

Uncovering Linguistic Patterns: Exploring Ensemble Learning and Low-Level Features for Identifying Spoken Arabic, English, Spanish, and German

Skander HAMDI

Dept. of Computer science
University of Ferhat Abbas, Setif 1
Setif, Algeria
skander.hamdi@univ-setif.dz

Abdelouahab MOUSSAOUI

Dept. of Computer science
University of Ferhat Abbas, Setif 1
Setif, Algeria
abdelouahab.moussaoui@univ-setif.dz

Mafaza CHABANE

Dept. of Computer science
University of Ferhat Abbas, Setif 1
Setif, Algeria
mafaza.chabane@univ-setif.dz

Ayoub LAOUAREM

Dept. of Computer science
University of Ferhat Abbas, Setif 1
Setif, Algeria
ayoub.laouarem@univ-setif.dz

Mohamed BERRIMI

Dept. of Computer science
University of Ferhat Abbas, Setif 1
Setif, Algeria
mohamed.berrimi@univ-setif.dz

Mourad OUSSALAH

Dept. of Computer Science
University of Oulu
Oulu, Finland
mourad.oussalah@oulu.fi

Abstract—This paper presents a novel approach to spoken language identification in Arabic, English, Spanish, and German languages using ensemble learning techniques. The study compares the performance of two well-known ensemble learning algorithms (Random Forest and XGBoost). Next, a Stacking Ensemble method with Logistic Regression is used as a meta-model, which combines the predictions of two configurations of Random Forest and two other configurations of XGBoost. We explore the utilization of the audio Low-Level Descriptors features, which have previously received limited attention in spoken language identification. Experimental results demonstrate the effectiveness of the ensemble learning algorithms, achieving high accuracy in accurately classifying spoken languages. Notably, the Stacking Ensemble method showcases its ability to reduce misclassification rate, emphasizing its potential for performance improvement. The stacking Ensemble method achieved the highest recorded classification rate of 97.22%, outperforming individual methods.

Index Terms—Audio analysis, Spoken Language Identification, Low-Level Descriptors, Stacking Ensemble Learning, Low-cost Machine Learning

I. INTRODUCTION

Language IDentification (LID) plays a crucial role in various speech and audio-related tasks, ranging from automatic speech recognition to multilingual translation systems. With the increasing availability of multilingual audio data [1]– [3], [7], the accurate and efficient identification of spoken languages has become an essential research area. Machine Learning (ML) and Deep Learning (DL) techniques have demonstrated remarkable success in tackling language audio-related models, offering promising performance and versatility [4]–[6]. In this work, we focus on the task of identifying spoken Arabic, along with English, Spanish, and German

from audio recording perspective. While there have been notable advancements in language identification research, few works have included Arabic as one of the target languages. Therefore, our work aims to address this gap and present an approach that incorporates Arabic, along with English, Spanish, and German languages. LID has diverse applications, including real-time language identification for fast speech translation and multilingual content indexing. However, fast and accurate detection of multiple languages poses challenges due to the complexity and variations in speech signals. To tackle this challenge, we propose to compare the performance of several ML and ensemble learning classifiers, and then, for further performance improvement, a meta-learner model has been trained to aggregate the best accurate base models by stacking them into one single classifier, by considering audio Low-Level Descriptors (LLD) as input features. Stacking Ensemble approach leverages the strengths of multiple base models to obtain an aggregated prediction with the aim of surpassing the performance of individual models Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). Our methodology is designed in the following. First, it creates a low-cost ML model by training several individual and accurate classifiers that maximize performance while minimizing computational complexity and resource requirements. Second, it combines the strengths of these classifiers using a meta-learner to enhance the overall performance. Intuitively, this enables our low-cost approach to be deployed on resource-constrained devices. With lightweight ML models, our system further enables real-time language detection on small devices, making it practical for applications such as mobile devices and embedded systems. This opens up new possibilities for on-the-go LID, facilitating seamless integration into various domains where fast and reliable LID is crucial. In this paper, we present

the experimental results of our proposed approaches on a combined and balanced dataset of spoken Arabic, English, Spanish, and German recordings. The paper is organized as follows: Section II provides an overview of related works in language identification and ensemble learning, emphasizing the limited inclusion of Arabic in previous studies. Section III describes the combined dataset used for our experiments and the data engineering steps employed. Section IV details the proposed methodology which will investigate the details and configuration of our base classifiers as well as the proposed stacking ensemble method. Section V presents the experimental results and performance evaluation of the approaches. Section VI discusses the achieved performance and provides an ablation study to highlight the advantages and weaknesses of the proposed study, and Section VII concludes the research.

II. RELATED WORKS

In the related works section, we provide an overview of existing studies in the field of spoken language identification using ML and give a particular attention to the works that included Arabic language. Gris and Candido [8] proposed a methodology for Automatic Spoken Language Identification and related tasks using Convolutional Neural Networks (CNNs) from spectrograms in Portuguese, Spanish, and English. A corpus is compiled mainly from audiobook recordings extracted from various freely available resources. The model uses spectrograms as features and several CNN architectures were evaluated on a corpus composed of five-second audios. An overall accuracy of 83% is achieved on unseen test data. Using the same methodology, Shrawgi et al. [9] used CNN and spectrograms to distinguish five languages, Deutsche, Dutch, English, French, and Portuguese, where recordings were extracted from VoxForge dataset. By comparing the proposed approach with classical ML and other existing DL approaches, it achieved the highest accuracy of 91.5%. Another study from Mukherjee et al. [10] aimed to classify seven Indian languages. Their model used line spectral frequency-based features which were extracted from the speech signals, and forwarded to a random forest classifier for training. They reported an overall accuracy of 99.71%, surpassing Support Vector Machine (SVM) and k-Nearest Neighbors (kNN). Using the same dataset for Indian languages and VoxForge for other languages, Garain et al. [11] employed Mel-frequency cepstral coefficients (MFCCs), Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Flatness, Spectral Roll-off, Poly Features, and Tonnetz as features in their FuzzyGCP architecture which consists of a fuzzy ensemble of Deep Neural Network (DNN), Deep CNN (DCNN), and Semi-Supervised Generative Adversarial Network (SSGAN). They compared the performance of their model, along with other classifiers such as SVM and Linear Discriminant Analysis (LDA). FuzzyGCP outperformed these classifiers, achieving an accuracy of 95%. Mukherjee et al. [12] used line spectral pair-grade (LSP-G) features to classify Tamil, Telugu, Malayalam, and Kannada. They extracted linear predictive coefficients to generate the LSPs, and computed band-wise grades from these to serve

as features. Their fuzzy classification approach achieved an accuracy of 96.46% on over 12000 clips. Extending this work to perform a seven-classes classification, another work has been proposed [13], using a CNN architecture that consists of two convolution layers followed each by a max pooling layer to classify spectrogram images. Several experiments have been performed using different frame sizes for Short-Time Fourier Transform (STFT). Although the results were quite close to each others, a frame of size 256 was reported to achieve the highest performance score. The work has been validated using different sources of noise to the audio signal. The overall performance achieved 99.96% with a slight decrease in the presence of aircraft cabin noise. By exploring the Mel Frequency Cepstral Coefficients (MFCCs) and CNN, Arla et al. [14] presented a CNN model to identify four Indian languages: Bengali, Gujarati, Tamil, and Telugu. The model operates on MFCC spectrogram images generated from short splits of audio data. The training was performed on a dataset of 5 hours per language, which yielded an accuracy of 88.82%, outperforming other ML models. In the context of low-resource Russian languages, Bedyakin et al. [15] explored 23 language classes of Siberian languages using a CNN model with a Self-Attentive Pooling layer. They achieved an accuracy of 80.31% on the validation set. Baba et al. [16] used Log-Mel spectrogram and CNN to build a model for English, French, German, and Spanish, achieving an accuracy of 93.4%. Regarding studies that included Arabic language, Alashban et al. [17] proposed an approach to classify seven languages: Arabic, German, English, Spanish, French, Russian, and Chinese, derived from Mozilla Common Voice (MCV) by taking a subset of 2000 recordings for each language. R2020b-MATLAB has been used to select speech boundaries and remove silent parts to prevent learning from silence. MFCCs and Gammatone Cepstral Coefficients (GTCCs) were used as key features of a convolutional recurrent neural network (CRNN) that has been used to extract both spatial and sequential patterns. An overall accuracy of 92.81% was achieved, 95.78% for Arabic language, 94.47% for Chinese, 92.35% for English, 92.49% for French, 95.25% for Russian, 93.49% for Spanish and, finally, 80.95% for German language. Kepecs and Beigi [18] proposed to train and evaluate a Time-Delay Neural Network (TDNN) in four languages: Arabic, Spanish, French, and Turkish from MediaSpeech dataset which consists of approximately 10 hours of speech with matched transcriptions. High-resolution MFCCs were extracted and iVectors were generated to capture speaker-specific information. iVectors, along with MFCC features, were used to train the TDNN classifier. The detection rate of French was very low with only 43.75% for standard French and 25.10% for African-accented French whereas Arabic, Turkish, and Spanish have recorded 99.08%, 67.39%, and 99.66%, respectively. Verma and Buduru [19] conducted a fine-grained LID using a multilingual CapsNet model, where a dataset of 10 spoken languages was created, however, only five have been used for classification; Arabic, Bengali, Chinese (Mandarin), English, and Hindi. For the experiments, a total of 500 hours of audio were used, with 70

hours per language for training the CapsNet model for the five languages and 20 hours for non-class detection. Spectrograms were computed and forwarded to a CapsNet, achieving an overall accuracy performance of 88.76%. Regarding Arabic detection performance, the proposed model achieved 89% for both precision and recall. Heracleous et al. [20] presented a LID system using DL and the i-vector paradigm. The effectiveness of deep neural networks (DNN) and CNN was studied and compared, and also, the integration into a complete system was investigated. The experiments were conducted on the NIST 2015 i-vector Machine Learning Challenge task for recognizing 50 in-set languages. The performance achieved by DNN was 3.55% of equal error rate (EER). The latter decreases to 3.48% when a CNN model was employed, and further decreased to 3.3% when a fusion model between DNN and CNN was employed.

III. DATASET AND FEATURE ENGINEERING

In this section, we describe the construction of our custom dataset and the feature engineering process, which consists of extracting meaningful representations from audio recordings. To conduct our language identification experiments, we created a custom dataset by curating samples from existing resources. We selected 2000 samples from English, German, and Spanish languages, which were obtained from the Spoken Language Identification dataset [22]. Additionally, we included 1914 samples from the Arabic Speech Corpus [21]. All samples for the three languages were randomly chosen from the original Spoken Language Identification as it includes more than 70.000 samples. This random selection ensured a diverse representation of the respective languages in our custom dataset. Table I presents an overview of the dataset, including the total duration in hours and the count of recordings for each language class.

	English	German	Spanish	Arabic
Total Duration (hours)	5.56	5.56	5.56	4.11
Recordings Count	2000	2000	2000	1914

TABLE I
OVERVIEW OF THE COMPILED DATASET

Furthermore, Figure 1 showcases the audio waveforms of representative samples from each language class. These audio waveforms provide a visual representation of the speech signals and serve as a preliminary illustration of the dataset’s diversity.

With the established dataset, the focus shifted to feature engineering. It is worth noting that while LLDs have exhibited promising performance in various audio classification tasks in different disciplines [23]–[26], their application in language detection remains unexplored in the literature. LLDs encompass a collection of features employed in audio signal processing. These features encapsulate fundamental characteristics of the audio signal, rendering them valuable and

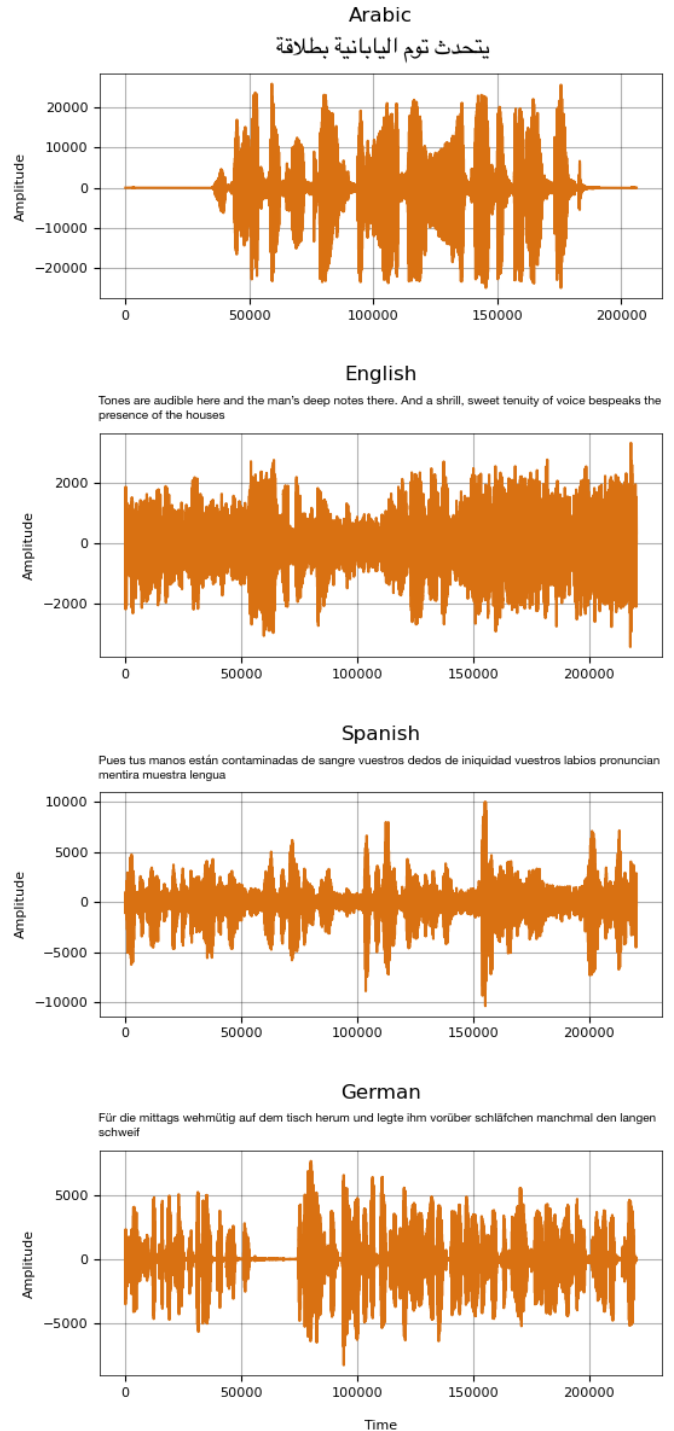


Fig. 1. Audio waveforms of representative samples from each language class with the corresponding transcript.

applicable in various domains, including speech recognition and general audio classification tasks. Put simply, LLDs represent a set of features extracted during the preprocessing phase to facilitate subsequent audio or speech analysis tasks, encompassing attributes like spectral roll-off, loudness, and temporal patterns. In this study, we used a sample rate of 22kHz. Furthermore, OpenSMILE toolkit was employed to extract LLDs. Especially, the popular *ComParE_2016 feature set*, derived from the Interspeech 2016 Computational Paralinguistics Challenge [27], was extracted. Unlike its smaller variants, such as GeMAPSv01a, GeMAPSv01b, eGeMAPSv01a, eGeMAPSv01b, and eGeMAPSv02, which offer 18 to 88 features, ComParE_2016 comprises a staggering 6373 low-level features. These features are organized into three primary groups: Energy-related LLD (4), Spectral-related LLD (55), and Voicing-related LLD (6). Additionally, ComParE_2016 feature set incorporates statistical functions that were applied to the LLD, Δ -LLD (velocity), and $\Delta\Delta$ -LLD (acceleration) features. These statistical functions encompass a range of metrics, including mean, standard deviation, linear regression slope, quadratic error, quadratic regression, and percentiles from 1% to 99% (as well as 6% and 94% percentiles). By combining these statistical functions, the ComParE_2016 feature set aims to uncover hidden patterns and explore the statistical distribution of the features. Table II shows a summary of the extracted features for each feature category.

Group	Features
Energy LLD	<ul style="list-style-type: none"> - Sum of auditory spectrum (loudness) - Sum of RASTA-style filtered auditory spectrum - Zero-Crossing Rate - RMS Energy - F0 (SHS and viterbi smoothing)
Spectral LLD	<ul style="list-style-type: none"> - RASTA-style auditory spectrum, bands 1–26 (0–8 kHz) - Spectral energy 250–650 Hz, 1 k–4 kHz - Spectral roll-off point 0.25, 0.50, 0.75, 0.90 - Spectral flux, centroid, entropy, slope - Psychoacoustic sharpness, harmonicity - Spectral variance, skewness, kurtosis
Voicing LLD	<ul style="list-style-type: none"> - Probability of voicing - Log. HNR, Jitter (local, delta), Shimmer (local)

TABLE II
SUMMARY OF COMPARE_2016 FEATURE SET [27]

IV. PROPOSED METHODOLOGY

Ensemble methods have gained prominence in machine learning for their ability to improve model performance and generalize well by combining multiple models. In audio-related tasks, such as COVID-19 diagnosis from cough sound [23], [24], ensemble methods have shown high performance as classifiers, as well as gender identification from Arabic speech [25] where Random Forest played a crucial role in the selection of the relevant features based on the importance. This is due to the complex nature of audio data, which contains diverse patterns and structures. In the proposed methodology, we aim to use LLD with ensemble methods,

specifically comparing the performance of Random Forest and XGBoost algorithms in addressing the spoken language detection problem, and then, the aim will be to combine the strengths of these algorithms to enhance the performance and improve the generalization capabilities by ensembling different configurations using stacking ensemble approach.

A. Random Forest classifier

Random Forest utilizes the concept of bagging [28], involving training multiple decision trees on different subsets of the training data, sampled with replacement. Each tree is trained independently, and the final prediction is obtained by aggregating the predictions from all the trees. A new prediction using Random Forest is defined by:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x_{new}) \quad (1)$$

where \hat{y} is the global prediction, $f_t(x)$ is the prediction from tree f_t , and T is the total number of trees in the forest.

B. XGBoost classifier

XGBoost combines the power of boosting, which sequentially improves the model by adding weak learners, with gradient descent optimization [29]. It constructs an ensemble of decision trees, where each subsequent tree corrects the errors made by the previous ones. Similar to Random Forest, a new prediction is computed. Updating the predicted value by gradient descent is generally, defined by the following equation:

$$\hat{y}_i^{(n+1)} = \hat{y}_i^{(n)} - \alpha \cdot \sum_{t=1}^T f_t(x_{new}) \quad (2)$$

where $\hat{y}_i^{(n+1)}$ is the updated predicted value for the i -th sample at iteration $(n+1)$, $\hat{y}_i^{(n)}$ is the current predicted value at iteration n , α is the learning rate that controls the step size of the update, f_t is the prediction of the t -th decision tree and T is the total number of decision trees or what we called estimators.

C. Proposed approach: Stacking Ensemble

The Stacking Ensemble (denoted SE) aims to leverage the strengths of multiple classifiers for enhanced performance in language detection. The ensemble consists of four classifiers, including two different configurations of Random Forest (RF1 and RF2) and two different configurations of XGBoost (XGB1 and XGB2). Each configuration is designed with a set of specific parameters to capture different aspects of the data. *RF1* is configured with bootstrap enabled, utilizing 200 decision trees and considering 100 randomly selected features. *RF2*, on the other hand, disables bootstrap while keeping the same number of decision trees and features. This variation allows the ensemble to explore the impact of bootstrap sampling on classification accuracy. In the case of XGBoost, *XGB1* employs 800 boosting iterations, while *XGB2* increases the number of iterations to 1200. The boosting iterations enable the model to iteratively improve its performance by fitting new trees to the residuals of the previous trees. With the

aim of combining the predictions of these individual models, we used a meta-learner, which is Logistic Regression. This choice is motivated by its effectiveness in capturing non-linear relationships between the base models' predictions and the target variable. SE approach computes a new prediction by utilizing the predictions of the base models as input to the meta-learner. We denote the predictions from RF1, RF2, XGBoost1, and XGBoost2 as P_1 , P_2 , P_3 , and P_4 , respectively. SE combines these predictions by fitting them as inputs to the meta-learner, which estimates the final prediction. A simple formulation of the meta-learner involves estimating the probabilities of class membership using a logistic function. We can denote the meta-learner's coefficients as β_0 , β_1 , β_2 , β_3 , and β_4 . Given the predictions P_1 , P_2 , P_3 , and P_4 , the meta-learner computes the logit Z of class membership as follows:

$$Z_i = \beta_0 + \beta_1 P_{1i} + \beta_2 P_{2i} + \beta_3 P_{3i} + \beta_4 P_{4i} \quad (3)$$

where Z_i presents a logit for the instance i , the logistic function is then applied to the logit Z_i to obtain the predicted probability:

$$p_i = 1/(1 + \exp(-Z_i)) \quad (4)$$

The logistic function or Sigmoid should produce a probability p_i of belonging to the class 0 or 1 using Equation 5. We shall note that we are dealing with a multi-class classification, thus, the particular case of *One vs. All* will be considered.

$$\hat{y}_i = \begin{cases} 0 & \text{if } p_i < 0.5 \\ 1 & \text{if } p_i \geq 0.5 \end{cases} \quad (5)$$

By training the Logistic Regression as a meta-learner using labeled data and optimizing the coefficients, SE can effectively combine the predictions of the base models to produce a final prediction with improved accuracy and generalizability. Each of the proposed methods, $RF1$, $RF2$, $XGB1$, $XGB2$, and SE approach is evaluated using four metrics: *accuracy*, *precision*, *sensitivity*, and *specificity*. Figure 2 presents a block diagram, summarizing the proposed approaches.

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results obtained from the evaluation of our proposed approaches. To ensure reliable and robust performance assessment, we employed 5-Fold cross-validation, which provides a comprehensive validation strategy. We report the aforementioned performance metrics for each of the four individual configurations $RF1$, $RF2$, $XGB1$, $XGB2$, and SE approach. The results highlight the effectiveness of ensemble algorithms in language detection, as well as the potential benefits of stacking ensemble methods. Table III presents the performance achieved by the two configurations of Random Forest algorithm; $RF1$, $RF2$ in terms of accuracy, precision, sensitivity, and specificity.

Furthermore, we present the classification report for only $RF2$ in Table IV, as it shows better performance than $RF1$. Also, Figure 3 shows the confusion matrix of $RF2$.

	Accuracy [%]	Precision [%]	Sensitivity [%]	Specificity [%]
$RF1$	93.60	93.96	93.75	97.86
$RF2$	94.61	94.88	94.73	98.29

TABLE III
RESULTS OF RANDOM FOREST CONFIGURATIONS ON THE TESTING DATA (15% OF THE WHOLE DATASET)

	Precision [%]	Sensitivity [%]	Support
Arabic	100	100	270
English	95.00	90.00	301
Spanish	94.00	94.00	298
German	90.00	95.00	319

TABLE IV
CLASSIFICATION REPORT FOR $RF2$ CLASSIFIER

Regarding XGBoost configurations, Table V highlights the performance achieved by $XGB1$ and $XGB2$, while Table VI presents the classification report of $XGB2$ that shows performance across the four classes as well as confusion matrix which is illustrated in Figure 4.

	Accuracy [%]	Precision [%]	Sensitivity [%]	Specificity [%]
$XGB1$	95.45	95.62	95.57	98.46
$XGB2$	96.21	96.36	96.31	98.72

TABLE V
RESULTS OF XGBOOST CONFIGURATIONS ON THE TESTING DATA (15% OF THE WHOLE DATASET)

	Precision [%]	Sensitivity [%]	Support
Arabic	100	100	270
English	97.00	93.00	301
Spanish	95.00	96.00	298
German	94.00	96.00	319

TABLE VI
CLASSIFICATION REPORT FOR $XGB2$ CLASSIFIER

Finally, we present the results of the proposed SE method, Table VII shows the achieved performance in terms of the studied evaluation metrics, Table VIII highlights the performance and each of the four classes, where the confusion matrix is illustrated in Figure 5.

VI. DISCUSSION

The experimental results show promising performance for all models, including $RF2$, $XGB2$, and the SE . Overall, the models achieved high accuracy, precision, sensitivity, and specificity, indicating their ability to classify the audio samples accurately across multiple languages (Arabic, English,

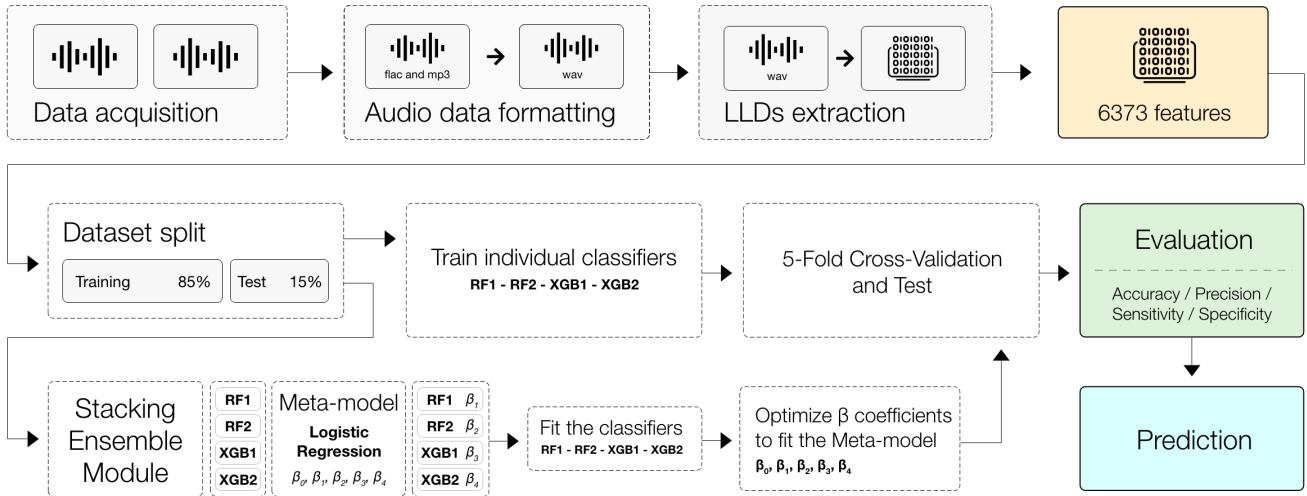


Fig. 2. Systematic Overview of the proposed approaches.

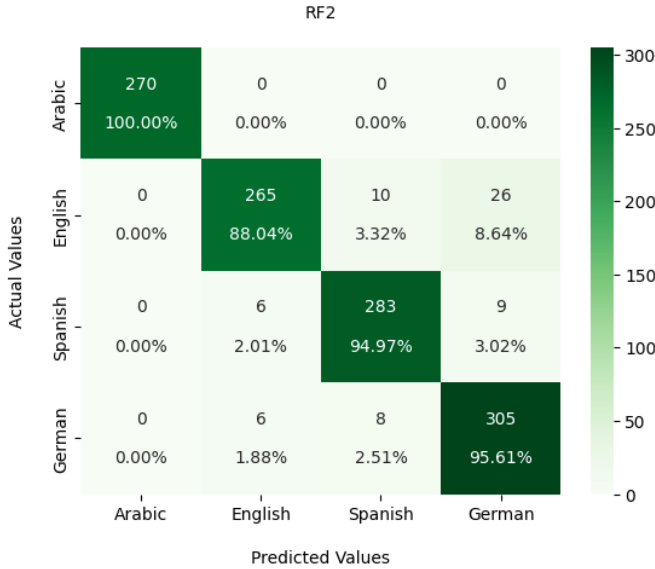


Fig. 3. Confusion matrix of the testing set using the model $RF2$.

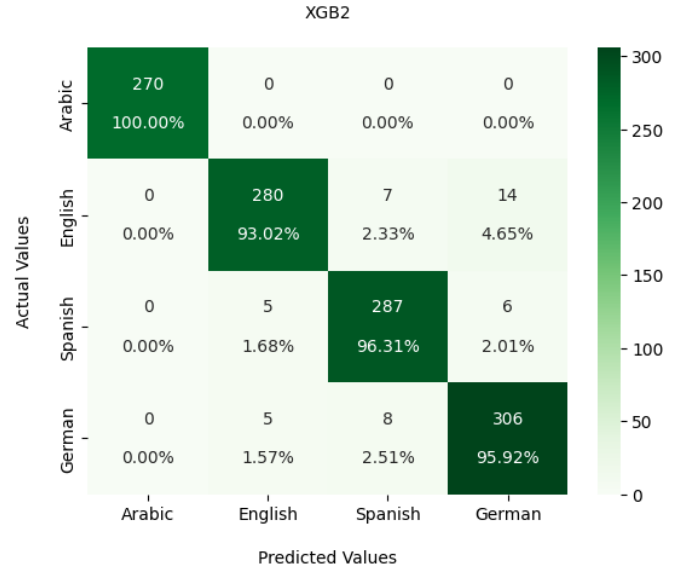


Fig. 4. Confusion matrix of the testing set using the model $XGB2$.

	Accuracy [%]	Precision [%]	Sensitivity [%]	Specificity [%]
SE	97.22	97.33	97.29	99.06

TABLE VII
RESULTS OF STACKING ENSEMBLE APPROACH ON THE TESTING DATA

	Precision [%]	Sensitivity [%]	Support
Arabic	100	100	270
English	97.00	95.00	301
Spanish	97.00	97.00	298
German	95.00	97.00	319

TABLE VIII
CLASSIFICATION REPORT FOR SE CLASSIFIER

Spanish, and German). Starting with $RF2$, it achieved an accuracy of 94.61% and demonstrated high precision values for each class, ranging from 90.00% to 100%. The sensitivity performance was also impressive, with scores ranging from 90.00% to 95.00%, indicating the model's ability to correctly identify positive instances for each class. Additionally,

the specificity scores were consistently high, ranging from 97.86% to 98.29%, indicating a low rate of false positives. Moving to $XGB2$, it outperformed $RF2$ with an accuracy

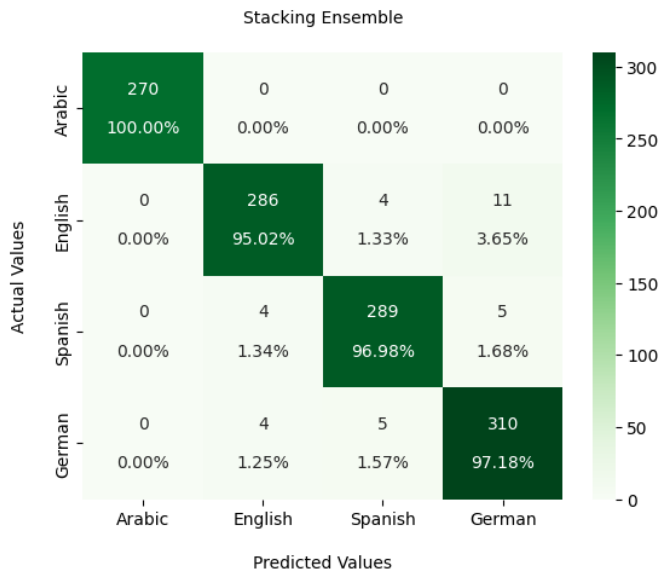


Fig. 5. Confusion matrix of the testing set using the model *SE*.

of 96.21%. *XGB2* demonstrated higher precision, sensitivity, and specificity values for each class, ranging from 94.00% to 100%, compared to *RF2*. *SE* approach achieved the highest accuracy of 97.22%, indicating its effectiveness in combining the predictions of *RF1*, *RF2*, *XGB1* and *XGB2* to make more accurate classifications. The precision values for each class in the *SE* are consistently high, ranging from 94.00% to 100%, indicating its ability to minimize misclassifications. The sensitivity values are also high, ranging from 93.00% to 96.00%, demonstrating the model's ability to correctly identify positive instances. Furthermore, it is important to consider the difficulties associated with each language and highlight the success of the proposed model in recognizing all Arabic language samples. Arabic, being a complex language with unique linguistic characteristics, presents challenges in automated language recognition. It is characterized by complex phonology, morphology, and orthography, which can make it more challenging to accurately classify. However, by analyzing *SE*'s confusion matrix (Figure 5), we observe that 270 Arabic samples have been correctly classified the proposed model demonstrates exceptional performance in correctly identifying all Arabic language samples, indicating its effectiveness in capturing the distinctive features of the Arabic language. Regarding misclassification, 1.33% and 3.65% of English samples are misclassified in Spanish and German, respectively, achieving the lowest classification rate among the four classes. However, if we observe the other confusion matrices, we can note that almost 12% and 7% of English samples are misclassified in Figure 3 and Figure 4, respectively, which has been enhanced by our proposed approach. We performed feature importance analysis using our stacking ensemble model and ranked the features accordingly. To visualize the discriminative power of the top six features, we plotted a Boxplot for each feature and for each class in Figure 6. The Boxplot

clearly showed that Arabic language samples have distinct feature distributions compared to other classes, confirming the separability achieved by our model.

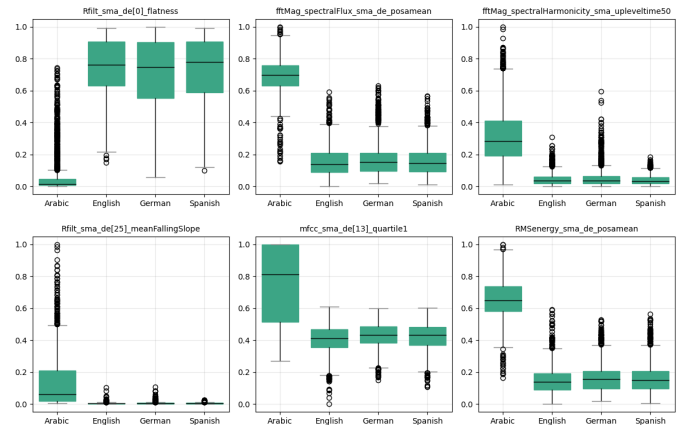


Fig. 6. Boxplot showing the distribution of the top six features for each class, highlighting the distinctiveness of Arabic language samples.

Moreover, to provide insights into the calibration performance of the models, each pair of classes' calibration curve is plotted in Figure 7, where we observe the absence of any curve below the ideal line which indicates that the model is generally well-calibrated and does not exhibit a systematic bias towards overconfidence or underconfidence. Regarding curves that present Arabic language, being significantly above the ideal line indicates that the model tends to be overconfident when predicting Arabic samples. This means that the models assign higher probabilities to Arabic samples than their true likelihood, which is due to the presence of distinct patterns or characteristics in Arabic language data that make it easier for the models to identify (see Figure 6). On the other hand, for the remaining curves that represent other languages, being slightly above the ideal line with a small deviation indicates that the models are able to provide reliable probability estimates for these classes, with only a slight bias towards higher probabilities.

VII. CONCLUSION

In summary, this work proposed a comprehensive approach to spoken language identification by employing two configurations of Random Forest and two configurations of XGBoost, followed by a Stacking Ensemble method with Logistic Regression as the Meta-model. The use of LLDs features, which had not been explored in the present context, contributed to the field of spoken language identification, specifically, classifying Arabic, English, Spanish and German. The results demonstrated the effectiveness of ensemble learning algorithms, such as Random Forest and XGBoost, in accurately classifying spoken languages. Although the enhancement achieved by the Stacking Ensemble method was not substantial, it showcased its potential in reducing misclassification rates and improving overall performance. These findings highlight the promise

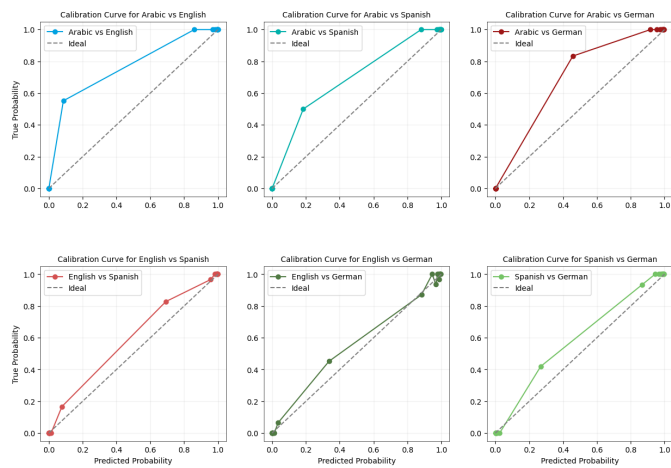


Fig. 7. Calibration curves showing confidence calibration of *SE* model.

of stacking ensembling as a viable approach for enhancing spoken language identification systems.

REFERENCES

- [1] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert. 'MLS: A Large-Scale Multilingual Dataset for Speech Research'. ArXiv, abs/2012.03411. 2020.
- [2] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audiobooks, 2015. <https://doi.org/10.1109/ICASSP.2015.7178964>.
- [3] N. Halabi, M. Wald: Phonetic inventory for an Arabic speech corpus, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).
- [4] C. Bartz, T. Herold, H. Yang, C. Meinel, Language Identification Using Deep Convolutional Recurrent Neural Networks, 2017. https://doi.org/10.1007/978-3-319-70136-3_93.
- [5] X. Chang, W. Zhang, Y. Qian, J. Le Roux, S. Watanabe, End-To-End Multi-Speaker Speech Recognition With Transformer, 2020. <https://doi.org/10.1109/ICASSP40776.2020.9054029>.
- [6] W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in: 2016 IEEE Int. Conf. Acoust. Speech Signal Process., 2016: pp. 4960–4964. <https://doi.org/10.1109/ICASSP.2016.7472621>.
- [7] J. Valk, T. Alumae, VOXLINGUA107: A Dataset for Spoken Language Recognition, 2021. <https://doi.org/10.1109/SLT48900.2021.9383459>.
- [8] L. Gris and A. Candido Junior. "Automatic Spoken Language Identification using Convolutional Neural Networks", in Proceedings of the XVII Latin American Congress of Free Software and Open Technologies , Online, 2020, pp. 16-20, doi: <https://doi.org/10.5753/latinoware.2020.18603>
- [9] H. Shrawgi, D.S. Sisodia, P. Gupta, Automated Spoken Language Identification Using Convolutional Neural Networks & Spectrograms BT - Key Digital Trends Shaping the Future of Information and Management Science, in: L. Garg, D.S. Sisodia, N. Kesswani, J.G. Vella, I. Brigui, S. Misra, D. Singh (Eds.), Springer International Publishing, Cham, 2023: pp. 152–163.
- [10] H. Mukherjee, S. Das, A. Dhar, S.M. Obaidullah, K.C. Santosh, S. Phadikar, K. Roy, An Ensemble Learning-Based Language Identification System BT - Computational Advancement in Communication Circuits and Systems, in: K. Maharatna, M.R. Kanjilal, S.C. Konar, S. Nandi, K. Das (Eds.), Springer Singapore, Singapore, 2020: pp. 129–138.
- [11] A. Garain, P.K. Singh, R. Sarkar, FuzzyGCP: A deep learning architecture for automatic spoken language identification from speech signals, Expert Syst. Appl. 168 (2021) 114416. <https://doi.org/https://doi.org/10.1016/j.eswa.2020.114416>.
- [12] H. Mukherjee, S. M. Obaidullah, S. Phadikar and K. Roy, "A Dravidian Language Identification System," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018, pp. 2654-2657, doi: 10.1109/ICPR.2018.8545406.
- [13] H. Mukherjee, S. Ghosh, S. Sen, O. Sk Md, K.C. Santosh, S. Phadikar, K. Roy, Deep learning for spoken language identification: Can we visualize speech signal patterns?, Neural Comput. Appl. 31 (2019) 8483–8501. <https://doi.org/10.1007/s00521-019-04468-3>.
- [14] L. R. Arla, S. Bonthu and A. Dayal, "Multiclass Spoken Language Identification for Indian Languages using Deep Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2020, pp. 42-45, doi: 10.1109/IBSSC51096.2020.9332161.
- [15] R. Bedyakin, N. Mikhaylovskiy, Low-Resource Spoken Language Identification Using Self-Attentive Pooling and Deep 1D Time-Channel Separable Convolutions, 2021.
- [16] M. Baba, N. Hoshikawa, H. Nakayama, T. Ito, A. Shiraki, Development of Spoken Language Identification System Using Directional Volumetric Display, Proc. Int. Disp. Work. (2020) 519. <https://doi.org/10.36463/idw.2020.0519>.
- [17] A.A. Alashban, M.A. Qamhan, A.H. Meftah, Y.A. Alotaibi, Spoken Language Identification System Using Convolutional Recurrent Neural Network, Appl. Sci. 12 (2022). <https://doi.org/10.3390/app12189181>.
- [18] B. Kepecs, H. Beigi, Automatic Spoken Language Identification using a Time-Delay Neural Network, 2022. <https://doi.org/10.13140/RG.2.2.21631.89763>.
- [19] M. Verma and A. B. Buduru, "Fine-grained Language Identification with Multilingual CapsNet Model," 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, India, 2020, pp. 94-102, doi: 10.1109/BigMM50055.2020.00023.
- [20] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad and A. Yoneyama, "Comparative Study on Spoken Language Identification Based on Deep Learning," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 2265-2269, doi: 10.23919/EUSIPCO.2018.8553347.
- [21] N. Halabi, M. Wald, Phonetic inventory for an Arabic speech corpus, 2016.
- [22] Spoken Language Identification. Kaggle, 2021, <https://www.kaggle.com/toponowicz/spoken-language-identification>.
- [23] S. Hamdi, A. Moussaoui, M. Oussalah, M. Saidi, Early COVID-19 Diagnosis from Cough Sound Using Random Forest and Low-Level Descriptors, in: Third Int. Conf. Comput. Inf. Sci. 2021, 2021: pp. 1–6.
- [24] S. Hamdi, A. Moussaoui, M. Oussalah, M. Saidi, Autoencoders and Ensemble-Based Solution for COVID-19 Diagnosis from Cough Sound BT - Modelling and Implementation of Complex Systems, in: S. Chikhi, G. Diaz-Descalzo, A. Amine, A. Chaoui, D.E. Saidouni, M.K. Kholadi (Eds.), Springer International Publishing, Cham, 2023: pp. 279–291.
- [25] H. Skander, A. Moussaoui, M. Oussalah, M. Saidi, Gender Identification from Arabic Speech Using Machine Learning, in: 2020: pp. 149–162. https://doi.org/10.1007/978-3-030-58861-8_11.
- [26] S. Klaylat, Z. Osman, L. Hamandi, R. Zantout, Emotion recognition in Arabic speech, Analog Integr. Circuits Signal Process. 96 (2018) 337–351. <https://doi.org/10.1007/s10470-018-1142-4>.
- [27] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird, A.C. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language, Interspeech. (2016).
- [28] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [29] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, 2016. <https://doi.org/10.1145/2939672.2939785>.