telephone Corpus, and the Maghrebian dataset provide valuable insights into real-world challenges. Similarly, with their crowd-sourced, diverse, and variable-quality recordings, Common Voice and the Speech Accent Archive are well-suited for practical applications such as ASR, voice assistants, and language learning tools. However, many benchmark studies still rely on clean, ideal data, limiting generalization. Future research should emphasize domain adaptation, data augmentation, and transfer learning while prioritizing datasets that reflect real-world conditions and demographic diversity to ensure robust and deployable accent recognition systems.

Another key direction is the development of advanced deep learning (DL) architectures. Future models should integrate approaches like hybrid neural networks, attention mechanisms, and transformer-based systems to capture complex speech data patterns better. Pre-trained is especially valuable in low-resource environments and can be combined with traditional machine learning classifiers to improve performance across different languages and accents. Additionally, exploring segmentation techniques at the phoneme or word level rather than the sentence level can further enhance classification precision.

The development of AI-driven accent classification systems must also incorporate ethical considerations, especially given their close relationship to speech, identity, and sociocultural context. Critical concerns include bias and fairness, where models trained on unbalanced or non-representative datasets may disproportionately underperform on certain accents, thereby reinforcing social or regional inequalities. Moreover, using publicly available voice data—such as those from YouTube or TV broadcasts—raises questions regarding informed consent and speaker privacy, particularly when data is repurposed without explicit permission. To ensure responsible development, future research should prioritize using ethically sourced open datasets, apply fairness-aware training strategies, and employ transparent model evaluation across all demographic groups.

In this context, explainable AI (XAI) techniques such as Grad-Transfer are gaining traction. These methods allow researchers to identify the specific time-frequency regions of spectrograms that influence model predictions, enhancing interpretability and helping to uncover hidden biases in model behavior. Such techniques are essential for building trustworthy and transparent systems, particularly when accent classification is applied in socially sensitive domains.

Further areas of exploration include handling multilingual and code-switched speech, cross-domain generalization, and optimizing models for deployment in low-resource environments. There is also a growing need for lightweight architectures capable of maintaining high performance while meeting the efficiency requirements of mobile and embedded devices. Table 7 demonstrates the future directions for accent classification identified in recent studies from 2022 to 2025. These directions highlight the importance of dataset enhancement, model innovation, fairness, explainability, and practical deployment, collectively paving the way toward more inclusive, accurate, and ethically responsible accent recognition systems.

Table 7. Future directions for accent classification in reviewed studies.

References	Year	Proposal for Future Directions
[12]	2022	A better understanding of the acoustic-phonetic properties of accents may also lead to the synthesis of a regional accented speech, in the future. Toward these studies, possible in future, this paper can be expected to be helpful, in developing a better understanding of the NLI task, especially for languages spoken in the southern part of a diverse country, India.
[48]	2022	Much research is required to gather and simulate our own dataset for the purpose of detection of lies in border control as well as the use for court proceedings, healthcare, border security and domestic lies. Another direction of research will involve the use of NIST for speaker recognition evaluation.
[49]	2022	Further studies may be helpful to expand the number of recognition classes, using an intermediate classifier to determine the L1 language group of the speaker before classifying a particular accent and using a dataset with a variety of spoken content.
[50]	2022	Few limitations in the proposed work have paved the path for future work. As speech data has a time-dependent recurrence relationship, advanced recurrence-based models like Graph Neural Networks and Transformer-based models can be used for accent classification. We can augment the proposed dataset with more non-native speakers from various backgrounds to make the MARS more robust. Speaker-based accent analysis can also be used to improve the system's reliability and generate more efficient results in real time.
[51]	2022	Extending this work using many non-native English accents speakers' countries and a huge dataset size is under consideration for future work with some technical improvements to the current methodology
[4]	2023	Finally, future work in this area could focus on testing the computational performance of Hilbert-based features for accent Recognition and other ASR applications, comparing the tradeoff of accuracy vs computational cost for various feature extraction techniques.
[54]	2023	The proposed method can be applied to other audio processing tasks since the premise of Grad-Transfer is to take advantage of the regions of the spectrograms (defined in terms of time and frequency) derived from the localization maps that are important to predict an audio category.
[40]	2023	All the trials, evaluations, and analyses conducted in this research have led us to different ideas for future work and improvement, especially in developing a database for various Quranic accents.
[57]	2024	In the subsequent studies, we will undertake more investigations, not only on Vietnamese speech, and lower the model size to accommodate low-resource devices. In addition, we will investigate the effectiveness of transfer learning approaches for speech accent recognition using pre-trained self-supervised models.
[56]	2024	For the future works, if the accent recognition will be performed by another kind of deep learning such as LSTM and compare the results with the actual best CNN model. It is preferable to consider more Maghrebian accents (Libyan and Mauritanian) for the next works.
[2]	2024	Further research could include classifying more than two accent groups, focusing on different sounds or including different languages.
[1]	2024	Our future approach will divide the dataset into two, one for males and one for females, then train the deep learning method on each to produce two models, which are then merged to improve recognition result In addition to collecting more data to increase the size of the dataset for better deep learning. Indeed, we are collecting data from two new countries, Egypt, and Morocco. The actual results will be compared to those of these two countries in future work.
[59]	2024	Future research should explore further enhancements in model training and dataset expansion to ensure even greater accuracy and applicability in real-world scenarios.
[58]	2024	In future research, we plan to address several important issues. Firstly, we aim to explore segmenting at the word or phoneme level instead of the sentence level to enhance classification accuracy. Additionally, we will evaluate the proposed model on larger datasets to ensure its generalizability. Moreover, we

		will investigate the effectiveness of using multi-resolution features, which combine long and short-term features, and consider incorporating information on formant position shifts. These endeavors will contribute to further advancing the field of automatic accents and dialect identification
[26]	2025	While acknowledging the significant memory resources required for training the DenseNet architecture combined with multi-task learning and attention mechanism, we propose exploring lighter-weight network architectures, such as CondenseNetV2.

8. Conclusion

This paper reviews various approaches to accent classification in multiple languages and applications. A new study on accent classification can be conducted using journal articles and conference papers published between 2015 and 2023 across academic databases and platforms, including Scopus, IEEE, Springer, MDPI, Google Scholar, and ResearchGate. This study examined various preprocessing approaches, feature extraction methods, evaluation metrics, and accent classification models, including traditional machine learning (TML) and deep learning (DL) frameworks.Furthermore, used datasets for accent classification were reviewed, highlighting their characteristics, strengths, and limitations. The regional tendency of accent classification may be seen due to the information acquired from publications. Additionally, this information can be utilized to explore new approaches and conduct further research on the topic.

After reviewing various methods, comparing traditional machine learning (TML), deep learning (DL), and hybrid approaches in accent classification revealed several important insights. The k-NN algorithm is the most effective TML method, as it outperformed other methods to classify six accents from the UK dialect containing 17,877 audio samples and used MFCC as an input feature. Despite this dataset being unbalanced in gender and the number of samples in each class, k-NN yielded an accuracy of 98.48%. For DL models, transfer learning with the xResNet18 model yielded an accuracy of 100% when utilizing a well-structured dataset (gender-balanced and balanced uniform content) to classify eight English accent-suppressing other DL models. This result ensures that a balanced and parallel dataset is important for accent classification accuracy. For a dataset that is relatively not large, CNN achieved a high accuracy of 96.4% to 98.7%, outperforming other methods. CNN achieved this by using amplitude Mel-spectrograms as input features and applying data augmentation techniques. The reported accuracy, the highest recorded on the Speech Accent Archive dataset, highlights its potential real-world applications. The fine-tuned transformer Wav2Vec2 achieves an overall accuracy of 99.74% utilizing a large, balanced, diverse dataset of six English accents, showing outstanding performance in raw audio-based accent classification.

The study concludes by identifying key research gaps and proposing future directions to advance accent recognition systems. Insights from previous publications also reveal regional tendencies in accent classification and offer valuable guidance for addressing current challenges and inspiring innovative approaches in future research.

Author Contributions

Conceptualization, S.J. and H.A.; methodology, S.J.; software, S.J.; validation, S.J. and H.A.; formal analysis, S.J.; investigation, S.J.; resources, H.A.; data curation, S.J.; writing—original draft preparation, S.J.; writing—review and editing, H.A.; visualization, S.J.; supervision, H.A.; project administration, H.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflict of interest.

Data Availability Statement

Data sharing is not applicable to this article, as no datasets were generated or analyzed during the current study

References

1. M. Habbash *et al.*, “Recognition of Arabic Accents From English Spoken Speech Using Deep Learning Approach,” *IEEE Access*, vol. 12, pp. 37219–37230, 2024, doi: 10.1109/ACCESS.2024.3374768.
2. J. Grigaliūnaitė and G. A. Melnik-Leroy, “Automatic Accent Identification Using Less Data: a Shift from Global to Segmental Accent,” *Arab. J. Sci. Eng.*, 2024, doi: 10.1007/s13369-024-09344-4.

3. W. O'Grady, J. Archibald, M. Aronoff, and J. Rees-Miller, *Contemporary Linguistics Analysis (Eight Edition)*. 2016.
4. D. Walsh, S. Dev, and A. Nag, "Hilbert-Huang-Transform Based Features for Accent Classification of Non-Native English Speakers," in *2023 34th Irish Signals and Systems Conference, ISSC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISSC59246.2023.10162075.
5. O. Aju, V. O.-C. J. of, and undefined 2022, "A Review of Accent-Based Automatic Speech Recognition Models for E-Learning Environment," *Journals.Covenantuniversity.Edu.Ng*, vol. 10, no. 2, 2022, [Online]. Available: https://journals.covenantuniversity.edu.ng/index.php/cjict/article/view/3146
6. J. J. Bird, A. Ekárt, E. Wanner, and D. R. Faria, "Accent Classification in Human Speech Biometrics for Native and Non-native English Speakers," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2019, pp. 554–560. doi: 10.1145/3316782.3322780.
7. Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, 2003, pp. 540–543. doi: 10.1109/icassp.2003.1198837.
8. P. Parikh, K. Velhal, S. Potdar, A. Sikligar, and R. Karani, "English Language Accent Classification and Conversion using Machine Learning." [Online]. Available: https://ssrn.com/abstract=3600748
9. D. Honnavalli and S. S. Shylaja, "Supervised Machine Learning Model for Accent Recognition in English Speech Using Sequential MFCC Features," in *Advances in Intelligent Systems and Computing*, Springer, 2021, pp. 55–66. doi: 10.1007/978-981-15-3514-7_5.
10. Y. Singh, A. Pillay, and E. Jembere, "Features of speech audio for accent recognition," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, icABCD 2020 - Proceedings*, 2020, pp. 1–6. doi: 10.1109/icABCD49160.2020.9183893.
11. "Accent recognition using i-vector , gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech Bahari , M . H .; Saeidi , R .; Van hamme , H .; Leeuwen , D . A . van 2013 , Article in monograph or in proceedi," pp. 7344–7348, 2024.
12. R. K. Guntur, K. Ramakrishnan, and M. Vinay Kumar, "An Automated Classification System Based on Regional Accent," *Circuits, Syst. Signal Process.*, vol. 41, no. 6, pp. 3487–3507, Jun. 2022, doi: 10.1007/s00034-021-01948-7.
13. N. J. Ibrahim, M. Y. I. Idris, M. Y. Z. M. Yusoff, N. N. A. Rahman, and M. I. Dien, "Robust Feature Extraction Based On Spectral And Prosodic Features For Classical Arabic Accents Recognition," *Malaysian J. Comput. Sci.*, vol. 2019, no. Special Issue 3, pp. 46–72, 2019, doi: 10.22452/mjcs.sp2019no3.4.
14. R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. Peter Campbell, "Introduction to machine learning, neural networks, and deep learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 1–12, 2020, doi: 10.1167/tvst.9.2.14.
15. P. D. McNicholas and P. A. Tait, "Supervised Learning," in *Data Science with Julia*, Springer, 2019, pp. 93–128. doi: 10.1201/9781351013673-5.
16. Krugman, *International Economics: theory and Policy: theory and Policy*. New York: Prentice Hall, 2003.
17. M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1993, pp. 399–402. doi: 10.1109/icassp.1993.319323.
18. L. M. Arslan and J. H. L. Hansen, "Improved HMM training and scoring strategies with application to accent classification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, 1996, pp. 589–592. doi: 10.1109/icassp.1996.543189.
19. T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *2001 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2001 - Conference Proceedings*, IEEE, 2001, pp. 343–346. doi: 10.1109/ASRU.2001.1034657.
20. S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Proceedings - Fourth IEEE Workshop on Automatic Identification Advanced Technologies, AUTO ID 2005*, IEEE, 2005, pp. 139–143. doi: 10.1109/AUTOID.2005.10.
21. H. Tang and A. A. Ghorbani, "Accent classification using support vector machine and hidden markov model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2003, pp. 629–631. doi: 10.1007/3-540-44886-1_65.
22. I. G. and Y. B. and A. Courville, "Deep learning 简介 一 、 什么是 Deep Learning ？," *Nature*, vol. 29, no. 7553, pp. 1–73, 2016, [Online]. Available: http://deeplearning.net/
23. K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, no. September, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.
24. Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3223444.
25. M. A. Humayun, J. Shuja, and P. E. Abas, "A review of social background profiling of speakers from speech accents," *PeerJ Comput. Sci.*, vol. 10, pp. 1–25, 2024, doi: 10.7717/peerj-cs.1984.
26. T. Song, L. T. H. Nguyen, and T. V. Ta, "MPSA-DenseNet: A novel deep learning model for English accent classification," *Comput. Speech Lang.*, vol. 89, Jan. 2025, doi: 10.1016/j.csl.2024.101676.
27. R. Patel and S. Patel, *Deep Learning for Natural Language Processing*, vol. 190. 2021. doi: 10.1007/978-981-16-0882-7_45.
28. J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey," *IETE Tech. Rev. (Institution Electron. Telecommun. Eng. India)*, vol. 32, no. 4, pp. 240–251, 2015, doi: 10.1080/02564602.2015.1010611.

29. Z. Ge, “Improved accent classification combining phonetic vowels with acoustic features,” in *Proceedings - 2015 8th International Congress on Image and Signal Processing, CISP 2015*, IEEE, 2016, pp. 1204–1209. doi: 10.1109/CISP.2015.7408064.
30. Y. Jiao, M. Tu, V. Berisha, and J. Liss, “Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2388–2392, 2016, doi: 10.21437/Interspeech.2016-1148.
31. E. Tverdokhleb, H. Dobrovolskyi, N. Keberle, and N. Myronova, “Implementation of accent recognition methods subsystem for eLearning systems,” in *Proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017*, IEEE, 2017, pp. 1037–1041. doi: 10.1109/IDAACS.2017.8095243.
32. M. Rizwan and D. V. Anderson, “A weighted accent classification using multiple words,” *Neurocomputing*, vol. 277, pp. 120–128, 2018, doi: 10.1016/j.neucom.2017.01.116.
33. F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, “Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors,” *J. Mater. Chem. C*, vol. 3, no. 4, pp. 10715–10722, 2015, doi: 10.1039/b000000x.
34. D. S. Widyowaty and A. Sunyoto, “Accent Recognition by Native Language Using Mel-Frequency Cepstral Coefficient and K-Nearest Neighbor,” *2020 3rd Int. Conf. Inf. Commun. Technol. ICOIACT 2020*, pp. 314–318, 2020, doi: 10.1109/ICOIACT50329.2020.9332026.
35. M. F. Hossain, M. M. Hasan, H. Ali, M. R. K. R. Sarker, and M. T. Hassan, “A machine learning approach to recognize speakers region of the united kingdom from continuous speech based on accent classification,” *Proc. 2020 11th Int. Conf. Electr. Comput. Eng. ICECE 2020*, pp. 210–213, 2020, doi: 10.1109/ICECE51571.2020.9393038.
36. Y. Korkmaz and A. Boyacı, “A comprehensive Turkish accent/dialect recognition system using acoustic perceptual formants,” *Appl. Acoust.*, vol. 193, May 2022, doi: 10.1016/j.apacoust.2022.108761.
37. K. Radha, M. Bansal, and S. M. Shabber, “Accent Classification of Native and Non-Native Children using Harmonic Pitch,” in *2022 2nd International Conference on Artificial Intelligence and Signal Processing, AISP 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/AISP53593.2022.9760588.
38. A. Purwar, H. Sharma, Y. Sharma, H. Gupta, and A. Kaur, “Accent classification using Machine learning and Deep Learning Models,” in *Proceedings of 2022 1st International Conference on Informatics, ICI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 13–18. doi: 10.1109/ICI53355.2022.9786885.
39. S. M. S. I. Badhon, H. Rahaman, F. R. Rupon, and S. Abujar, “Bengali Accent Classification from Speech Using Different Machine Learning and Deep Learning Techniques,” in *Advances in Intelligent Systems and Computing*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 503–513. doi: 10.1007/978-981-15-7394-1_46.
40. N. J. Ibrahim, M. Y. I. Idris, M. Y. Z. M. Yusoff, R. Ramli, and R. J. Raja Yusof, “The Study of Malay’s Prosodic Features Impact on Classical Arabic Accents Recognition,” *IEEE Access*, vol. 11, pp. 94589–94612, 2023, doi: 10.1109/ACCESS.2023.3299814.
41. F. Weninger, Y. Sun, J. Park, D. Willett, and P. Zhan, “Deep learning based Mandarin accent identification for accent robust ASR,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 510–514. doi: 10.21437/Interspeech.2019-2737.
42. A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, “VFNet: A Convolutional Architecture for Accent Classification,” Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.06697
43. U. Singh, A. Gupta, D. Bisharad, and W. Arif, “Foreign accent classification using deep neural nets,” in *Journal of Intelligent and Fuzzy Systems*, IOS Press, 2020, pp. 6347–6352. doi: 10.3233/JIFS-179715.
44. R. Kethireddy, S. R. Kadiri, and S. V. Gangashetty, “Learning Filterbanks from Raw Waveform for Accent Classification,” *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9206778.
45. P. Berjon, A. Nag, and S. Dev, “Analysis of French phonetic idiosyncrasies for accent recognition,” *Soft Comput. Lett.*, vol. 3, no. June, p. 100018, 2021, doi: 10.1016/j.socl.2021.100018.
46. Z. Zhang, Y. Wang, and J. Yang, “Accent Recognition with Hybrid Phonetic Features,” *Sensors (Basel).*, vol. 21, no. 18, Sep. 2021, doi: 10.3390/s21186258.
47. Z. Lou and Y. Ren, “Investigating Issues with Machine Learning for Accent Classification,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1738/1/012111.
48. Z. Al-Jumaili, T. Bassiouny, A. Alanezi, W. Khan, D. Al-Jumeily, and A. J. Hussain, “Classification of Spoken English Accents Using Deep Learning and Speech Analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 277–287. doi: 10.1007/978-3-031-13832-4_24.
49. V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, “Language Accent Detection with CNN Using Sparse Data from a Crowd-Sourced Speech Archive,” *Mathematics*, vol. 10, no. 16, Aug. 2022, doi: 10.3390/math10162913.
50. S. Darshana, H. Theivaprakasham, G. Jyothish Lal, B. Premjith, V. Sowmya, and K. Soman, “MARS: A Hybrid Deep CNN-based Multi-Accent Recognition System for English Language,” in *2022 1st International Conference on Artificial Intelligence Trends and Pattern Recognition, ICAITPR 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICAITPR51569.2022.9844177.
51. Y. A. Wubet, D. Balram, and K. Y. Lian, “Intra-Native Accent Shared Features for Improving Neural Network-Based Accent Classification and Accent Similarity Evaluation,” *IEEE Access*, vol. 11, no. November 2022, pp. 32176–32186, 2023, doi: 10.1109/ACCESS.2023.3259901.

52. O. Cetin, “Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network,” *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 1973–1990, Feb. 2023, doi: 10.1007/s13369-022-07086-9.
53. E. Alsharhan and A. Ramsay, “Robust automatic accent identification based on the acoustic evidence,” *Int. J. Speech Technol.*, vol. 26, no. 3, pp. 665–680, Sep. 2023, doi: 10.1007/s10772-023-10031-2.
54. A. Carofilis, E. Alegre, E. Fidalgo, and L. Fernandez-Robles, “Improvement of Accent Classification Models Through Grad-Transfer from Spectrograms and Gradient-Weighted Class Activation Mapping,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2859–2871, 2023, doi: 10.1109/TASLP.2023.3297961.
55. W. Ke, “Study on recognition and classification of English accents using deep learning algorithms,” *J. Intell. Syst.*, vol. 32, no. 1, Jan. 2023, doi: 10.1515/jisys-2023-0174.
56. K. Mebarkia and A. Reffad, “CNN models for Maghrebian accent recognition with SVM silence elimination,” *Signal, Image Video Process.*, vol. 18, no. 6–7, pp. 5089–5098, Aug. 2024, doi: 10.1007/s11760-024-03217-x.
57. B. T. Ta, N. M. Le, and V. H. Do, “Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language,” *Eng. Appl. Artif. Intell.*, vol. 132, Jun. 2024, doi: 10.1016/j.engappai.2024.107895.
58. K. Kashif, A. Alwan, Y. Wu, L. De Nardis, and M. G. Di Benedetto, “MKELM based multi-classification model for foreign accent identification,” *Heliyon*, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e36460.
59. O. Ozturk, H. Kilimci, H. H. Kilinc, and Z. H. Kilimci, “Spoken Accent Detection in English Using Audio-Based Transformer Models,” *UBMK 2024 - Proc. 9th Int. Conf. Comput. Sci. Eng.*, pp. 539–544, 2024, doi: 10.1109/UBMK63289.2024.10773414.
60. Linguistic Data Consortium, “Foreign Accented English (LDC2007S08), Philadelphia: University of Pennsylvania,” 2007, [Online]. Available: https://catalog.ldc.upenn.edu/LDC2007S08
61. L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa TIMIT: Acoustic-Phonetic Continuous Speech Corpus CD-ROM,” pp. 1–79, 1990, [Online]. Available: http://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf
62. K. J. van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, “The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles,” *Lang. Speech*, vol. 53, no. 4, pp. 510–540, 2010, doi: 10.1177/0023830910372495.
63. B. Schuller *et al.*, “The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2001–2005, 2016, doi: 10.21437/Interspeech.2016-129.
64. R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” *arXiv Prepr. arXiv1912.06670*, 2019.
65. V. Community, “VoxForge Speech Corpus”, [Online]. Available: https://www.voxforge.org/home/about
66. P. Ahamad, Afroz and Anand, Ankit and Bhargava, “AccentDB,” *Accent. A Database Non-Native English Accent. to Assist Neural Speech Recognit.*, pp. 5351--5358, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.659
67. S. Weinberger, *George Mason University’s Speech Accent Archive*. 2013. [Online]. Available: http://accent.gmu.edu/about.php
68. C. Veaux, J. Yamagishi, and K. MacDonald, *English multi-speaker corpus for CSTR voice cloning toolkit*. 2017. [Online]. Available: https://doi.org/10.7488/ds/1994
69. I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, “Opensource multispeaker corpora of the english accents in the british isles,” *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. 979-10-95546-34–4, pp. 6532–6541, 2020, [Online]. Available: https://aclanthology.org/2020.lrec-1.804/
70. X. Shi *et al.*, “The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, 2021, pp. 6918–6922. doi: 10.1109/ICASSP39728.2021.9413386.
71. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.
72. S. Mnasri, “ArL2Eng dataset to recognize Arabic accents from English speech”, [Online]. Available: osf.io/thv6p
73. K. M. Meghan Glenn, Haejoong Lee, Stephanie Strassel, “GALE Phase 3 Arabic Broadcast News Transcripts,” *Linguist. Data Consort.*, 2016, [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016T17
74. S. N. Mohammed and A. K. Hassan, “Automatic voice activity detection using fuzzy-neuro classifier,” *J. Eng. Sci. Technol.*, vol. 15, no. 5, pp. 2854–2870, 2020.
75. S. Graf, T. Herbig, M. Buck, and G. Schmidt, “Features for voice activity detection: a comparative analysis,” *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, 2015, doi: 10.1186/s13634-015-0277-z.
76. S. Wei, S. Zou, F. Liao, and W. Lang, “A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification,” *J. Phys. Conf. Ser.*, vol. 1453, no. 1, 2020, doi: 10.1088/1742-6596/1453/1/012085.
77. M. H. Tanveer, H. Zhu, W. Ahmed, A. Thomas, B. M. Imran, and M. Salman, “Mel-spectrogram and Deep CNN Based Representation Learning from Bio-Sonar Implementation on UAVs,” *2021 Int. Conf. Comput. Control Robot. ICCCR 2021*, pp. 220–224, 2021, doi: 10.1109/ICCCR49711.2021.9349416.
78. K. Murakami, K. Araki, M. Hiroshige, and K. Tochinai, “Effectiveness of a direct speech transform method

using inductive learning from laryngectomee speech to normal speech," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2003, pp. 686–698. doi: 10.1007/978-3-540-24581-0_59.
79. J. H. L. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification," *Speech Commun.*, vol. 52, no. 10, pp. 777–789, 2010, doi: 10.1016/j.specom.2010.05.004.
80. S. Michieletto, F. Stival, and E. Pagello, *A probabilistic approach to reconfigurable interactive manufacturing and coil winding for Industry 4.0*. INC, 2020. doi: 10.1016/B978-0-12-818906-1.00003-6.
81. A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," *Odyssey 2012 - Speak. Lang. Recognit. Work.*, no. August, pp. 28–33, 2012.
82. H. Hallawi, A. F. Almukhtar, D. A. Nasrawi, A. D. Salah, and T. Z. Faisal, "Gender Classification Based on Iraqi Names Using Machine Learning," *Iraqi J. Sci.*, vol. 65, no. 11, pp. 6725–6737, 2024, doi: 10.24996/ijs.2024.65.11.42.
83. M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.
84. L. Zhang, D. Zhang, and F. Tian, "SVM and ELM: Who Wins? Object Recognition with Deep Convolutional Features from ImageNet," in *Proceedings of ELM-2015 Volume 1: Theory, Algorithms and Applications (I)*, Springer, 2016, pp. 249–263. doi: 10.1007/978-3-319-28397-5_20.
85. J. Chorowski, J. Wang, and J. M. Zurada, "Review and performance comparison of SVM- and ELM-based classifiers," *Neurocomputing*, vol. 128, pp. 507–516, 2014, doi: 10.1016/j.neucom.2013.08.009.
86. X. Liu, C. Gao, and P. Li, "A comparative analysis of support vector machines and extreme learning machines," *Neural Networks*, vol. 33, pp. 58–66, 2012, doi: 10.1016/j.neunet.2012.04.002.
87. L. J. Su and M. Yao, "Extreme learning machine with multiple kernels," in *IEEE International Conference on Control and Automation, ICCA*, IEEE, 2013, pp. 424–429. doi: 10.1109/ICCA.2013.6565148.
88. S. Mohammed, L. George, and H. Dawood, "The Effect of Classification Methods on Facial Emotion Recognition Accuracy," *Br. J. Appl. Sci. Technol.*, vol. 14, no. 4, pp. 1–11, 2016, doi: 10.9734/bjast/2016/23090.
89. H. A. Abdulmohsin, B. Al-Khateeb, S. S. Hasan, and R. Dwivedi, "Automatic illness prediction system through speech," *Comput. Electr. Eng.*, vol. 102, no. July, p. 108224, 2022, doi: 10.1016/j.compeleceng.2022.108224.
90. J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey," *IETE Tech. Rev. (Institution Electron. Telecommun. Eng. India)*, vol. 32, no. 4, pp. 240–251, 2015, doi: 10.1080/02564602.2015.1010611.
91. R. Gemello, D. Albesano, and F. Mana, "Multi-source neural networks for speech recognition," *Proc. Int. Jt. Conf. Neural Networks*, vol. 5, no. 10, pp. 2946–2949, 1999, doi: 10.1109/ijcnn.1999.835942.
92. H. M. Ahmed and H. H. Mahmoud, "Effect of successive convolution layers to detect gender," *Iraqi J. Sci.*, vol. 59, no. 3, pp. 1717–1732, 2018, doi: 10.24996/IJS.2018.59.3C.17.
93. A. A. R. Hussien and N. A. Z. Abdullah, "A Review for Arabic Sentiment Analysis Using Deep Learning," *Iraqi J. Sci.*, vol. 64, no. 12, pp. 6572–6585, 2023, doi: 10.24996/ijs.2023.64.12.37.
94. P. C. Vakkantula and W. Virginia, *Speech Mode Classification using the Fusion of CNNs and LSTM Networks Speech Mode Classification using the Fusion of CNNs and LSTM Networks Lane Department of Computer Science and Electrical Engineering*. West Virginia University, 2020.
95. T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Anal. Appl.*, vol. 24, no. 2, pp. 423–431, 2021, doi: 10.1007/s10044-020-00921-5.
96. K. Zaman, M. Sah, and C. Direkoglu, "Classification of Harmful Noise Signals for Hearing Aid Applications using Spectrogram Images and Convolutional Neural Networks," *4th Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2020 - Proc.*, 2020, doi: 10.1109/ISMSIT50672.2020.9254451.
97. I. Ozer, O. Cetin, K. Gorur, and F. Temurtas, "Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset," *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14975–14989, 2021, doi: 10.1007/s00521-021-06133-0.
98. M. Ismail *et al.*, "Development of a regional voice dataset and speaker classification based on machine learning," *J. Big Data*, vol. 8, no. 1, pp. 1–18, 2021, doi: 10.1186/s40537-021-00435-9.
99. G. R. Krishna, R. Krishnan, and V. K. Mittal, "A System for Automatic Regional Accent Classification," in *2020 IEEE 17th India Council International Conference, INDICON 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/INDICON49873.2020.9342577.