Heliyon 10 (2024) e36460

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

MKELM based multi-classification model for foreign accent identification

Kaleem Kashif [a,*], Abeer Alwan [b], Yizhi Wu [c], Luca De Nardis [a], Maria-Gabriella Di Benedetto [a]

[a] Department of Information Engineering, Electronics and Telecommunication, Sapienza University Rome, Rome, 00184, Italy
[b] Electrical and Computer Engineering Department, University of California, Los Angeles, Los Angeles, CA 90095, USA
[c] Information Science & Technology, Donghua University, Shanghai, 201620, PR China

ARTICLE INFO

Keywords:
Foreign accent identification (FAID)
Multi-kernel extreme learning machine (MKELM)
Weighted classification scheme (WCS)

ABSTRACT

The automatic identification of foreign accents can play a crucial role in various speech systems, including speaker identification, e-learning, telephone banking, and more. Additionally, it can greatly enhance the robustness of Automatic Speech Recognition (ASR) systems. Non-native accents in speech signals are characterized by distinct pronunciations, prosody, and voice characteristics of the speaker. However, automatically identifying foreign accents poses significant challenges, particularly in the context of multi-class modeling. Multi-classification models face difficulties in achieving high performance and dealing with computational challenges when confronted with multi-dimensional and unbalanced datasets, such as those with more than two accents. Furthermore, the choice of features remains a bottleneck problem for Foreign Accent Identification (FAID), further hindering performance in these tasks. Consequently, the accuracy of current systems is typically low. To address these challenges, this paper proposes a framework based on the Multi-Kernel Extreme Learning Machine (MKELM) model for the multi-classification of FAID. The MKELM model utilizes a novel weighted scheme to classify various non-native English accents, including Arabic, Chinese, Korean, French, and Spanish. The model first combines Mel-frequency cepstral coefficients (MFCCs) and prosodic features as input, trains pairwise binary classifiers independently, and subsequently employs a weighting scheme to distinguish between classes and identify accents. Through experiments, the proposed model achieves an accuracy rate of 84.72% using a paired weighting scheme. In contrast, the accuracy rate drops to 66.5% when employing the traditional non-weighted multi-classification scheme. A comparison with other models demonstrates the significant advantages of the proposed model in FAID multi-class classification, showcasing improved accuracy, reduced computational complexity (requiring fewer computations, faster learning rates, and shorter training time), and enhanced stability compared to state-of-the-art classification methods.

* Corresponding author.
E-mail address: kaleem.kashif@uniroma1.it (K. Kashif).

https://doi.org/10.1016/j.heliyon.2024.e36460
Received 9 December 2023; Received in revised form 15 August 2024; Accepted 15 August 2024

2405-8440/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Foreign accent identification (FAID) involves the task of determining the native language (L1) of non-native speakers when they speak a second language (L2) [1]. This task has gained significant attention within the speech community due to the adverse impact accents have on the accuracy of conventional automatic speech recognition (ASR) systems (e.g., [2]). Most existing ASR systems are designed for native speech, and their recognition rates significantly decline when words or sentences are pronounced with a foreign accent [3]. Foreign accent variations also have negative effects on automatic speaker and language recognition systems [4,5]. Additionally, FAID is a subject of immense interest in intelligence and security domains, such as immigration screening and border monitoring systems. Other applications such as voice conversion, soft biometrics, telephone-based assistance, e-learning, voice mail, voice dialing, and online banking can also benefit from FAID. [6]. It's worth noting that accent variations can also arise from regional or dialectal pronunciations. Regional accents pertain to changes in pronunciation and speaking style [7,8] predominantly among native speakers of a language. On the other hand, dialect variations can involve differences in lexicon and grammar. Prior research suggests that foreign accents may have assimilated regional or dialectal accents, as they deviate from the established reference language in terms of phonetic realization of vowels and consonants, rhythmic characteristics, prosody, and speaking style [9–12].

Foreign accents are distinct from regional accents or dialects, as the deviation from standard pronunciation is influenced by the speaker's native language (L1) and its impact on their second language (L2) proficiency [13]. In the case of foreign accents, the pronunciation of words can vary significantly based on the speaker's L1 and their level of proficiency in L2, which further complicates the problem.

Non-native speakers often modify certain phonetic features when producing words in L2, as they may only possess partial mastery of its pronunciation. For example, Arabic native speakers from the Gulf region frequently do not aspirate bilabial voiced /b/ and its voiceless counterpart /p/, as well as voiced alveolar /d/ and its counterpart /t/ when speaking English [14]. Additionally, non-native speakers often substitute an unfamiliar L2 phoneme with the closest equivalent from their L1 phoneme inventory [15].

Furthermore, the degree of foreign accent can vary within the same native language depending on the non-native speaker's proficiency in L1 [6,16]. This paper focuses on addressing the problem of Foreign Accent Identification (FAID). FAID is a pivotal area in the domain of speech processing, attracting extensive research interest due to its applications in linguistics, language learning, and automated speech systems. This review aims to provide an exhaustive analysis of the computational methodologies employed in FAID, charting their evolution from traditional machine learning models to cutting-edge deep learning techniques. Early FAID research predominantly utilized traditional machine learning methods, such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and Linear Discriminant Analysis (LDA). These models were chosen for their robustness in handling sequential data and their effectiveness in classification tasks. HMMs have been a cornerstone in FAID research due to their proficiency in modeling temporal sequences of speech signals. Kat et al. [17] utilized HMMs with prosodic features to identify foreign accents in English spoken by Cantonese speakers. Similarly, Kumpf and King [18] employed phonotactic features with HMMs to distinguish between Lebanese Arabic and South Vietnamese accents in English. Hansen et al. [1] further extended this approach by incorporating source generator-based prosodic features to classify Turkish, Chinese, and German accents, demonstrating the versatility of HMMs in handling diverse linguistic backgrounds. GMMs have been favored for their probabilistic framework, which effectively models the distribution of speech features. Phapatanaburi et al. [19] used Mel-frequency cepstral coefficients (MFCCs) as features with GMMs to identify Japanese speakers' accents in English. Fohr and Illina [20] applied GMMs with prosodic features such as pitch and energy to classify accents from French, Italian, and Greek speakers. The probabilistic nature of GMMs allows for robust modeling of the variability in speech signals. Choueiter et al. [21] combined heteroscedastic LDA with Maximum Mutual Information (MMI) for multi-class accent classification, using the FAE corpus and 13-dimensional PLP-based feature vectors. This approach underscored the potential of LDA in enhancing the discriminative power of FAID systems. To address the limitations of traditional models, researchers explored Support Vector Machines (SVM) for FAID. SVMs have shown superior performance in high-dimensional feature spaces and are effective for both binary and multi-class classification. Kashif et al. [14] conducted binary classification using SVMs with MFCC features to identify Arabic accents in English. This study demonstrated the high discriminative capability of SVMs in accent classification. Bahari et al. [22] extended this approach to multiple languages, creating a 60-dimensional feature vector including energy and its derivatives for classifying Russian, Hindi, Thai, Vietnamese, and Cantonese accents. The study highlighted the scalability of SVMs in handling multi-class FAID tasks. The i-vector representation has become a prominent technique in speaker and accent recognition due to its compact and discriminative nature. Behravan et al. [23] utilized i-vector modeling to successfully identify seven different foreign accents with English as the L2 language. This approach marks a significant advancement in FAID, offering a robust and efficient representation of speech features for accent identification. The advent of deep learning has revolutionized FAID, with models such as Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Artificial Neural Networks (ANN), and Deep Belief Networks (DBN) demonstrating substantial improvements in performance. These models leverage large datasets and complex architectures to capture intricate patterns in speech data. Sheng and Edmund [24] used CNNs with MFCC features to identify Chinese and Korean accents in English. The spatial hierarchies captured by CNNs have proven effective in accent classification tasks. Jiao et al. [25] combined DNNs and RNNs to develop a FAID model, leveraging the strengths of both architectures in handling sequential data. This hybrid approach illustrates the power of integrating deep learning models for enhanced performance. Purwar et al. [26] proposed a hybrid model using LSTM and CNN for multi-class accent classification, considering various native languages including Arabic, Dutch, French, Hindi, Korean, Mandarin, Portuguese, Spanish, and Russian. This study showcases the effectiveness of combining temporal and spatial modeling capabilities for FAID. Upadhyay and Lui [27] applied DBNs with MFCC features to classify accents from speakers of six different native languages. The hierarchical feature learning in DBNs contributes to their robust performance in FAID tasks. The techniques used in FAID have

Table 1
Summary of the number of Identified Accents, L2, Acoustic Features, Classification Model used and type of Accent Class for several previous studies and this study.

Study	L2	Accents Identified	Features	Model	Performance%
Kat L.W. et al. [17]	English	2	Prosodic	HMM	73.38
Kumpf K. et al. [18]	English	2	Phonotactic	HMM	76.06
Hansen J.H. et al. [1]	English	3	Prosodic	HMM	88.09
Phapatanaburi K. et al. [19]	English	2	MFCCs based	GMM, DNN	93.00
Fohr D. et al. [20]	English	3	Prosodic	GMM	83.03
Choueiter G. et al. [21]	English	23	plp-based vector	HLDA+MMI	32.07
Kashif K. et al. [14]	English	2	MFCCs based	SVM	88.00
Bahari M.H. et al. [22]	English	5	MFCCs based	SVM	58.00
Behravan H. et al. [38]	English	7	Attributes	i-Vector	57.03
Sheng L.M.A. et al. [24]	English	3	MFCCs based	CNN, MLP	88.00
Jiao Y. et al. [25]	English	11	MFCCs based	DNN+RNN	52.48
Upadhyay R. et al. [27]	English	6	MFCCs based	DBN	90.02
Purwar A. et al. [26]	English	9	MFCCs based	CNN+LSTM	97.36
Rizwan M. et al. [29]	English	7	MFCCs based	ELM, SVM	76.92
Bryant M. et al. [31]	English	5	MFCCs based	GDA, NB	63.86
Widyowaty D.S. et al. [32]	English	5	MFCCs based	CNN	51.96
Singh Y. et al. [33]	English	5	MFCCs based	CNN	70.38
Widyowaty D.S. et al. [34]	English	6	MFCCs based	KNN	57.00
Ensslin A. et al. [35]	English	3	Spectrogram based	CNN	61.00
Parikh P. et al. [36]	English	3	MFCCs based	CNN, DNN, RNN	68.67
Weninger F. et al. in [30]	Chinese	3	i-vector-based	DBN, bLSTM	76.00
Chen T. et al. [28]	Chinese	4	MFCCs based	GMM	65.00
Berjon P. et al. [37]	French	5	Spectrogram based	2-Layer CNN	70.65
Abbas K. et al. [39]	Swiss German	7	phoneme-to-grapheme based	wav2vec	52.08
Eiman A. et al. [40]	Arabic	5	HMM phoneme based	DNN	86.00
Present Study	**English**	**6**	**MFCCs+Prosodic**	**MKELM**	**84.72**

also been successfully applied to regional accent identification, highlighting their versatility. Chen et al. [28] used GMMs with prosodic features for Chinese Mandarin regional accent identification. Rizwan et al. [29] employed Extreme Learning Machines (ELM) with MFCC features for identifying US regional accents. Weninger et al. [30] investigated SVMs, DNNs, and BLSTMs for Mandarin regional accent identification, incorporating i-vector-based features. These studies underscore the adaptability of FAID methodologies to various dialect and regional accent identification tasks. Recent studies have diversified the feature sets used in FAID, including MFCCs, spectrograms, and fundamental frequency (F0) features. These features have been employed across various datasets to enhance the robustness and accuracy of accent identification models [31–37]. The use of diverse features facilitates the capture of different acoustic properties, thereby improving model performance. The evolution of FAID methodologies from traditional machine learning models to advanced deep learning approaches underscores the continuous improvement in accuracy and robustness. Each method's unique strengths and applications have contributed to a more comprehensive understanding and identification of foreign accents. As the field progresses, integrating multiple models and feature sets, and leveraging large, diverse datasets will likely drive further advancements in FAID, enhancing the capability of automated systems to accurately identify and analyze foreign accents. This comprehensive review highlights the significant strides made in FAID research and sets the stage for future innovations and explorations.

Table 1 provides a summary of the different features, languages, and approaches employed in various studies on foreign accent and regional accent identification, as well as the approach proposed in this paper.

Despite the promising outcomes obtained from the aforementioned classification models, Foreign Accent Identification (FAID) remains a challenging task. GMM models have certain limitations, such as their computational constraints, which make them less suitable for handling multidimensional data. Additionally, GMMs require a relatively longer training time compared to other techniques like SVMs, leading to computational inefficiency [41,42]. On the other hand, SVM models suffer from complexity issues, particularly during the parameter-tuning phase. This algorithm involves several crucial parameters that need to be carefully adjusted in order to achieve optimal classification performance. However, what may work well for one class might not be suitable for another [43]. Furthermore, kernel-based SVMs pose challenges in selecting an appropriate kernel function, which can be a difficult and time-consuming task. Deep learning approaches have demonstrated promising results in FAID; however, they are not without limitations. Challenges associated with deep learning include overfitting, model complexity, and lengthy training times [25]. Therefore, while various models have shown potential in FAID, each approach has its own set of limitations that need to be considered for effective implementation. Extreme Learning Machine (ELM) models have demonstrated promising performance in multi-classification applications [44]. However, ELMs can be complex, particularly when dealing with high-dimensional data, as tuning the hidden layer neurons becomes challenging. Kernel-based ELM, on the other hand, offers an efficient alternative as it eliminates the need to handle hidden layers. However, selecting an appropriate kernel function poses a complex and time-consuming problem. Moreover, in the case of multi-class classification, the use of multiple kernels becomes necessary [45]. Nevertheless, employing a multi-kernel ELM approach in Foreign Accent Identification (FAID) offers two main advantages. Firstly, the multi-kernel function enables the selection of optimal kernels and parameters from a larger set, thereby reducing bias associated with kernel choice. Secondly, by incorporating

multiple kernels, the model can effectively utilize multiple sources of information, leading to improved performance in multi-class classification tasks. Consequently, this paper proposes the utilization of a multi-kernel-based ELM model (MKELM) for FAID. Recently, the Multi-Kernel based ELM model (MKELM) has shown successful applications in various classification tasks. For instance, Zhang et al. [46] utilized MKELM for EEG classification in brain-computer interfaces, while Zhao and Guo [47] proposed an MKELM classifier for Interval Forecasting in Integrated Energy Systems, specifically for Combined Electricity-Heat-Cooling-Gas Loads. Ahuja and Vishwakarma [48] applied MKELM for pattern classification. Although MKELM-based approaches have proven competitive in multi-class classification, to the best of our knowledge, they have not yet been explored in the context of Foreign Accent Identification (FAID).

In classification problems with more than two classes, there are primarily two methods commonly employed. The first approach involves breaking down the multi-class problem into a series of binary problems. Conversely, the second approach tackles the multi-class problem directly. The first approach poses challenges in terms of combining data from each binary classifier and ensuring sufficient representation of each class. On the other hand, the second approach may lead to over-training and a bias towards classes that are more prevalent in the sample or easier to separate [49]. Furthermore, the dimensionality of features needs to be carefully considered as the number of classes increases [50]. In this study, our research objectives are twofold. Firstly, we aim to assess the effectiveness of utilizing a multi-Kernel based Extreme Learning Machine (MKELM) model for Foreign Accent Identification (FAID) by comparing its performance with existing state-of-the-art methods such as SVM, ANN, LSTM, ELM, MLELM, and KELM. Secondly, we seek to develop and evaluate a weighted scheme integrated with the MKELM model to effectively address the challenges associated with multi-classification, particularly when handling high-dimensional data in FAID. By enhancing the model's capability to manage multi-classification issues, we aim to improve its overall accuracy and efficiency. Lastly, we intend to investigate how the proposed MKELM-based FAID model addresses computational constraints, training time, and model stability issues encountered in traditional classification approaches such as SVM, ANN, LSTM, ELM, MLELM, and KELM. This paper introduces a novel architecture for Foreign Accent Identification (FAID) based on a weighted classification scheme. The proposed approach utilizes pairwise weighted schemes and employs majority voting to classify the target class. We evaluate the performance of our MKELM-based FAID model and compare it with other state-of-the-art models such as SVM, ANN, LSTM, ELM, MLELM, and KELM. Our algorithm demonstrates advantages in terms of reduced computations, faster learning rates, shorter training time, and improved model stability compared to existing classification methods.

The remainder of this paper is structured as follows: Section 2 details the research methods employed. Section 3 presents the results of our study. Section 4 offers a discussion of these findings. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Research methods

2.1. The multi-kernel extreme learning machine (MKELM)

In this section, we provide an overview of the fundamental principles behind the Extreme Learning Machine (ELM) algorithm and its kernelized version, known as KELM. Building upon the strengths of KELM, we further enhance its capabilities by introducing the Linear Combination of Kernels (KLC) technique. This technique is based on the Multi-Kernel ELM algorithm (MKELM) and is designed to address multi-classification problems effectively.

The kernel learning method was introduced into the ELM to enhance the stability and generalization capability [51]. The kernel matrix Ω_{ELM}, constructed to replace HH^T, can be defined as shown in equation (1):

$$\Omega_{ELM} = HH^T : \Omega_{ELM_{i,j}} = h(x_i)h(x_j) = K(x_i, x_j) \tag{1}$$

where $h(x)$ represents the hidden layer mapping. The output of KELM can be expressed as follows in equation (2) and (3):

$$f(x) = h(x)H^T\left(I/C + HH^T\right)^{-1} Y \tag{2}$$

$$= \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T (I/C + \Omega_{ELM})^{-1} Y \tag{3}$$

2.2. KLC (kernel linear combination) based proposed approach for multi-kernel ELM

To address the challenge of multi-classification, we extend the Kernel Extreme Learning Machine (KELM) algorithm and propose the Kernels Linear Combination (KLC) approach within the framework of the Multi-Kernel ELM (MKELM) algorithm. While the single-kernel learning approach is retained, the KELM algorithm combines the advantages of ELM and the generalization capability of Support Vector Machine (SVM) methods.

The performance of a classification algorithm is influenced by the choice of kernel parameters and the type of kernel function used. Since KELM relies on a single kernel function, it has limitations in terms of detection accuracy, robustness, and the ability to select the most suitable kernel function for a given multi-classification scenario. Multi-kernel functions offer advantages as they enhance mapping performance by combining kernel functions with diverse features [52].

Mercer's theorem [52] provides a sufficient condition for constructing kernel functions, stating that any semi-positive definite symmetric function can be used as a kernel function. Different kernel functions yield varying effects on the performance of the constructed MKELM model. Fig. 1 illustrates a schematic diagram of this algorithm.

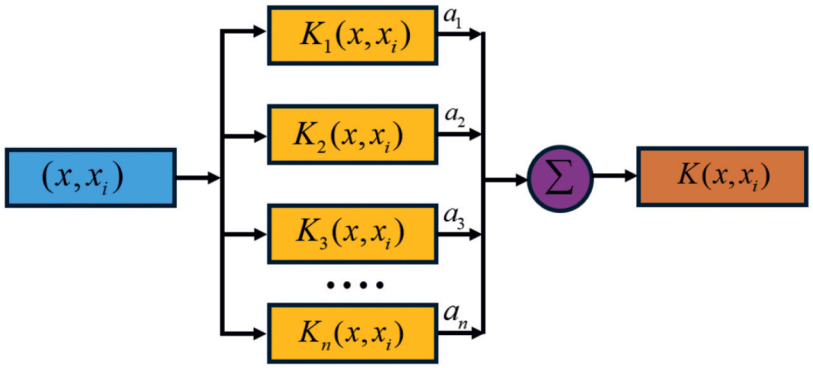

Fig. 1. A diagram of the KLC (Kernel Linear Combination) functions to obtain multi-kernel functions.

Suppose $K(x,x_i)$ is a kernel function that is known, and $\hat{K}(x,x_i)$ is the normalized form of the kernel function, the kernel function can be normalized as follows: $\sqrt{K(x,x)K(x_1,x_1)}$. The algorithm is then created in accordance with the Mercer property of the kernel function, combining the multi-kernel learning method with ELM and the proposed multi-kernel extreme learning machine model. The algorithm is derived as follows:

Typically, the linear combination of multi-kernel functions uses a variety of kernels. e.g., linear kernel $K(x,x_i)=x.x_i$, Gaussian kernel $K(x,x_i)=\exp(-\|x-x_i\|^2/\delta^2)$, and polynomial kernel $K(x,x_i)=(x.x_i+1)^d$. The obtained multi-kernel function overcomes the deficiency present in single-kernel functions, and the form of the multi-kernel function is as follows in equation (4) and (5):

$$K(x,x_i)=a_1K_1(x,x_i)+a_2K_2(x,x_i)+a_3K_3(x,x_i)+...+a_nK_n(x,x_i) \tag{4}$$

$$\text{s.t.}\ \sum_{k=1}^{n}a_k=1,\ \ \forall a_k\geq 0 \tag{5}$$

The optimization problem of the multi-kernel extreme learning machine can be described as in equation (6), (7) and (8):

$$\min L_{MKELM}=\frac{1}{2}\sum_k\frac{1}{a_k}\|w_k\|^2+C\frac{1}{2}\sum_{i=1}^{N}\zeta_i^2 \tag{6}$$

$$s.t.\ \sum\nolimits_k K(x,x_i)w_k=t_i-\xi_i,\ \ i=1,...,N, \tag{7}$$

$$\sum_k\frac{1}{a_k}=1 \tag{8}$$

where w_k is feature weight corresponding to the adopted kernel function $K(x,x_i),\xi_i$ is the predicted error of sample i and C is the regularization parameter to balance model complexity and predictive performance. By replacing the kernel matrix of equation (3) in the ELM with the newly constructed multi-kernel function, the multi-kernel extreme learning machine (MKELM) Algorithm is obtained by the process is shown in Algorithm 1.

Algorithm 1 Multi-kernel ELM algorithm.

Step1: Initialize sample set N,

$$N=\{(x_i,t_i)\,|x_i\in R^n,\ t_i\in R^n, i=1,2,...,N\},$$

Step 2: The best MKELM is created by combining various single-kernel functions according to the multi-kernel formula (4), choosing the best kernel function combination, and figuring out the regularization parameter C and kernel parameters.
Step 3: Weights w and bias b between the input layer and hidden layer that was produced at random using the MKELM algorithm,
Step 4: With training samples as input, hidden layer output matrix H and layer weight matrix β are calculated using equations (3) and (5),
Step 5: Comparing the performance of MKELM made up of various single-kernel functions, evaluating the effectiveness of MKELM using a number of experiments,
Step 6: Return the MKELM classifier along with the classification outcome.

2.3. Experimental setup

This section provides an overview of the corpus utilized in this paper, along with the proposed Foreign Accent Identification (FAID) model and the weighted scheme. Additionally, it outlines the methodology for combining features and describes the attributes of the selected features.

Table 2
Details of the Speech Accent Archive. (GMU) corpus containing 6 different native languages [53].

Native language	Total speakers	Males	Females	Birthplace	Time duration (hh:mm:ss)
English	100	46	54	USA	38:47
Arabic	100	55	45	Saudi Arabia: 97 (54M, 43F) U.A.E: 3 (1M, 2F)	56:20
Chinese	100	34	66	China	53:10
Korean	100	38	62	South Korea	51:20
French	80	41	39	France: 27 (13M, 14F) Canada: 8 (5 M, 3 F) Belgium: 20 (12M, 8F) Switzerland: 15 (8M, 7F) Portugal: 10 (3M, 7F)	37:35
Spanish	100	57	43	Spain: 36 (20M, 16F) Argentina: 17 (12M, 5F) el Salvador: 25 (15M, 10F) Mexico: 22 (10M, 12F)	48:20
Total	**580**	**271**	**309**		**4:45:32**

2.4. Speech corpus

The "Speech Accent Archive" is a collection curated by George Mason University (GMU) [53], comprising of spoken English utterances from individuals with diverse linguistic backgrounds. Each participant recites a paragraph in English that encompasses commonly used English words and challenging pronunciation sounds. The paragraph consists of 69 words, encompassing all English phonemes. The audio files are available in (.mp3) format. The text of the paragraph is as follows:

- *"Please call Stella. Ask her to bring these things with her from the store. Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station".*

According to the GMU website, the recordings were made using Sony TC-D5M along with Radio Shack 33-3001 recording equipment, which includes a Sony mini-disk recorder MDR-70 and a unidirectional dynamic microphone with Sony ECM-MS907 stereo microphone. In this paper, a total of 580 speakers were considered, representing different native languages including English, Arabic, Chinese, Korean, French, and Spanish. The speaker subgroups are labeled based on their respective L1 languages: English (EN-A1), Arabic (AR-A2), Chinese (CN-A3), Korean (KR-A4), French (FR-A5), and Spanish (SP-A6). More information about the corpus can be found in Table 2.

2.5. Feature attributes

Speech features consisted of Mel-frequency cepstrum (MFCCs) and Prosodic features as described in the following subsections.

2.5.1. Mel-frequency cepstrum (MFCC)

MFCCs (Mel-frequency cepstral coefficients) are computed using a series of steps that involve pre-emphasis, windowing, discrete Fourier transform, Mel filter bank, logarithm calculation, and discrete cosine transform. In this paper, the extraction of MFCC features is performed using a Python-based library for speech feature extraction [54].

To apply to the spectrum obtained from the discrete Fourier transform, a bank of triangle filters is constructed. These filters are evenly distributed below 1000 Hz and spaced logarithmically above 1000 Hz. The first 13 cepstral coefficients, along with their first and second derivatives, are considered as part of the feature set. The concatenation of MFCC features and their derivatives is accomplished using the HCopy function from the HTK (Hidden Markov Model Toolkit).

2.5.2. Prosodic features

Prosodic features refer to suprasegmental aspects of speech that capture the manner in which individuals speak [55]. These features play a significant role in identifying accents or dialects as they encompass various stress patterns. In this study, we focus on utilizing pitch, and energy as the selected prosodic features.

- Pitch:

Pitch, also known as fundamental frequency (F0), refers to the rate at which the vocal cords vibrate during the production of voiced sounds. The pitch contour of an utterance holds significant value in various applications, such as voice activity detection, speaker identification, and emotional state recognition. Several methods exist for pitch extraction, with one popular approach being the autocorrelation method in the time domain [56]. The pitch (F0) of a speech signal is computed using Praat's short-time autocorrelation method and pitch stylization functions, utilizing a frame size of 25 ms. The "To Pitch (ac)" function is employed to analyze the pitch

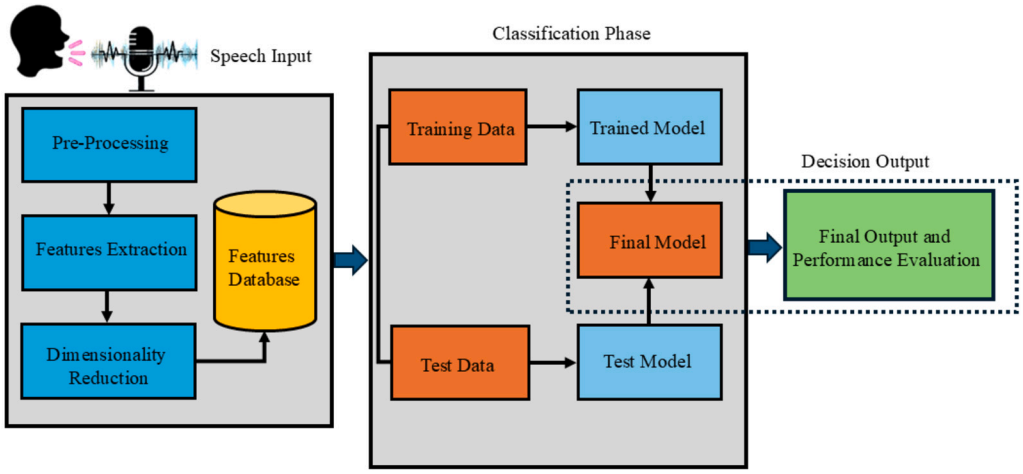

Fig. 2. A block diagram of the proposed system. Input speech features are in (.mp3) files, and after preprocessing (.wav) features are used for feature extraction. The output generated by MFCCs and prosodic features are concatenated to produce the final feature vector. Principal component analysis (PCA) is applied for feature reduction and stored in a database. The classification phase consists of traditional multi-classification methods and a weighted scheme-based pairwise multi-classification method. The last phase is the decision-making phase.

of each frame in the speech signal. Subsequently, the pitch values are converted into a pitch tier using the "To PitchTier" function, enabling the calculation of minimum, maximum, and mean F0 values. These values are obtained through the "Get Mean," "Get Minimum," and "Get Maximum" functions in Praat, respectively.

- Short-time Energy:

Normalized frame-level energy is used as a prosodic feature of accents as follows in equation (9):

$$E_{norm}(i) = \frac{1}{f_L} \sum_{n=1}^{f_L} |s_i(n)|^2 \tag{9}$$

where, $s_i(n),\ n = 1, \ldots f_L$ are the audio samples in the i^{th} frame and f_L is the length of the frame.

2.6. Feature combination

The evaluations in this study incorporated MFCCs (13 features) along with their first and second derivatives, as well as Prosodic features (F0 and Energy) with their respective first and second derivatives. Additionally, the F0 maximum, minimum, and mean values were included as separate features. We conducted a comparative analysis to assess the performance of these feature sets individually and in combination with each other. This led to a 4-way comparison, namely: MFCCs alone, MFCCs combined with F0 features, MFCCs combined with Energy, and MFCCs combined with both F0 features and Energy. The concatenation of these features was performed frame-by-frame.

2.7. System architecture

The block diagram of the proposed system is illustrated in Fig. 2. In this study, we employed the Multi-Kernel ELM Model, as discussed in Section 2, for performing multi-class classification. The input to the model comprises feature vectors extracted from the data. To evaluate the model's performance, the dataset is divided into training (80 percent) and testing (20 percent) sets. Hyperparameter tuning is conducted to optimize the model, and an initial model is constructed based on the selected parameters.

2.8. Weighted scheme architecture

We propose a novel weighted scheme for multi-classification tasks, which utilizes pairwise input data. The block diagram in Fig. 3 illustrates the integration of this weighted scheme with our proposed MKELM model. The algorithm involves three stages:

In the first stage, multiple models are trained using a pair-wise classification approach. This technique helps identify effective decision boundaries [23]. The hyperparameters of each model are learned through cross-validation. In the second stage, the outputs of the multiple models are combined to obtain a classification score. This is achieved by tallying the number of times each class is selected by the models. Subsequently, a weight is assigned to each class based on these counts. In the final stage, the classification score is computed, and the output class decision is made based on the highest score. The maximum count that any class can have is determined by the number of pairwise combinations involving that class, the process is shown in Algorithm 2.

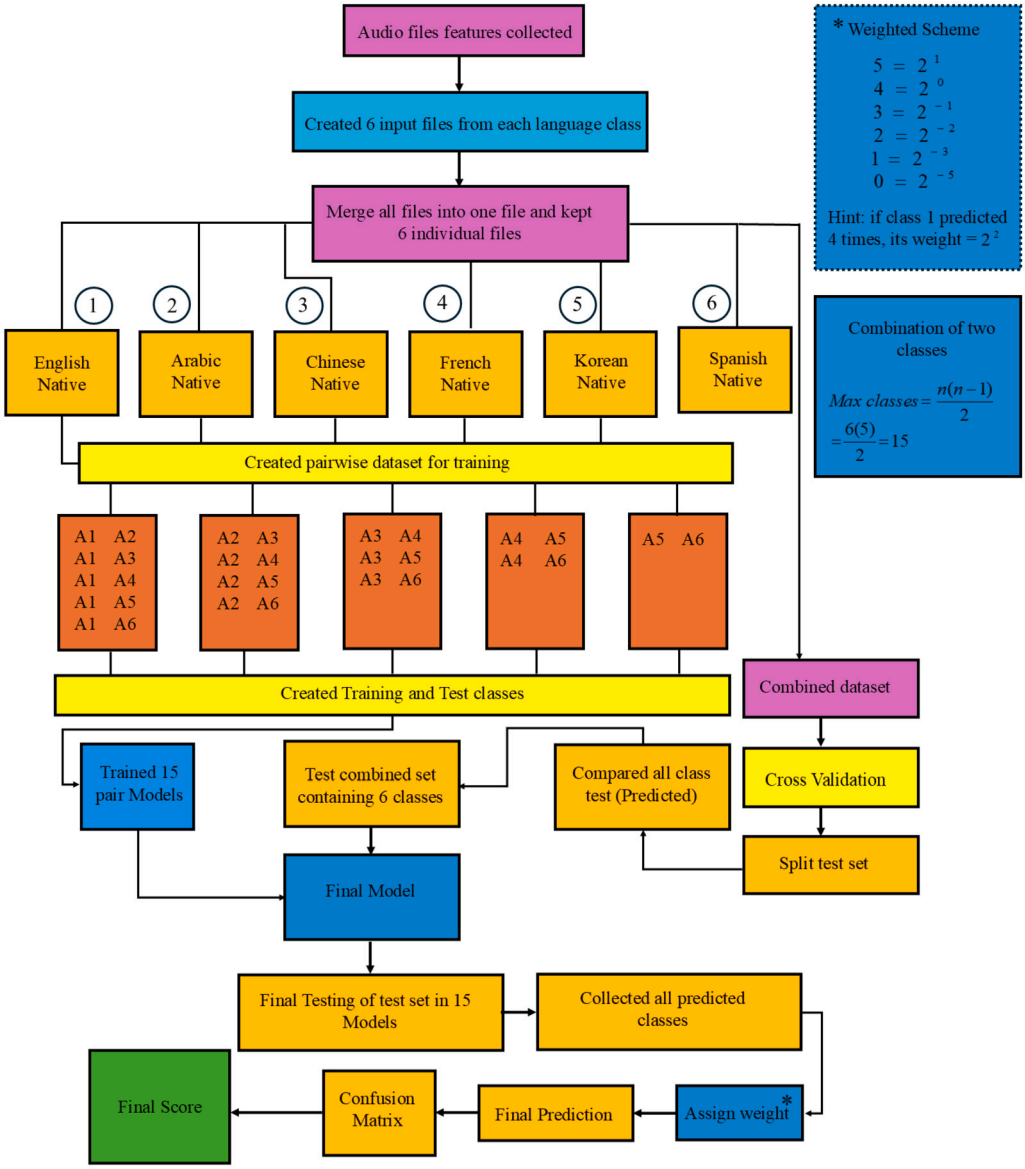

Fig. 3. A block diagram of the weighted scheme for multi-classification using the proposed MKELM model. The input consists of 6 data files (A1, A2, A3, A4, A5, A6) which are then merged into one file, and then 15 pairwise classes are trained and tested using the proposed model. Cross-validation and weight assignment, according to the frequency of each class, are used to make a final decision.

This algorithm has been successfully implemented on various models for accent classification, demonstrating improved overall performance compared to other approaches. Detailed results and analysis will be provided in the subsequent sections.

2.9. Pre-processing

The “Speech Accent Archive” [53] comprises audio files in (.mp3) format, which were converted to (.wav) format specifically for this study. To ensure consistency, the data underwent several preprocessing steps. First, the sampling rate was down-sampled from 44.1 kHz to 16 kHz. Additionally, the audio was converted to mono and adjusted to 16 bits/sample using Winamp. To standardize the maximum amplitude value across all files, Audacity 2.1.1 was utilized.

2.10. Hardware and software tools

The experiments were conducted on a hardware system with the following configuration: Core ™ i-7 7500U CPU (2.70 GHz - 2.90 GHz). All the codes were implemented using Python version 3.10.

Algorithm 2 Weighted scheme algorithm with weight computation.

Require: Input data with n instances and k classes
Ensure: Predicted class for each instance
Divide the data into p pairwise subsets
for $i = 1$ to p **do**
    Train a binary classifier on the i^{th} pairwise subset
    Tune the hyper-parameters of each classifier using cross-validation
end for
Initialize an array W to store weights for each pairwise classifier
for $i = 1$ to p **do**
    Predict the class probabilities for each instance using the i^{th} binary classifier
    Calculate the accuracy of the binary classifier on the validation set
    Calculate the weight w_i for the i^{th} binary classifier based on its accuracy
    $W[i] \leftarrow w_i$
end for
Normalize the weights: $W \leftarrow \frac{W}{\sum_{i=1}^{p} W[i]}$
Initialize an array P to store the combined class probabilities for each instance
for $i = 1$ to p **do**
    Predict the class probabilities for each instance using the i^{th} binary classifier
    Combine the class probabilities using the pairwise weighted scheme: $P \leftarrow P + W[i] \times$ Class Probabilities from i^{th} classifier
end for
Choose the class with the highest score in P as the predicted class for each instance

Table 3
Classification accuracy for all models (SVM, ANN, LSTM, ELM, ML-ELM, KELM, and MKELM) based on different combinations of MFCCs and Prosodic features. All classification results are based on the traditional multi-classification method.

	Accuracy achieved (%)				
Models	MFCCs	MFCCs + Pitch	MFCCs + Energy	MFCCs + Pitch + Energy	(%) Increment due to prosodic features
SVM	34.30	36.0	37.50	38.56	4.26
ANN	16.31	18.70	19.90	20.83	4.52
LSTM	22.0	24.05	25.88	26.93	4.93
ELM	27.40	28.76	29.02	31.88	4.48
ML-ELM	32.29	33.17	35.38	37.0	4.71
KELM	49.70	52.18	53.97	55.0	5.3
MKELM (Proposed)	60.0	62.23	63.99	65.5	5.5

2.11. Implementation of the models

This section presents the results of a series of experiments conducted in two phases. In the first phase, we evaluated the performance of our designed MKELM model using a traditional multi-classification method (one-vs-all classes). We compared its performance with other state-of-the-art models, including SVM, ANN, LSTM, ELM, ML-ELM, and KELM, on a dataset consisting of six classes. The accuracy and performance of each model were recorded and analyzed.

In the second phase, we incorporated a weighted scheme based on pairwise inputs into the multi-classification process of the MKELM model. The results of this phase were also recorded and analyzed. Furthermore, we applied the same novel weighted scheme to the traditional classifiers, and their accuracy and overall performance were recorded and compared.

2.12. Research flowchart

The flowchart shown in the accompanying Fig. 4 provides a systematic overview of the research framework, detailing each stage from sample set initialization to final classification outcomes. Through parameter optimization and rigorous model training, the MKELM model effectively captures accent variations, aided by the Weighted Scheme Algorithm for precise weight computation, ensuring robust and standardized accent identification.

3. Results

The experiments were conducted in four steps, each focusing on a different combination of features: (1) baseline features consisting of only MFCCs, (2) MFCCs combined with Pitch, (3) MFCCs combined with Energy, and (4) all features combined (MFCCs, Pitch, Energy). The results of implementing SVM, ANN, LSTM, ELM, ML-ELM, and KELM models on the six-class dataset are presented and summarized in Table 3. When the SVM model was implemented using only MFCCs as baseline features, an accuracy of 34.30% was achieved. The accuracy improved to 36.0% when MFCCs were combined with Pitch features, 37.50% when combined with Energy features, and reached 38.56% when all features (MFCCs, Pitch, Energy) were combined. The overall increment in accuracy was 4.26% as the prosodic features (Pitch and Energy) were added to the MFCC features.

Similarly, the accuracies of other models improved when MFCCs were combined with prosodic features (F0 and Energy): ANN improved from 16.31% to 20.83%, LSTM from 22.0% to 26.93%, ELM from 27.40% to 31.88%, ML-ELM from 32.29% to 37.0%, and

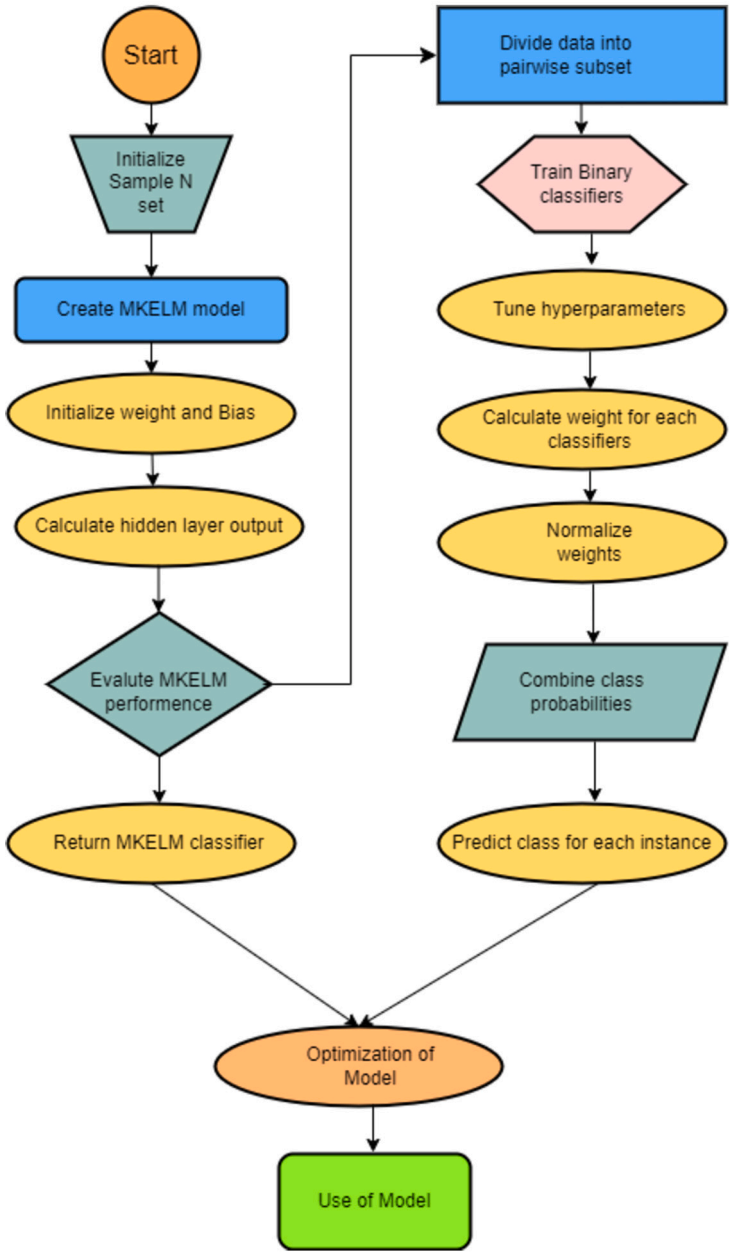

Fig. 4. Flowchart illustrating the systematic methodology for Foreign Accent Identification. The diagram delineates each stage from sample set initialization to final classification outcomes, showcasing the integration of the Multi-Kernels Extreme Learning Machine (MKELM) model and the Weighted Scheme Algorithm with Weight Computation for robust accent identification.

the KELM model's accuracy improved from 49.70% to 55.0% with the combined features. The proposed MKELM model achieved an accuracy of 60.0% with only MFCCs, 62.23% when combined with Pitch features, 63.99% when combined with Energy features, and 65.5% when all features were combined.

In a second experiment, we implemented the proposed MKELM model along with other state-of-the-art models using the novel pairwise weighted scheme architecture. During this phase, the accuracy of each model was evaluated using different combinations of features, and the results are presented in Table 4. The models trained using the pairwise weighted scheme demonstrated improved performance compared to the traditional multi-class classification scheme. For instance, SVM's accuracy improved from 38.56% to 69.90%, ANN from 20.83% to 41.3%, LSTM from 26.93% to 55%, ELM from 31.88% to 68.45%, ML-ELM from 37.0% to 73.0%, KELM from 55.0% to 77.79%, and MKELM from 65.6% to 84.72%.

In summary, the incorporation of prosodic features led to performance improvements of approximately 4-5% for all systems. Notably, the proposed weighted scheme resulted in remarkable enhancements of around 20%. Across all cases, the proposed MKELM model achieved the best results.

Table 4
Classification accuracy for all models (SVM, ANN, LSTM, ELM, ML-ELM, KELM and MKELM) based on different combinations of MFCCs and Prosodic features. All classification results are based on the pairwise weighted multi-classification scheme.

	Accuracy achieved (%)				
Models with Weighted Scheme	MFCCs	MFCCs + Pitch	MFCCs + Energy	MFCCs + Pitch + Energy	(%) Increment due to prosodic features
SVM	65.40	67.90	68.70	69.90	4.5
ANN	37.02	39.10	40.01	41.3	4.3
LSTM	50.91	52.04	53.88	55.0	4.09
ELM	63.80	65.02	66.91	68.45	4.65
ML-ELM	68.01	70.0	71.90	73.0	4.99
KELM	72.10	75.30	76.95	77.79	5.69
MKELM (Proposed)	79.0	82.40	83.19	84.72	5.72

Table 5
Accent-wise Performance Metrics for the MKELM model using the best of a combination of the MFCCs and prosodic features (MFCCs, pitch, and energy) (%).

Accents	Tested (No. of appearance)	Predicted (No. of appearance)	Accuracy (%)	Precision	Recall	F1-score
English (A1)	26	25	96.15	0.86	0.96	0.90
Arabic (A2)	21	19	90.47	0.82	0.90	0.86
Chinese (A3)	24	21	87.50	0.84	0.87	0.85
French (A4)	14	11	78.57	0.73	0.78	0.75
Korean (A5)	16	12	75.0	1.0	0.75	0.85
Spanish (A6)	15	11	73.73	0.91	0.73	0.81
		Macro avg		0.84	0.83	0.83
		Weighted avg		0.84	0.84	0.84
		Final accuracy			**84.72**	

3.1. Prediction accuracy for individual accents

Table 5 presents the accuracy and precision values for each accent obtained through the proposed MKELM model with the weighted scheme. The accuracy of identifying a specific accent relies on the availability of similar accents within a region. For instance, Arabic demonstrates higher accuracy compared to Spanish and French since it encompasses two similar accents (SA and UAE), whereas French and Spanish speakers come from different countries with distinct accent variations. The results indicate that Accent A1 (L1 English) achieved an accuracy of 96.15%, A2 (L1 Arabic) achieved 90.47%, A3 (L1 Chinese) achieved 87.50%, A4 (L1 Korean) achieved 78.57%, A5 (L1 French) achieved 75.0%, and A6 (L1 Spanish) achieved 73.33%.

4. Discussion

In this section we discussed MKELM model's performance through K-Fold cross-validation, emphasizing its robustness after evaluation. It highlights the model's expedited training times and accuracy in accent identification, showcasing its efficacy. Comparative analysis with prior research underscores the model's advancements and acknowledges avenues for improvement.

4.1. Evaluation of model based on K-fold cross-validation

The evaluation of a design model is a crucial step in the classification phase. Once the model has been trained, it is necessary to assess its capabilities and performance. In machine learning, classification models are commonly used to make predictions on input data. During the training phase, a model is trained to predict unknown labels. There are various algorithms with different approaches and techniques for training and prediction. However, it is essential to determine the effectiveness of our proposed model. To evaluate algorithms, we often employ cross-validation, which involves dividing the data into two portions: a test set and a training set. Typically, 20% of the data is reserved for testing, while the remaining 80% is used for training. The training set is utilized to train the designed model, while the test set is used to evaluate the model's performance. Cross-validation, specifically K-Fold-based cross-validation, is a valuable technique in machine learning as it helps identify if a designed model is overfitting or performing adequately.

To ensure the robustness and generalizability of our models, we employed K-Fold cross-validation with varying values of K (5, 10, 15, and 20). In this process, the dataset is divided into K subsets, and the model is trained and tested K times, each time using a different subset as the test set and the remaining subsets as the training set. This process is repeated until all subsets have been utilized for cross-validation. Our proposed MKELM model demonstrated superior stability compared to SVM, LSTM, ANN, ELM, ML-ELM, and KELM models. The stability of the models was evaluated by plotting box plots showing the "mean," "median," and "standard deviation" of the accuracies obtained across the different KFold iterations is shown in Fig. 5 and 6. Fig. 7 illustrates the diagram of the cross-validation model.

We performed statistical analysis on KFold cross validation that demonstrates that the MKELM model significantly outperforms other models in terms of mean accuracy, with all p-values indicating highly significant differences. The confidence intervals for the

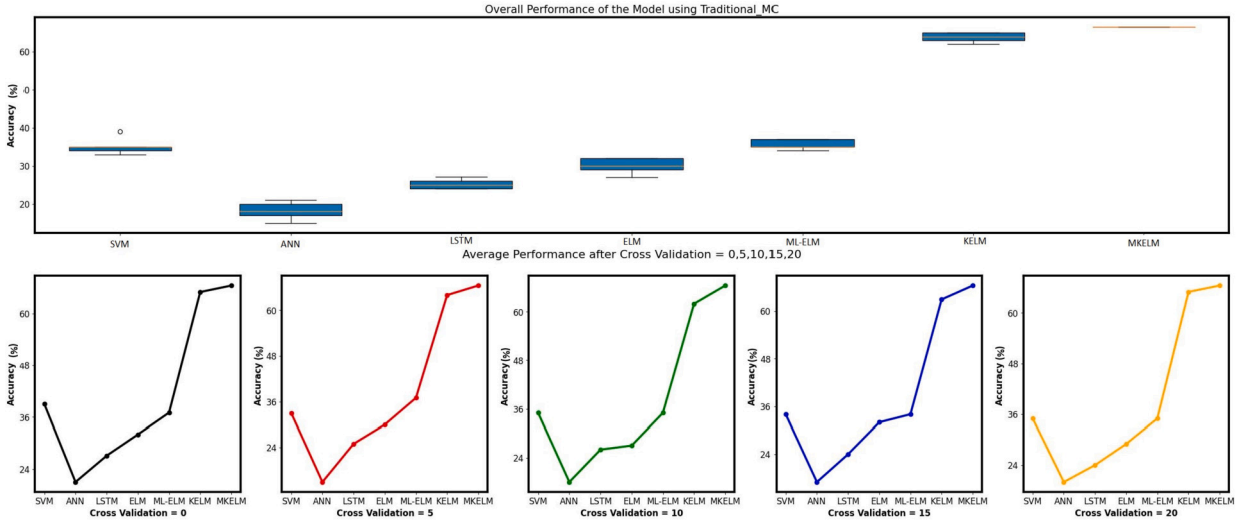

Fig. 5. Performance comparison based on the stability of various models (SVM, ANN, LSTM, ELM, MLELM, KELM, and the proposed MKELM) while using the traditional way of multi-classification. Performance is measured based on the cross validation K0,K5,K10,K15,K20 iterations. Accuracy values vary between 33%-39% for the SVM, 15%-21% for ANN, 24%-27.1% for LSTM, 27%-32% for ELM, 34%-37% for MLELM, 62%-65% for KELM, and for the proposed model, MKELM, it is 65.5% at all K-fold levels.

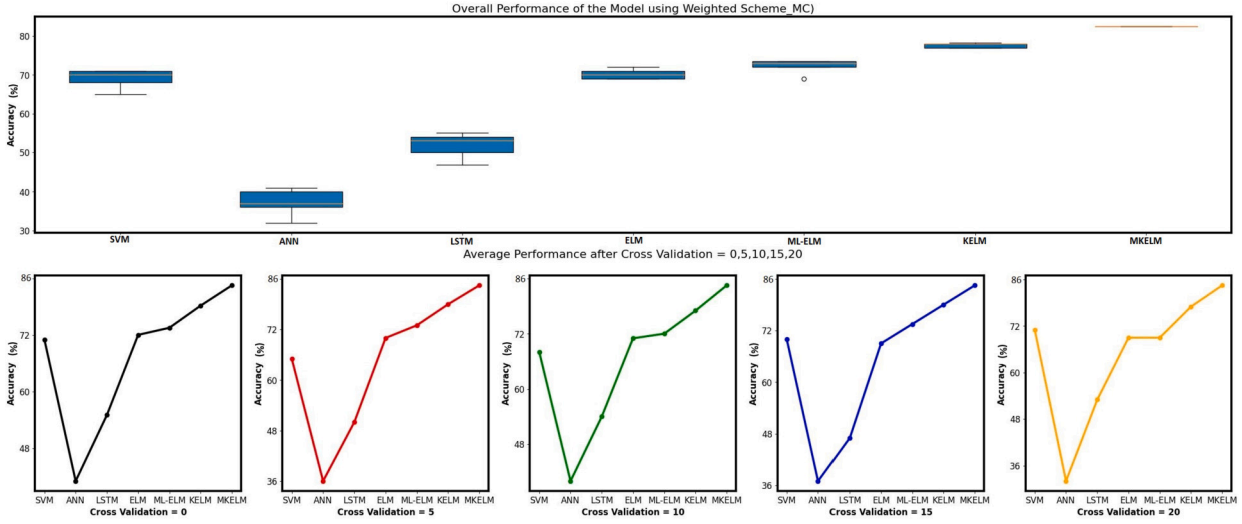

Fig. 6. Performance comparison based on the stability of various models (SVM, ANN, LSTM, ELM, MLELM, KELM, and the proposed MKELM) while using the weighted scheme multi-classification method. Performance is measured based on the cross-validation K0, K5, K10, K15, and K20 iterations. Accuracy values vary between 65%-71% for SVM, 32%-41% for ANN, 47%-55% for LSTM, 69%-72% for ELM, 69%-73.5% for MLELM, 77%-78.2% for KELM and for the proposed model MKELM 84.72% at all K-fold levels.

MKELM model are precise, reflecting its stable performance across different K-Fold iterations. These results underscore the robustness and superior performance of the MKELM model, particularly in the context of cross-validation with varying K-Fold iterations. The zero standard deviation for MKELM indicates consistent performance, unaffected by changes in the K-Fold values, further highlighting its reliability and effectiveness. The results of our statistical analysis, summarized in Table 6, demonstrate that the MKELM model significantly outperforms other models in terms of mean accuracy. The paired t-tests and confidence intervals confirm that the differences in performance are statistically significant. MKELM achieved a mean accuracy of 84.72% with a standard deviation of 0.00, indicating exceptional stability across different K-Fold values. In contrast, models like SVM and ANN showed lower mean accuracies and higher standard deviations, highlighting their less consistent performance.

The confusion matrix for the proposed system is generated when employing both traditional multi-classification and weighted scheme-based multi-classification with the MKELM model. Fig. 8 illustrates the confusion matrix obtained from multi-classification using our proposed weighted scheme with the MKELM model.

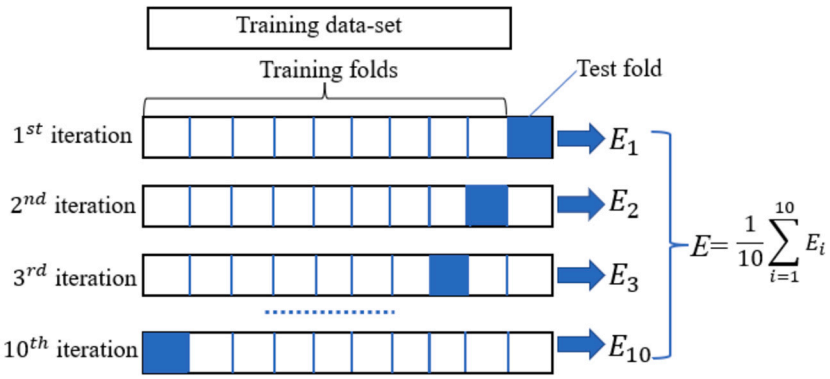

Fig. 7. KFold-based cross-validation structural diagram. To evaluate the performance of all models including MKELM, 80% of the data is considered for training and 20% for testing.

Table 6
Statistical Analysis of Model Accuracies (K = 0, 5, 10, 15, 20). Paired t-tests compare each model's accuracy with the MKELM model's accuracy. The p-values indicate the statistical significance of the differences.

Models	Mean Accuracy (%)	Standard Deviation	95% Confidence Interval	Paired t-test (vs. MKELM)	p-value (vs. MKELM)
SVM	68.00	2.10	[66.83, 69.17]	23.96	< 0.001
ANN	36.50	3.10	[34.52, 38.48]	45.14	< 0.001
LSTM	51.00	2.90	[49.17, 52.83]	36.79	< 0.001
ELM	70.50	1.50	[69.60, 71.40]	24.64	< 0.001
ML-ELM	71.75	2.10	[70.57, 72.93]	25.34	< 0.001
KELM	77.60	0.80	77.17, 78.03]	31.77	< 0.001
MKELM	84.72	0.00	[84.72, 84.72]	-	-

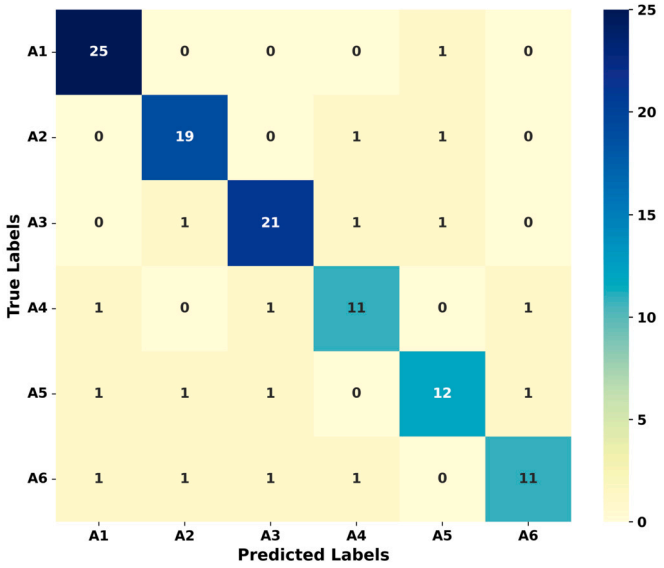

Fig. 8. Confusion matrix based on the frequency of predicted classes/accents vs frequency of actual classes by applying the MKELM model for Multi-classification (weighted Scheme).

4.2. Model's performance evaluation based on computation time

A comparison was conducted between the proposed MKELM model and other state-of-the-art models to evaluate their training times. Two techniques were employed: traditional multi-classification and weighted scheme-based multi-classification. The evaluation, using the hardware specifications revealed the following training times for each model under the traditional multi-classification approach: SVM (740 s), ANN (1540 s), LSTM (1620 s), ELM (500 s), ML-ELM (510 s), KELM (495 s), and the proposed MKELM model (490 s).

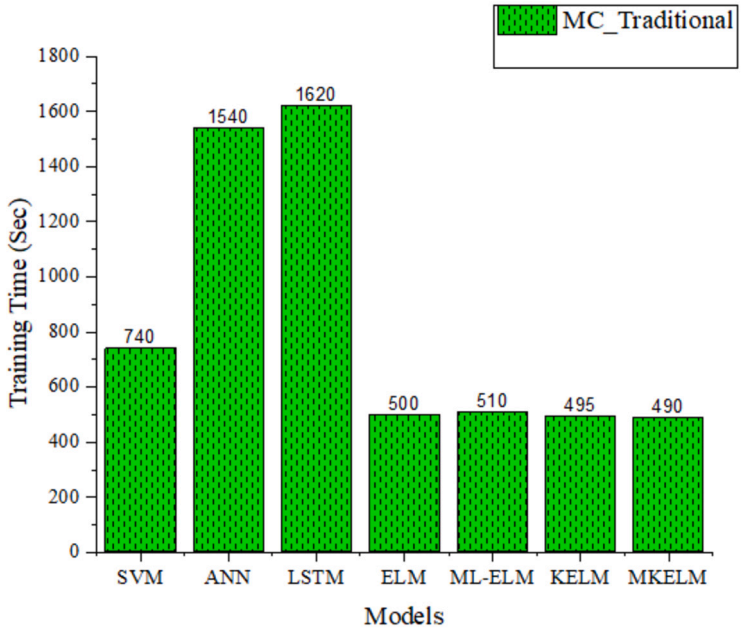

Fig. 9. Training time (in sec) for each model using traditional multi-classification, in which the proposed model MKELM has the shortest training time (490 sec).

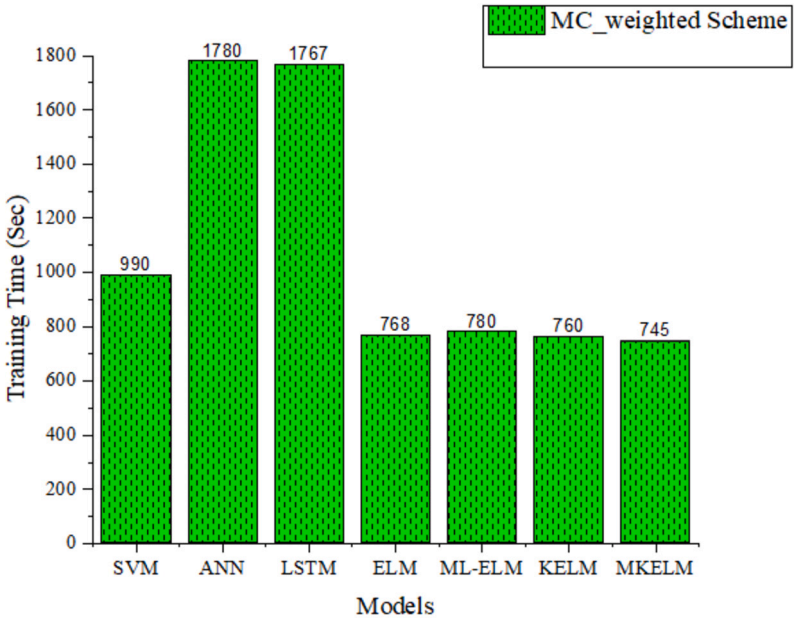

Fig. 10. Training time (sec) of each model using the pairwise weighted scheme of multi-classification. The proposed model, MKELM, has the lowest training time (745 sec).

In contrast, when the weighted scheme-based multi-classification technique was employed, the training times for SVM, ANN, LSTM, ELM, ML-ELM, KELM, and the proposed MKELM model were 990 s, 1780 s, 1767 s, 768 s, 780 s, 760 s, and 745 s, respectively. These findings demonstrate that the proposed MKELM model is computationally more efficient compared to the other models, as evidenced by the reduced training times. A graphical representation of these results can be found in Figs. 9 and 10.

4.3. Comparison with previous results

We compared the performance of our proposed model with other studies on foreign accent classification that utilized the same data-set (Accent George Mason University (GMU) data-set). The results of previous studies are shown in Table 7. In [31] Bryant et al. achieved 42% accuracy on the GMU data-set by proposing a Gaussian Discriminant Analysis (GDA) and a Naive Bayes model

Table 7
A comparison of different foreign accent identification (FAID) studies that use the same data set from GMU. The table shows the advantage of the proposed model MKELM that results in an 84.72% accuracy for 6 accents.

Study on Foreign Accents Identification	Data-set	Models	Accents Identified	Accuracy achieved (%)
M Bryant et al. [31]	Speech Accent Archive (GMU)	GDA and Naive Bayes	5	42.00
DS Widyowaty et al. [32]	Speech Accent Archive (GMU)	CNN	5	51.96
Y Singh et al. [33]	Speech Accent Archive (GMU)	CNN	5	53.92
DS Widyowaty et al. [34]	Speech Accent Archive (GMU)	KNN	6	57.00
A Ensslin et al. [35]	Speech Accent Archive (GMU)	CNN	3	61.00
P Parikh et al. [36]	Speech Accent Archive (GMU)	CNN, DNN, RNN	3	68.67
e P Berjon et al. [37]	Speech Accent Archive (GMU)	2-layer CNN	5	70.65
V Mikhailava et al. [57]	Speech Accent Archive (GMU)	CNN	9	98.70
S Deshpande et al. [58]	Speech Accent Archive (GMU)	GMM	2	85.00
Y Singh et al. [33]	Speech Accent Archive (GMU)	CNN	5	53.92
Duduka et al. [59]	Speech Accent Archive (GMU)	CNN	3	62.81
Y Jiao et al. [25]	Speech Accent Archive (GMU)	DNN RNN	11	46.66
P Parikh et al. [36]	Speech Accent Archive (GMU)	DNN RNN CNN	3	68.67
A Ahmed et al. [60]	Speech Accent Archive (GMU)	CNN	3	70.33
Current study	**Speech Accent Archive (GMU)**	**MKELM (Proposed)**	**6**	**84.72**

and identified 5 foreign accents. In [32] Widyowaty et al. and [33] Y. Singh et al. proposed a CNN-based approach for 5 foreign accents in the data set and achieved an overall accuracy of 51.96% and 53.92%, respectively. In [34] Widyowaty et al. proposed a KNN model and 6 different accents were identified with an accuracy of 57%. In [35] Ensslin et al. used a DNN and identified 3 accents with a 61% accuracy. In [36] Parikh et al. use an approach based on a CNN and LSTM to identify 3 accents with an accuracy of 68.67%. In [37] Berjon et al. used a 2-layer based CNN model to identify 5 accents with a 70.65% accuracy. In [27] Upadhyay et al. proposed a DBN model to identify 6 accents achieving an accuracy of 71.09%. In [31], Bryant et al. achieved an accuracy of 42% on the GMU dataset by proposing a Gaussian Discriminant Analysis (GDA) and Naive Bayes model, successfully identifying 5 foreign accents. Widyowaty et al. [32] and Singh et al. [33] utilized a CNN-based approach for 5 foreign accents in the dataset, obtaining overall accuracies of 51.96% and 53.92%, respectively. In [34], Widyowaty et al. proposed a KNN model that identified 6 different accents with an accuracy of 57%. Ensslin et al. [35] employed a DNN to identify 3 accents with an accuracy of 61%. Parikh et al. [36] utilized a CNN and LSTM approach to identify 3 accents, achieving an accuracy of 68.67%. Berjon et al. [37] proposed a 2-layer CNN model that successfully identified 5 accents with an accuracy of 70.65%. Upadhyay et al. [27] proposed a DBN model to identify 6 accents, achieving an accuracy of 71.09%. S. Deshpande et al. [58] proposed a Gaussian Mixture Model (GMM) that successfully identified two accents with an accuracy of 85.00%. Y. Singh et al. [33] utilized a Convolutional Neural Network (CNN) to identify five accents, achieving an accuracy of 53.92%. S. Duduka et al. [59] also employed a CNN, identifying three accents with an accuracy of 62.81%. Y. Jiao et al. [25] adopted a combined Deep Neural Network (DNN) and Recurrent Neural Network (RNN) approach, identifying eleven accents with an accuracy of 46.66%. P. Parikh et al. [36] proposed a method combining DNN, RNN, and CNN to identify three accents, achieving an accuracy of 68.67%. Lastly, A. Ahmed et al. [60] utilized a CNN model to identify three accents with an accuracy of 70.33%. In comparing our “Multi-Kernel Extreme Learning Machine (MKELM)” model with a CNN-based approach for foreign accent identification it is observed that the CNN-based model with the Mel-spectrogram features selection approach achieves higher reported accuracy [57]. However, a nuanced analysis reveals several considerations. The CNN model effectively utilizes amplitude Mel-spectrograms, emphasizing the significance of diverse features. MKELM, relying on MFCCs and prosodic features, offers a distinctive feature set, particularly highlighting the potential enhancement through further exploration and innovation in leveraging prosodic features. CNN’s use of the Speech Accent Archive dataset contributes to its success, particularly for specific European accents. In contrast, a deeper exploration of the diversity and representativeness of MKELM’s dataset may uncover opportunities for improvement or specialization. MKELM demonstrates reduced computational complexity, faster learning rates, and shorter training times, suggesting practical advantages for real-world applications. Its efficiency and scalability are critical for resource-constrained environments. While CNN excels in recognizing specific accents, MKELM’s paired weighting scheme addresses class imbalances, indicating the potential for improved generalization across a broader range of accents. The prospect of hybrid models, combining the strengths of both MKELM and CNN architectures, emerges as a promising avenue for future research. This approach could lead to a more robust and accurate foreign accent identification system. Evaluation in real-world scenarios, consideration of additional metrics beyond accuracy, and continuous refinement based on evolving datasets are crucial aspects of both models’ practical applicability and adaptability. In the final discussion, despite CNN’s higher reported accuracy, MKELM presents distinctive advantages, particularly with its emphasis on prosodic features, efficiency, potential for generalization, and adaptability. The pursuit of hybrid approaches and continuous refinement positions MKELM as a promising candidate for advancing the field of foreign accent identification.

In our study, we employed the MKELM model with the pairwise weighted scheme for multi-classification, achieving better performance on 6 different accents with an accuracy of 84.72%.

4.4. Research limitations

This study facing some limitations as well. The GMU dataset may not fully represent the global diversity of accents, indicating the need for larger and more varied datasets. The features used, primarily MFCCs and prosodic features, could be expanded with

additional feature sets, such as deep learning-based embeddings, mel spectrograms based on word and phonemes level to potentially enhance performance. The controlled environment of pre-recorded samples does not account for real-world noise and variability, requiring increased robustness and noise-handling capabilities. The model's accuracy varies across different accents, with some (e.g., L1 French and L1 Spanish) showing lower accuracy, indicating the need for further refinement in feature extraction and model training. Addressing these limitations can improve the robustness, accuracy, and applicability of the MKELM model.

5. Conclusions

In recent years, there has been a growing interest in Foreign Accent Identification (FAID) within the speech community. Previous studies in this field employed various techniques, including statistical and probabilistic methods with Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). However, more recent research has shifted towards utilizing Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), particularly Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs) for FAID and Speaker Accent Recognition (DID) tasks.

In this paper, we propose a novel approach for FAID using the Multiple Kernel Extreme Learning Machine (MKELM) and a pair-wise weighted scheme for multi-classification of foreign accents. The proposed model combines Mel-Frequency Cepstral Coefficients (MFCCs) and Prosodic features as input. To evaluate the effectiveness of our model, we compared it with state-of-the-art methods, including SVM, ANN, LSTM, ELM, MLELM, and KELM. Experimental analyses demonstrated that the MKELM model with the weighted classifier effectively distinguishes between different foreign accents. It outperformed all other tested models in terms of accuracy and computational complexity.

However, it is important to acknowledge that variations in speaker characteristics can influence the formation of accents. For a given set of speakers and/or foreign accents, achieving perfect foreign accent classification performance might not be feasible due to inconsistencies in accent-sensitive features within speech signals from certain speakers.

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 [61].

CRediT authorship contribution statement

Kaleem Kashif: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Abeer Alwan:** Writing – review & editing, Investigation, Conceptualization. **Yizhi Wu:** Methodology, Formal analysis, Conceptualization. **Luca De Nardis:** Conceptualization. **Maria-Gabriella Di Benedetto:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kaleem Kashif reports was provided by University of Rome La Sapienza. Kaleem Kashif reports a relationship with University of Rome La Sapienza that includes: employment. Kaleem Kashif has patent pending to No. Sapienza University Rome. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data are available from corresponding author upon reasonable request.

References

[1] J.H. Hansen, L.M. Arslan, Foreign Accent Classification Using Source Generator Based Prosodic Features, 1995 International Conference on Acoustics, Speech, and Signal Processing, vol. 1, IEEE, 1995, pp. 836–839.
[2] V. Gupta, P. Mermelstein, Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer, J. Acoust. Soc. Am. 71 (6) (1982) 1581–1587.
[3] S. Goronzy, S. Rapp, R. Kompe, Generating non-native pronunciation variants for lexicon adaptation, Speech Commun. 42 (1) (2004) 109–123.
[4] L.M. Arslan, J.H. Hansen, Language accent classification in American English, Speech Commun. 18 (4) (1996) 353–367.
[5] P. Angkititrakul, J.H. Hansen, Advances in phone-based modeling for automatic accent classification, IEEE Trans. Audio Speech Lang. Process. 14 (2) (2006) 634–646.
[6] H. Behravan, V. Hautamäki, T. Kinnunen, Factors affecting I-vector based foreign accent recognition: a case study in spoken Finnish, Speech Commun. 66 (2015) 118–129.
[7] C. Woehrling, P.B.d. Mareüil, Identification of regional accents in French: perception and categorization, in: Ninth International Conference on Spoken Language Processing, 2006.
[8] A. Leemann, Comparative Analysis of Voice Fundamental Frequency Behavior of Four Swiss German Dialects, Selbstverl., 2009.
[9] C.G. Clopper, D.B. Pisoni, K. De Jong, Acoustic characteristics of the vowel systems of six regional varieties of American English, J. Acoust. Soc. Am. 118 (3) (2005) 1661–1676.
[10] L.M. Hyman, In defense of prosodic typology: a response to Beckman and Venditti, Linguist. Typol. 16 (3) (2012) 341–385.

[11] H.N. Andreassen, I. Racine, Schwa et variation inter-régionale: une analyse de trois points d'enquête suisses, in: Journées PFC 2013 «Regards Croisés sur les Corpus Oraux, 2013.
[12] P. Rickard, R. Lodge Anthony, French: From Dialect to Standard, Routledge, London and New York, 1993, x+285 pp., 0 415 08071 1 J. Fr. Lang. Stud. 3 (2) (1993) 243–244.
[13] J. Nerbonne, Linguistic variation and computation (invited talk), in: 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003.
[14] K. Kashif, Y. Wu, A. Michael, Consonant phoneme based extreme learning machine (elm) recognition model for foreign accent identification, in: Proceedings of the 2019 the World Symposium on Software Engineering, 2019, pp. 68–72.
[15] H. You, A. Alwan, A. Kazemzadeh, S. Narayanan, Pronunciation variations of Spanish-accented English spoken by young children, in: Ninth European Conference on Speech Communication and Technology, 2005.
[16] J.E. Flege, C. Schirru, I.R. MacKay, Interaction between the native and second language phonetic subsystems, Speech Commun. 40 (4) (2003) 467–491.
[17] L.W. Kat, P. Fung, Fast accent identification and accented speech recognition, in: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, in: Proceedings. ICASSP99 (Cat. No. 99CH36258), vol. 1, IEEE, 1999, pp. 221–224.
[18] K. Kumpf, R.W. King, Automatic accent classification of foreign accented Australian English speech, in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, vol. 3, IEEE, 1996, pp. 1740–1743.
[19] K. Phapatanaburi, L. Wang, R. Sakagami, Z. Zhang, X. Li, M. Iwahashi, Distant-talking accent recognition by combining gmm and dnn, Multimed. Tools Appl. 75 (9) (2016) 5109–5124.
[20] D. Fohr, I. Illina, Text-independent foreign accent classification using statistical methods, in: 2007 IEEE International Conference on Signal Processing and Communications, IEEE, 2007, pp. 812–815.
[21] G. Choueiter, G. Zweig, P. Nguyen, An empirical study of automatic accent classification, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 4265–4268.
[22] M.H. Bahari, R. Saeidi, D. Van Leeuwen, et al., Accent recognition using I-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7344–7348.
[23] H. Behravan, V. Hautamäki, S.M. Siniscalchi, T. Kinnunen, C.-H. Lee, i-vector modeling of speech attributes for automatic foreign accent recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (1) (2016) 29–41, https://doi.org/10.1109/TASLP.2015.2489558.
[24] L.M.A. Sheng, M.W.X. Edmund, Deep learning approach to accent classification, in: CS229, 2017.
[25] Y. Jiao, M. Tu, V. Berisha, J.M. Liss, Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features, in: Interspeech, 2016, pp. 2388–2392.
[26] A. Purwar, H. Sharma, Y. Sharma, H. Gupta, A. Kaur, Accent classification using machine learning and deep learning models, in: 2022 1st International Conference on Informatics (ICI), IEEE, 2022, pp. 13–18.
[27] R. Upadhyay, S. Lui, Foreign English accent classification using deep belief networks, in: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), IEEE, 2018, pp. 290–293.
[28] T. Chen, C. Huang, E. Chang, J. Wang, Automatic accent identification using Gaussian mixture models, in: IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01, IEEE, 2001, pp. 343–346.
[29] M. Rizwan, B.O. Odelowo, D.V. Anderson, Word based dialect classification using extreme learning machines, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, 2016, pp. 2625–2629.
[30] F. Weninger, Y. Sun, J. Park, D. Willett, P. Zhan, Deep learning based mandarin accent identification for accent robust asr, in: INTERSPEECH, 2019, pp. 510–514.
[31] M. Bryant, A. Chow, S. Li, Classification of accents of English speakers by native language, 2014.
[32] D.S. Widyowaty, A. Sunyoto, H.A. Fatta, Accent recognition using mel-frequency cepstral coefficients and convolutional neural network, in: Proceedings of the International Conference on Innovation in Science and Technology (ICIST 2020), Atlantis Press, 2021, pp. 43–46.
[33] Y. Singh, A. Pillay, E. Jembere, Features of speech audio for accent recognition, in: 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), IEEE, 2020, pp. 1–6.
[34] D.S. Widyowaty, A. Sunyoto, Accent recognition by native language using mel-frequency cepstral coefficient and k-nearest neighbor, in: 2020 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE, 2020, pp. 314–318.
[35] A. Ensslin, T. Goorimoorthee, S. Carleton, V. Bulitko, S.P. Hernandez, Deep learning for speech accent detection in videogames, in: Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference, 2017.
[36] P. Parikh, K. Velhal, S. Potdar, A. Sikligar, R. Karani, English language accent classification and conversion using machine learning, in: Proceedings of the International Conference on Innovative Computing & Communications (ICICC), 2020.
[37] P. Berjon, A. Nag, S. Dev, Analysis of French phonetic idiosyncrasies for accent recognition, Soft Comput. Lett. 3 (2021) 100018.
[38] H. Behravan, V. Hautamäki, S.M. Siniscalchi, T. Kinnunen, C.-H. Lee, I-vector modeling of speech attributes for automatic foreign accent recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (1) (2015) 29–41.
[39] A. Khosravani, P.N. Garner, A. Lazaridis, Modeling dialectal variation for Swiss German automatic speech recognition, in: Interspeech, 2021, pp. 2896–2900.
[40] E. Alsharhan, A. Ramsay, Robust automatic accent identification based on the acoustic evidence, Int. J. Speech Technol. 26 (3) (2023) 665–680.
[41] J. Padmanabhan, M.J. Johnson Premkumar, Machine learning in automatic speech recognition: a survey, IETE Tech. Rev. 32 (4) (2015) 240–251.
[42] A. Tomar, Various classifiers based on their accuracy for age estimation through facial features, Int. Res. J. Eng. Technol. 3 (07) (2016).
[43] K. Aida-zade, A. Xocayev, S. Rustamov, Speech recognition using support vector machines, in: 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2016, pp. 1–4.
[44] J. Chorowski, J. Wang, J.M. Zurada, Review and performance comparison of svm- and elm-based classifiers, Neurocomputing 128 (2014) 507–516.
[45] X. Li, W. Mao, W. Jiang, Multiple-kernel-learning-based extreme learning machine for classification design, Neural Comput. Appl. 27 (2016) 175–184.
[46] Y. Zhang, Y. Wang, G. Zhou, J. Jin, B. Wang, X. Wang, A. Cichocki, Multi-kernel extreme learning machine for eeg classification in brain-computer interfaces, Expert Syst. Appl. 96 (2018) 302–310.
[47] H. Zhao, S. Guo, Uncertain interval forecasting for combined electricity-heat-cooling-gas loads in the integrated energy system based on multi-task learning and multi-kernel extreme learning machine, Mathematics 9 (14) (2021) 1645.
[48] B. Ahuja, V.P. Vishwakarma, Deterministic multi-kernel based extreme learning machine for pattern classification, Expert Syst. Appl. 183 (2021) 115308.
[49] G. Forman, et al., An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res. 3 (Mar 2003) 1289–1305.
[50] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, vol. 207, Springer, 2008.
[51] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 42 (2) (2011) 513–529.
[52] M. Girolami, Mercer kernel-based clustering in feature space, IEEE Trans. Neural Netw. 13 (3) (2002) 780–784.
[53] S. Weinberger, Speech Accent Archive, George Mason University, 2015, Online, http://accent.gmu.edu.
[54] J. Lyons, D.Y.-B. Wang Gianluca, H. Shteingart, E. Mavrinac, Y. Gaurkar, W. Watcharawisetkul, S. Birch, L. Zhihe, J. Hölzl, J. Lesinskis, H. Almér, C. Lord, A. Stark, jameslyons/python_speech_features: release v0.6.1, https://doi.org/10.5281/zenodo.3607820, Jan. 2020.
[55] N. Dehak, P. Dumouchel, P. Kenny, Modeling prosodic features with joint factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 15 (7) (2007) 2095–2103.

[56] L. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process. 25 (1) (1977) 24–33.
[57] V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, E. Pyshkin, Language accent detection with cnn using sparse data from a crowd-sourced speech archive, Mathematics 10 (16) (2022) 2913.
[58] S. Deshpande, S. Chikkerur, V. Govindaraju, Accent classification in speech, in: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), IEEE, 2005, pp. 139–143.
[59] S. Duduka, H. Jain, V. Jain, H. Prabhu, P.M. Chawan, A neural network approach to accent classification, Int. Res. J. Eng. Technol. 8 (03) (2021) 1175–1177.
[60] A. Ahmed, P. Tangri, A. Panda, D. Ramani, S. Karmakar, Vfnet: a convolutional architecture for accent classification, in: 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1–4.
[61] A. Johnson, K. Everson, V. Ravi, A. Gladney, M. Ostendorf, A. Alwan, Automatic dialect density estimation for African American English, in: Proc. Interspeech 2022, 2022, pp. 1283–1287.