Cureus Journal of Computer Science

Open Access Original Article

Comparative Study of Various Machine Learning Approaches for Marathi Dialect Recognition

Ashwini G. Pawar [1] , Ajay S. Patil [1] , Nita V. Patil [1]

1. School of Computer Science, KBC North Maharashtra University, Jalgaon, IND

Corresponding author: Ashwini G. Pawar, ashwini.gulabrao.pawar@gmail.com

Received 03/18/2025
Review began 03/25/2025
Review ended 07/02/2025
Published 07/18/2025

DOI:
https://doi.org/10.7759/s44389-025-03648-8

Abstract

Identifying dialects accurately in the digital age to preserve linguistic diversity has become increasingly important. This research focuses on developing a strong Marathi dialect identification system that uses sound data and artificial intelligence techniques. Various acoustic features such as spectral centroid, bandwidth, roll-off, root mean square energy, zero-crossing rate, and mel-frequency cepstral coefficients were extracted from audio recordings representing distinct Marathi dialects spoken in different regions. The dataset was preprocessed by dividing the recordings into frames that can be analyzed and normalizing audio formats. Several machine learning algorithms were utilized to classify the dialects, such as K-nearest neighbors, decision trees, naïve Bayes, AdaBoost, and gradient boosting. Multiple metrics were used to compare the performance of the algorithms, which resulted in the best output at 94.56% accuracy using the gradient boosting algorithm. Precision, recall, and F1-score were also used to evaluate the model's performance, where gradient boosting outperformed other methods in all cases. This work demonstrates the potential of advanced machine learning techniques in dialect recognition, especially for resource-scarce languages like Marathi. It opens the way for further improvements in natural language processing applications.

Categories: AI applications, Machine Learning (ML), Natural Language Processing (NLP)
Keywords: language dialect identification (ldi), mel-frequency cepstral coefficients (mfcc), zero crossing rate (zcr), naïve bayes (nb), gradient boosting (gb), k-nearest neighbor (knn)

Introduction

People can distinguish between spoken languages naturally, often relying on their knowledge of other languages. Although these conclusions may not always be accurate, they demonstrate how language skills identify different linguistic groups. Automatic language identification aims to use natural language processing (NLP) to mimic this capacity. NLP allows computers to understand spoken and written human language through artificial intelligence. Dialects, social or geographical variants of a language, are characterized by vocabulary, grammar, and pronunciation differences [1]. Dialect is speech that differs from the culture's language or speech patterns. This kind may be localized. Konkanis and Vidarbhas speak Konkani and Vidarbha Marathi, respectively. Dialects of a language can be understood by each other. Dialects, which are regional speech patterns, are influenced by geography, mother tongue, surrounding languages, socioeconomic class, education, and culture. Language dialect identification (LDI) automatically identifies dialects from spoken utterances, enabling speaking-style-based search and information retrieval. Language identification distinguishes language families, while LDI distinguishes regional speech varieties within the same language. The third most native speakers in India speak Marathi after Hindi and Bengali. Of the 42 Marathi varieties spoken, 22 are officially recognized by the Indian constitution. LDI helps to preserve linguistic diversity, improve NLP applications, obtain sociolinguistic insights, improve communication, enhance education, aid forensic linguistics, and benefit commercial applications [2]. These objectives emphasize its importance in improving technological, social, and cultural knowledge and applications. Dialects are community-specific linguistic traits. Phonemes, pronunciation, loudness, nasality, and tone can identify dialects. Dialect recognition helps automatic speech recognition (ASR), e-health and remote access, e-learning for older and housebound individuals, etc. Thus, language recognition is more accessible than dialect recognition. Therefore, dialect identification is a fascinating and essential topic of study.

Lee et al. [3] presented Korean Dialect Identification (K-DID). Based on theories and prior research in Korean dialectology, this model learns the prosody (intonation patterns) of different dialects of the Korean language. This work aims to classify dialectal and nondialectal speech in a way that can identify dialectal speech patterns, even when those patterns are embedded within nondialectal intervals [3]. Patil et al. [4] concentrated on developing voice recognition for the Marathi dialect using Hidden Markov Models. Nanmalar et al. [5] took up the challenge of classifying Tamil dialects using only the acoustic characteristics of the speech. They employed mel-frequency cepstral coefficients (MFCC) and Gaussian mixture models. With this approach, they classified the Tamil dialects across various speakers without prior annotation. Their model proposed a significant and scalable way to handle Tamil dialects across tongue and text in a computational framework. Das et al. [6] developed a new method for extracting prosodic language features

How to cite this article
Pawar A G, Patil A S, Patil N V (July 18, 2025) Comparative Study of Various Machine Learning Approaches for Marathi Dialect Recognition. Cureus J Comput Sci 2 : es44389-025-03648-8. DOI https://doi.org/10.7759/s44389-025-03648-8

of Indian languages. In this paper, a unique neural network Q-learning technique for optimal feature extraction and classification of Indian languages is presented. The approach examines parametric and essential speech signal types, particularly the excitation sources. Utilizing a deep perception-based neural network effectively gathers important prosodic properties. Fourier transformations are employed to address the fundamental non-stationarity of voice signals. The usefulness and validity of the proposed model were empirically evaluated using a limited database consisting of 80 utterances, each containing eight words from a single Indian language sentence. This served as an evaluation platform for discerning the language among the eight spoken in the speech corpus [6]. Ismail and Singh [7] developed a method significant for distinguishing between two dialects of the Assamese language The two tools applied in this study, the Universal Background Model and the Gaussian Mixture Model, are conventional tools in the domain but are strikingly different; hence, applying them sequentially happens to be a fortunate selection. Chittaragi et al. [8] developed word-based language order paradigm based on the acoustic properties of speech. This method was applied for testing the Intonational Variations in English corpus that contains data of nine British English dialects. The research aimed at understanding intonational patterns concerning word formation by adopting the use of spectral and prosodic variables. XGBoost, supported with algorithms of the kind tree-based methods and support vector machines (SVM), was implemented in determining language order design, thereby making up [8] to the procedure in the language identification method formulated for seven dialects of the Indians by Madhu et al [9]. This process is a front-end phonetic engine that interprets spoken words through phonetic symbolization. Its method includes sylla counting of the organization process into banks where phones form each syllable. The relationship between unbroken segments of all syllables linked by phonotactic rules to allow for 'natural lip-transition' formations does this. A frame is two consecutive syllables, quantitatively expressed to indicate the optimal tongue position for articulating syllables, which is neither too far forward nor too far back [9]. Ankit et al. [10] studied automated speech recognition using the Sphinx engine, a research library designed for this purpose. To enhance automatic speech recognition systems, Shigli et al. [11] sought to automatically recognize the accents or dialects of South Indian English speakers. The primary signal processing method used to extract features from speech signals is MFCC. Models developed by Chaudhari et al. [12] detect emphasis in Hindi and highlight certain features using MFCC, linear predictive coding, and acoustic elements. Liu and Hansen [13] described the terms "emphasize" and "vernacular" as distinctive examples of elocution and jargon used by local speakers in a specific geographic area. A feature selection technique was created to enhance the Prosodic Language Identification system's performance, according to the evaluation of Ng et al. [14]. Using this technique, they could identify prosody-related elements and provide creative front-end and back-end solutions for language identification problems. Etman and Beex [15] utilized feature vectors from the acoustic segment model in the back-end based on data and the co-occurrences of acoustic segment model. They implemented a unique vector space modeling strategy to differentiate between spoken languages.

The literature survey reveals that there has been extensive research on dialect detection using machine learning for various languages, including Korean, Tamil, Assamese, and British English. However, there is a notable gap in the application of these techniques to Marathi dialects. While Patil et al. [4] developed voice recognition for Marathi using Hidden Markov Models, a comprehensive study comparing different machine learning approaches for Marathi dialect detection is lacking. The survey carried by us highlights various techniques such as prosodic feature analysis, acoustic characteristics, and phonotactic elements used in other languages. These methods, including MFCC, Gaussian mixture models, neural networks, and SVM, have shown promising results in dialect identification. So, given the linguistic diversity of the Marathi language and the success of machine learning in other language dialects, there is a clear need for a comparative study applying these advanced techniques to Marathi dialect detection.

This approach identifies Marathi dialects using phonemes, spectral features, and prosodic aspects like speaker intonation and tonality. These are the major components of the natural language processing system being developed to classify and organize Marathi dialects. Processing and assessing the system's primary components uses machine learning and feature extraction methods. This research presents a novel approach to automatically identify Marathi dialects using key acoustic features extracted from audio recordings. The primary contribution of this work includes:

- Acoustic Feature Extraction: The study incorporates advanced signal processing techniques to extract essential acoustic features such as spectral centroid, bandwidth, roll-off, root mean square energy (RMSE), zero-crossing rate, and MFCC. These features are critical for capturing the unique acoustic characteristics of different Marathi dialects.

- Machine Learning Algorithms: Multiple machine learning classifiers, including K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Trees (DT), AdaBoost, and Gradient Boosting (GB), were utilized for dialect classification. A comprehensive evaluation was performed to determine their effectiveness in dialect identification, with GB emerging as the top-performing model.

- Performance Evaluation: All proposed models are thoroughly tested regarding precision, recall, F1-score, and accuracy. Here, the GB model scored the highest when all algorithms are considered, providing a high value of 94.56%, and thereby showing machine learning's excellence in dialect recognition.

- Impact on Sociolinguistics: This research further contributes to NLP and sociolinguistics by overcoming the dialect identification problem in low-resource languages like Marathi. The findings are quite insightful for various machine learning techniques to differentiate regional dialects and serve as a basis for further research.

Materials And Methods

The detailed flow and dataset preparation for the suggested system are presented in this section. Figure *1* displays the proposed system's block diagram.

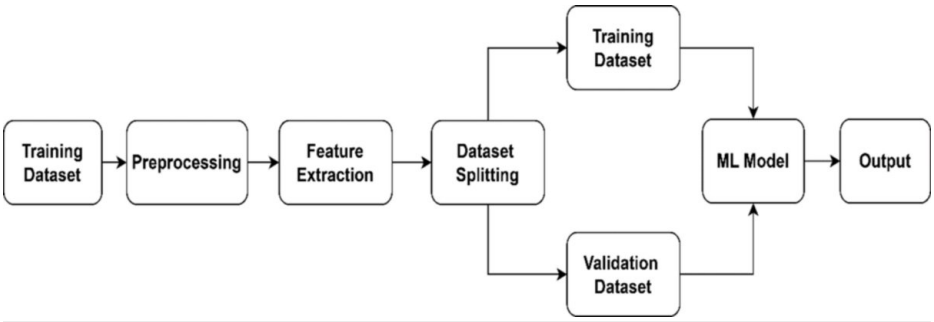

FIGURE 1: Block diagram of the proposed system

Dataset

Audio samples representing the linguistic diversity within Marathi language dialects were first gathered during the research phase. Audio data was collected from voluntary participants who are native speakers of the target dialects, ensuring a diverse representation by age and gender. Each sample was recorded in a noise-minimized setting at a uniform sampling rate and then segmented for further analysis. The samples were gathered from different individuals residing in various regions, representing four different dialects: Ahirani, Khandeshi Marathi, Standard Marathi, and Varhadi. The rich dataset serves as a significant part of the dialect detection system, ensuring that the system adequately captures the full range of linguistic variations found in Marathi. The dataset serves as a solid foundation for the training of dialect identification machine learning models. Dialects belonging to different geographic regions are part of the dataset. It was divided into 10-fold cross-validation, ensuring the proper development and robust evaluation of the model at multiple iterations. A fixed random seed was set for all data splitting and cross-validation procedures to ensure the results are consistent and reproducible across runs. Table *1* shows the sample distribution across the four dialects with the total number of samples. To ensure the authenticity of dialect labeling, participant origin and language background were verified before data collection. A linguistics expert reviewed a subset of the recordings to confirm dialect classification accuracy.

Dialects	No. of total samples
Ahirani	305
Khandeshi Marathi	159
Standard Marathi	251
Varhadi	112

TABLE 1: Dataset distribution

This process for gathering and prepping the data helps ensure the collected dataset will be diverse and well-formed to allow dialect recognition systems to gain accurate performance within the system through training and validation.

Preprocessing

In developing machine learning models, preprocessing is a critical process. Audio-based tasks, such as dialect recognition, rely heavily on successful preprocessing. This research's preprocessing stage went through several processes to prepare the raw audio data for feature extraction and subsequent model training. Ultimately, it aimed to standardize the audio inputs, eliminate noise or irrelevant information, and

maintain uniformity across the dataset.

These are considered the baseline procedures to make accurate model predictions and avoid overfitting or underfitting at the Open Set training time. After obtaining the dataset, each audio sample is then preprocessed. This includes breaking every audio recording into standardized clips of 20 seconds. This ensures consistency in the data length, which is very useful later on for the evaluation and feature extraction. The standardization of the clip duration is important in order to maintain uniformity over the dataset and prepare the dataset for further processing. This study was composed of several critical steps, in order to standardize the audio recordings before feature extraction.

- Segmentation: Each audio recording was segmented into 20-second clips to maintain uniformity and facilitate consistent analysis.

- Normalization: The audio signals were normalized to ensure uniform volume levels, reducing the impact of variations in recording conditions.

- Noise Reduction: Background noise was filtered out to enhance the clarity of the speech signals, improving the accuracy of the feature extraction process.

- Format Standardization: All audio files were converted to a standard format (.wav), including uniform sampling rates (11,000 Hz) and bit depths to ensure consistency across the dataset.

- Silence Removal: The recordings removed Silent segments to focus on meaningful speech data, enhancing the training process. Silence segments were detected and removed based on short-term energy thresholds using Librosa's *'effects.trim'* function, thereby focusing only on active speech for further processing.

These preprocessing steps ensured that the dataset was clean, consistent, and ready for effective feature extraction and machine-learning model development.

Feature extraction

Feature extraction essentially extracts meaningful information from raw data, which helps in analysis and interpretation for use in decision-making in different sectors. Its relevance in pattern recognition and data analysis is fundamental for machines to understand what to process, which can explain the different forms of Marathi dialects accordingly. Various features were extracted from the preprocessed audio clips describing the different types of Marathi dialects appropriately. These features include chromatogram, spectral features, RMSE, zero crossing rate, MFCC, chroma, and tonal features. Each feature captures different aspects of the audio signal, thereby comprehensively representing the linguistic characteristics inherent in the various dialects.

Feature	Description	No. of sub-features
Chromagram	Captures harmonic content of the signal by computing chroma features using Short-Time Fourier Transform [16].	1
Spectral Centroid	Represents the center of mass of the frequencies in the signal, giving insight into the "brightness" of the sound.	1
Spectral Bandwidth	Measures the spread of frequencies around the spectral centroid, indicating the signal's timbral width.	1
Spectral Rolloff	The frequency at which a defined percentage (e.g., 85%) of the total spectral energy is encompassed.	1
Root Mean Square Energy (RMSE)	Measures the signal's energy and relates to the loudness of the sound.	1
Zero Crossing Rate (ZCR)	Determines the frequency at which the signal intersects zero, signifying the audio's noisiness.	1
MFCC (13 coefficients)	MFCC encapsulates the timbre of the signal and is extensively utilized in speech and audio analysis.	13
CQT Chromagram	Another chroma feature of the Constant-Q Chromagram was capturing harmonic content using CQT.	1
Chroma Energy Normalized (CENS)	Captures harmonic content using chroma features normalized for energy, providing robustness against dynamics.	1
Spectral Contrast	Quantifies the spectrum's disparity between peaks and troughs, offering information into the sound's quality.	7
Spectral Flatness	Indicates how flat or peaky the spectrum is, helping distinguish between tonal and noise-like sounds [17].	1
Polynomial Coefficients	Represents the polynomial coefficients of the signal, capturing its overall shape and trends.	1
Tonal Centroid (Tonnetz)	Describes the tonal relations of the signal, capturing harmonic structure using a tonal centroid method.	6

TABLE 2: Features extracted for the proposed system

CQT, Constant-Q Transform; MFCC, Mel-Frequency Cepstral Coefficients

The detailed features extracted for the proposed system are tabulated in Table *2*. In this approach, the 'Libros' library extracts the different features [18]. It is a Python package specifically designed for music and audio analysis, offering a range of powerful tools for tasks such as time-series transformation, filtering, and feature extraction. The ‘Libros’ library extracts 36 various acoustic features from audio recordings in this approach.

Dataset splitting

After feature extraction is applied, there is 10-fold cross-validation across the dataset - a method that stratifies the dataset across 10 equal partitions or folds in each iteration for testing, setting aside one while using the remainder to train machine learning models. These allow the different models to correctly learn the variations in each form of the Marathi dialect by taking the majority vote of the subset in each loop. It also ensures unbiased assessment of model performance, providing a balanced framework whereby the algorithms are trained on a substantial portion of the data and evaluated on an independent subset.

Training and classification

The extracted features serve as input for training machine learning algorithms. Several classification methods are used to find patterns and relationships in the data, such as KNN, DT, NB, Adaboost, and GB algorithms. These algorithms aim to classify the audio samples into their respective dialect categories. The KNN algorithm is a non-parametric algorithm that does not assume any specific data distribution. It is a lazy learning algorithm that saves untrained instances and uses them to forecast new instances by comparing them to stored ones. KNN anticipates similar output variable values for feature values close to

the feature space. The algorithm uses a distance metric, such as Euclidean, Manhattan, Minkowski, or Hamming Distance, to measure how similar or dissimilar data points are. However, the performance of KNN is highly dependent on the quality of the chosen distance metric. Additionally, the number of neighbors (k), a crucial hyperparameter, significantly affects KNN's performance. A small k value may make the model sensitive to noise, while a considerable k value could cause it to over smooth the decision boundary, leading to underfitting [19]. DT is a popular machine learning technique frequently used for regression and classification tasks. Because they are straightforward and intuitive, they are simple to comprehend and use in the evaluation and analysis of data. DT models data using a naturally occurring tree structure. Based on the value of a feature, every internal node reflects a decision. In classification, each leaf node signifies a class label; in regression, it denotes a continuous value. Each branch signifies the outcome of that decision [20]. Probabilistic classifiers in the NB family are based on Bayes' theorem and assume feature independence. Despite its simplicity, the approach has proven effective across various applications. Fundamentally, an NB classifier classifies a feature by calculating the likelihood that it belongs to a specific class. The ensemble learning method, Adaptive Boosting, or AdaBoost, is particularly effective at enhancing the performance of simple models. When applied, AdaBoost combines several weak learners, almost learners who do not perform much better than random chance. The algorithm weights the instances in the training set in each iteration, and the weak learner makes predictions. If the learner predicts incorrectly, the algorithm increases the weight of the mispredicted instance. If the learner is correct, the algorithm keeps the example at its current weight. GB is an ensemble learning technique that builds models in succession, with each new model correcting the flaws of its predecessors. The essence of GB is to combine weak predictive models to form a "strong" overall predictor. GB has two essential components: (1) sequentially summing weak models to get an overall model and (2) constructing the weak models.

Performance evaluation

The concluding phase is assessing the efficiency of the constructed models. The efficiency of the dialect detection algorithm is evaluated using many measures, including average accuracy, weighted averages of precision, recall and F1-score. Precision is a measure of total correctness, representing the accuracy of positive predictions. The F1-score, on the other hand, strikes a balance between recall and precision. The capacity to correctly select positive examples is evaluated through recall. The system's capacity to recognize and differentiate between different Marathi dialects is thoroughly evaluated using these measures. A dialect detection system's performance is assessed using multiple critical parameters. Here is a brief explanation of each metric and its significance:

- Precision: It measures how accurate the model is in producing a positive output. The rate of precision was calculated by counting the number of true positives relative to the number of all correct predictions, where the latter contains both true and false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

- Recall: This measures the sensitivity or true positive rate, which is the percentage of all pertinent events that the model actually detects. Sometimes, it is referred to as a proportion: the ratio of actual positives to false negatives, or true positives to true negatives.

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

- F1-score: F1-score is the harmonic mean between precision and recall. In cases of imbalanced class distribution, this kind of score is particularly useful as it provides a balanced measure. Both false positives and false negatives are taken into account for evaluating the performance of the model holistically.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- Accuracy: The accuracy metric calculates the proportion of correct predictions in relation to the total number of predictions, measuring the overall correctness of the model. It is made up of both true positives and true negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- Weighted Average Precision, Recall, and F1-score: These metrics are calculated as the class-weighted average of the respective scores for each class, using weights proportional to the number of instances in a particular class. It addresses class imbalance and ensures a better, more accurate assessment of the model's performance across the dialects.

Results

This section documents our results pertaining to the task of identifying the Marathi dialect using multiple machine learning models. The performances of the models are evaluated through metrics such as average accuracy and the weighted averages for precision, recall, and F1-score corresponding to each of the four classes, providing profound insight into how well each performs. The comparison study describes both the merits and demerits of a few different models. Table *3* and Figure *2* summarize the comparison of performance across algorithms, whereas Figure *3* is a confusion matrix of the GB algorithm.

ML algorithms	Parameters	Precision	Recall	F1-Score	Accuracy (%)
KNN	k = 3	0.7655	0.7656	0.7523	76.56
	k = 5	0.7467	0.7560	0.7383	75.60
	k = 7	0.7454	0.7487	0.7315	74.87
Decision Tree	-	0.8927	0.8900	0.8890	89.00
Naive Bayes	-	0.7849	0.7812	0.7793	78.12
AdaBoost	-	0.7866	0.7665	0.7585	76.65
Gradient Boosting	-	0.9483	0.9456	0.9447	94.56
MFCC+GMM [5]	-	0.82	0.89	0.85	88
SVM and XGBoost, Ensemble [8]	-	-	-	-	SVM = 75.9, XGBoost = 80.5

TABLE 3: Comparative analysis of the performance of different ML algorithms

ML, Machine Learning; KNN, K-Nearest Neighbors; MFCC, Mel-Frequency Cepstral Coefficients; GMM, Gaussian Mixture Models; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting

Table *3* and Figure *2* present a comparative study of several machine learning algorithms applied to dialect recognition, measured in terms of four metrics: precision, recall, F1-score, and accuracy. An experiment was conducted using the KNN algorithm, where several values of k were tested: k = 3, k = 5, and k = 7. KNN performed the best with k = 3, achieving a precision of 0.7655, recall of 0.7656, F1-score of 0.7523, and accuracy of 76.56%. The performance decreased with k = 5 and k = 7, indicating that k = 3 is the most optimal parameter for this dataset. The DT algorithm emerged as one of the top performers, with a precision of 0.8927, recall of 0.8900, F1-score of 0.8890, and accuracy of 89.00%. NB showed relatively better performance compared to AdaBoost, with a precision of 0.7849, recall of 0.7812, F1-score of 0.7793, and accuracy of 78.12%, though it lagged behind the top models. AdaBoost had average performance, with a precision of 0.7866, recall of 0.7665, F1-score of 0.7585, and accuracy of 76.65%. Despite producing good results, more reliable algorithms like DT were preferred for the task. In the comparative analysis, it was observed that the proposed GB and NB approaches performed better than the existing system.

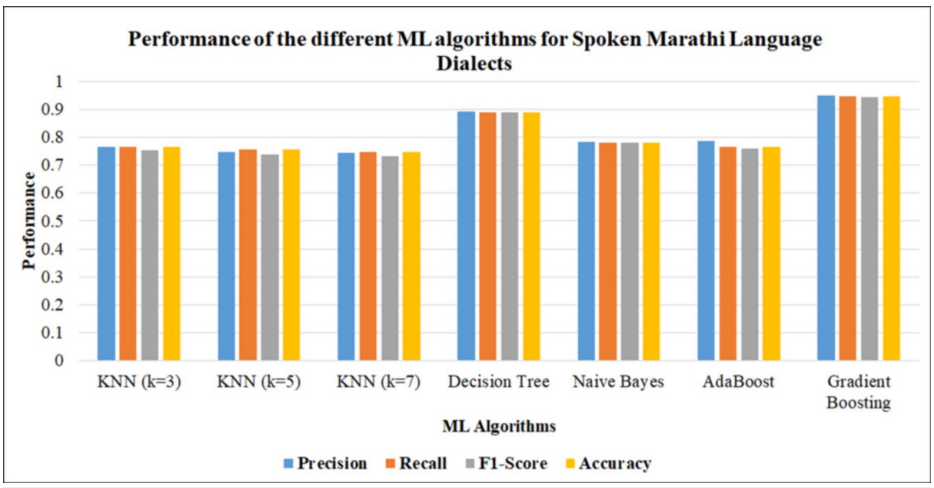

FIGURE 2: Performance of different algorithms for spoken Marathi dialect recognition

ML, Machine Learning; KNN, K-Nearest Neighbors

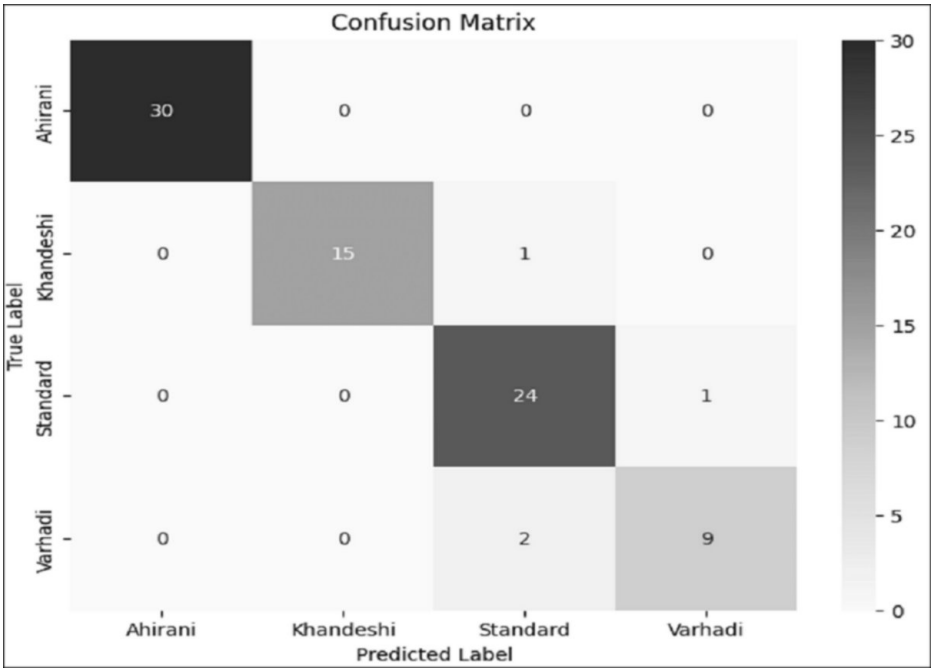

FIGURE 3: Confusion matrix of gradient boosting algorithm for spoken Marathi dialect recognition

GB emerged as the most effective algorithm, achieving the highest precision (0.9483), recall (0.9456), F1-score (0.9447), and accuracy (94.56%). These results demonstrate that GB is this study's most suitable model for dialect recognition, outperforming all other algorithms. The sample confusion matrix for the GB algorithm illustrates the model's performance in predicting the four Marathi dialects: Ahirani, Khandeshi Marathi, Standard Marathi, and Varhadi. The model achieved excellent results in recognizing Ahirani, with 30 out of 30 samples classified correctly. For Khandeshi, the model performs well by classifying 15 out of 16 samples, with only one sample incorrectly classified. In the case of Standard Marathi, the model achieved good results; 24 out of 25 samples were correctly classified, with only 1 sample incorrectly classified. Lastly, for Varhadi, the model correctly predicted 9 out of 11 samples, with two samples incorrectly classified. The model shows strong performance, with only a few misclassifications, suggesting effective handling of dialect recognition.

Discussion

The results of this study affirm the effectiveness of various machine learning models for Marathi dialect classification, where GB performed the best with a 94.56% accuracy rate. This affirms ensemble learning techniques to improve classification by progressively making more accurate predictions. DT is an instance that performs at its best accuracy of 89.00%, showing competence in understanding dialect variation in rule-based classification. The KNN algorithm implemented for various values of k tested proved to be at its best functioning for k = 3 with the correctness rate of 76.56%, showing evidence of a suitable size neighborhood being discovered within smaller sizes under these settings. NB and AdaBoost were less accurate (78.12% and 76.65%), likely due to their inability to handle complex feature relationships and weak forms, respectively. Looking at the GB confusion matrix revealed minor misclassifications between Standard Marathi and Khandeshi Marathi due to their similar phonetic structures. Also, the relatively lower accuracy in classifying the Varhadi dialect shows the need for a balanced dataset with higher samples from minority dialects. The imbalance in sample sizes, particularly the lower representation of the Varhadi dialect, is explicitly acknowledged. Weighted evaluation metrics are used to account for this during performance analysis. In future work, data augmentation methods such as pitch shifting, time-stretching, or synthetic sample generation will be explored to further balance the dataset and enhance recognition performance for underrepresented dialects.

The work emphasizes the significance of feature extraction in the classification of dialects, where MFCC, spectral features, and chroma features notably helped differentiate dialects. Although conventional machine learning models have been effective, future work should investigate deep learning methods like convolutional neural networks or recurrent neural networks to extract hierarchical features automatically. Increasing the dataset size to cover additional dialects and speakers would further enhance model generalization. The results have real-world applications in NLP tasks, such as speech-to-text translation, voice-enabled technologies, and linguistic studies. Synthesizing dialect identification with NLP systems has the potential to improve automatic speech recognition for low-resource languages such as Marathi. Other aspects, such as real-time processing and hybrid models combining conventional machine learning with deep learning to achieve higher classification accuracy and scalability, may be addressed in future work. Applying the same strategy to other Indian languages with rich dialects may contribute further to language preservation and improvements in speech-processing technologies.

Conclusions

This paper successfully analyzed audio recordings' acoustic data to check the performance of different machine learning algorithms in classification to distinguish between various Marathi dialects. In the approaches applied, the GB algorithm was found the most effective, and remarkable metrics were an accuracy of 94.56%, precision of 94.83%, recall of 94.56%, and an F1-score of 94.47%. The DT algorithm also presented robust performance as it achieved an accuracy of 89.00%. The KNN algorithm performed the best at k = 3, reaching the highest accuracy of 76.56%. NB and AdaBoost showed generally medium results, with accuracy values at 78.12% and 76.65%, respectively. These results prove that the new, sophisticated machine learning techniques, specifically GB, have much promise for dialect identification. This research draws attention to the importance of careful model selection and parameter tuning adapted to the peculiarities of the data. Future research could consider deep learning-based feature extraction, expanding the size of the database to include dialects and more speakers, as well as providing real-time processing capabilities. Another area of work would be extending these techniques into other low-resource languages, and this would lead to a significant breakthrough in natural language processing and even linguistic studies.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ashwini G. Pawar

Acquisition, analysis, or interpretation of data: Ashwini G. Pawar, Ajay S. Patil, Nita V. Patil

Drafting of the manuscript: Ashwini G. Pawar, Ajay S. Patil, Nita V. Patil

Critical review of the manuscript for important intellectual content: Ajay S. Patil, Nita V. Patil

Supervision: Ajay S. Patil, Nita V. Patil

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.
Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Banafa A: Natural language processing (NLP). Introduction to Artificial Intelligence (AI). River Publishers, New York; 2024. 15-19. 10.1201/9781003499527-3
2. Humayun MA, Yassin H, Abas PE: Native language identification for Indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions. Speech Communication. 2022, 139:92-101. 10.1016/j.specom.2022.03.007
3. Lee J, Kim K, Chung M: Korean dialect identification based on intonation modeling. 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore. 2021, 168-173. 10.1109/O-COCOSDA202152914.2021.9660537
4. Patil SB, Patil NV, Patil AS: Speaker independent isolated word recognition using HTK for Varhadi - A dialect of Marathi. International Journal of Engineering and Advanced Technology. 2020, 9:748-751. 10.35940/ijeat.b3832.029320
5. Nanmalar M, Vijayalakshmi P, Nagarajan T: Literary and colloquial dialect identification for Tamil using acoustic features. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India. 2019, 1303-1306. 10.1109/TENCON.2019.8929499
6. Das HS, Roy P: Optimal prosodic feature extraction and classification in parametric excitation source information for Indian language identification using neural network based Q-learning algorithm. International Journal of Speech Technology. 2019, 22:67-77. 10.1007/s10772-018-09582-6
7. Ismail T, Singh LJ: Dialect identification of Assamese language using spectral features. Indian Journal of Science and Technology. 2017, 10:1-7. 10.17485/ijst/2017/v10i20/115033
8. Chittaragi NB, Koolagudi SG: Acoustic features based word level dialect classification using SVM and ensemble methods. 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, India. 2017, 1-6. 10.1109/IC3.2017.8284315
9. Madhu C, George A, Mary L: Automatic language identification for seven Indian languages using higher level features. 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kollam, India. 2017, 1-6. 10.1109/SPICES.2017.8091332
10. Ankit A, Mishra SK, Shaikh R, Gupta CK, Mathur P, Pawar S, Cherukuri A: Acoustic speech recognition for Marathi language using Sphinx. ICTACT Journal on Communication Technology. 2016, 7:1361-1365.
11. Shigli A, Patel I, Rao KS: Automatic dialect and accent speech recognition of South Indian English. International Journal of Latest Trends in Engineering and Technology,. 2016, 103-111.
12. Chaudhari RH, Waghmare K, Gawali BW: Accent recognition using MFCC and LPC with acoustic features. International Journal of Innovative Research in Computer and Communication Engineering. 2015, 3:2128-2134.
13. Liu G, Hansen JHL: A systematic strategy for robust automatic dialect identification. 2011 19th European Signal Processing Conference, Barcelona, Spain. 2011, 2138-2141.
14. Ng RWM, Lee T, Leung CC, Ma B, Li H: Analysis and selection of prosodic features for language identification. 2009 International Conference on Asian Language Processing, Singapore. 2009, 123-128. 10.1109/IALP.2009.34
15. Etman A, Beex AAL: Language and dialect identification: A survey. 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK. 2015, 220-231. 10.1109/IntelliSys.2015.7361147
16. Geraei H, Rodriguez EAV, Majma E, Habibi S, Al-Ani D: A noise invariant method for bearing fault detection and diagnosis using adapted local binary pattern (ALBP) and short-time Fourier transform (STFT). IEEE Access. 2024, 12:107247-107260. 10.1109/ACCESS.2024.3438106
17. Lu W-k, Zhang Q: Deconvolutive short-time Fourier transform spectrogram. IEEE Signal Processing Letters. 2009, 16:576-579. 10.1109/LSP.2009.2020887
18. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O: librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference. 2015, 18-25. 10.25080/Majora-7b98e3ed-003
19. Manfaati IM, Abdurakhman A: Analysis of social vulnerability in Java Island using K-medoids algorithm with variation of distance measurements (Euclidean, Manhattan, Minkowski). Indonesian Journal of Artificial Intelligence and Data Mining. 2024, 7:467-467. 10.24014/ijaidm.v7i2.31111
20. Mokarram R, Emadi M: Classification in non-linear survival models using Cox regression and decision tree. Annals of Data Science. 2017, 4:329-340. 10.1007/s40745-017-0105-4