



ARTICLE



<https://doi.org/10.1057/s41599-024-03113-2>

OPEN

Multi-class identification of tonal contrasts in Chokri using supervised machine learning algorithms

Amalesh Gope¹✉, Anusuya Pal², Sekholu Tetseo¹, Tulika Gogoi¹, Joanna J¹ & Dinkur Borah¹

This study examines and explores the effectiveness of various Machine Learning Algorithms (MLAs) in identifying intricate tonal contrasts in Chokri (ISO 639-3), an under-documented and endangered Tibeto-Burman language of the Sino-Tibetan language family spoken in Nagaland, India. Seven different supervised MLAs, viz., [Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB)], and one neural network (NN)-based algorithms [Artificial Neural Network (ANN)] are implemented to explore five-way tonal contrasts in Chokri. Acoustic correlates of tonal contrasts, encompassing fundamental frequency fluctuations, viz., f0 height and f0 direction, are examined. Contrary to the prevailing notion of NN supremacy, this study underscores the impressive accuracy achieved by the RF. Additionally, it reveals that combining f0 height and directionality enhances tonal contrast recognition for female speakers, while f0 directionality alone suffices for male speakers. The findings demonstrate MLAs' potential to attain accuracy rates of 84-87% for females and 95-97% for males, showcasing their applicability in deciphering the intricate tonal systems of Chokri. The proposed methodology can be extended to predict multi-class problems in diverse fields such as image processing, speech classification, medical diagnosis, computer vision, and social network analysis.

¹Department of Linguistics and Language Technology, Tezpur University, Tezpur, Assam 784028, India. ²Graduate School of Arts and Sciences, The University of Tokyo, Komaba 4-6-1, Meguro, Tokyo 153-8505, Japan. ✉email: amaleshtezu@gmail.com

Introduction

Natural languages employ (different combinations of) consonants and vowels to maintain meaning contrasts. In addition to this strategy, an intriguing subset of languages exploits pitch variations to control the same purpose, i.e., to represent words (meanings) in their language. Pitch variations refer to the auditory sensation that our ears perceive. The phonetic correlate of pitch is the fundamental frequency (f_0) that refers to a (speech) signal and the number of pulses per second that the signal contains. Each pulse refers to a single vibration of the vocal cord measured in Hertz (Hz), where each Hz refers to one cycle per second (Gope 2016). Based on the functions, pitch can be categorized as tone and intonation. Tone refers to pitch patterns that intricately distinguish individual words, while intonations convey broader contextual nuances within sentences. This interplay between tones and intonations is most vividly seen in tonal languages, where variations in pitch on individual words alter their word meanings (Yip 2002). Interestingly, pitch variations are universally employed in all human languages to convey information, either at the lexical (meaning contrasts of individual words), post-lexical (accent or stress), or discursal level (sentential meaning differences) (Gope 2021).

Traditionally, researchers working on tonal languages relied on f_0 perturbation (viz., f_0 height, f_0 direction, f_0 slope), duration, and intensity, followed by suitable statistical tests and modeling to establish the tonal contrasts in a given language (Gope and Mahanta 2014). The visual interpretation of the f_0 tracks generated through a production experiment aids in predicting the tone's quality, and suitable statistical modeling helps us quantify it. This paper assesses and appraises the effectiveness of various machine learning algorithms (henceforth MLAs) in identifying intricate tonal contrasts, albeit a complex five-way tonal contrast in Chokri (ISO 639-3) (VanDriem 2007). Chokri is a Tibeto-Burman language of the Sino-Tibetan language family spoken in Nagaland, India. A production experiment was carried out to examine the tonal contrasts in this language. The visual interpretation of the f_0 tracks generated through the production experiment followed by a repeated measure ANOVA and a subsequent post-hoc Bonferroni test indicates a potential 5-way tonal contrast in this language viz. four level tones- extra high (EH), high (H), mid (M), and low (L); and one contour tone-mid-rising (MR) (Gogoi et al. 2023). The other acoustic components, viz., duration and intensity, are observed to be non-significant factors in the realization of tonal contrasts in Chokri. Therefore, we left out the duration and intensity values in our further analysis.

Recently, studies have shown promise in a departure from the conventional approach of tone analysis and adapted machine learning algorithms (MLAs) (Gogoi et al. 2021; Li et al. 2006; Wang et al. 2008), a subset of artificial intelligence, to achieve far superior and accurate results. MLAs are a group of computational algorithms adept at uncovering latent patterns (rooted in data or images) and making predictions based on their training (previous) experiences. Selecting a generalized statistical model when there are increased features and possible interactions is challenging (Gope 2021). However, the same could be considered an advantage when incorporating MLAs. These can efficiently process extensive data volumes and manage numerous features, rendering them highly suitable and remarkably effective for addressing intricate multi-dimensional research challenges.

Beyond the supervised and unsupervised categorization of MLAs, algorithms can further be classified into three distinct categories— linear, non-linear, and ensemble-based (Boehmke and Greenwell 2019). The choice of the most suitable algorithm hinges upon the specifics of the dataset and its inherent characteristics. For instance, when the data exhibits a linear trend, we

can use linear algorithms like Simple Linear Regression (SLR), Multivariate Linear Regression (MLR), Logistic Regression (LR), or Perceptron (Brownlee 2016b). On the other hand, if the data boasts numerous intricate features and discerning a linear trend is challenging, non-linear algorithms, such as Decision Trees (DT), can be adopted. Algorithms such as Naive Bayes (NB) and Hidden Markov model (HMM) are based on probabilities. In contrast, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Learning Vector Quantization (LVQ) help uncover intricate patterns within complex data points (Brownlee 2016b). Furthermore, ensemble algorithms like Bootstrap Aggregation, Random Forest (RF), and Stacked Generalization offer an effective approach when the data's features are abundant and intricate, necessitating a holistic understanding of the underlying relationships. Diving deeper, we encounter the realm of Deep Learning (DL), which can handle both linearity and non-linearity and encapsulates multi-layer perceptron (MLP), neural networks (NN), convolutional neural networks (CNN), artificial neural networks (ANN) (Brownlee 2016a), and more. Deep Learning has been proven to handle even the most intricate patterns and intricate relationships in data, making it a dynamic tool in the MLA arsenal.

Linguistic research has recently integrated MLAs into its methodologies to examine tonal contrasts (Gogoi et al. 2020; Lemus-Serrano et al. 2021; Ramdinmawii and Nath 2022) and other linguistic properties (Krasnyuk et al. 2023; Liu et al. 2023; Wang et al. 2023). To identify and explore the tonal properties in various languages, researchers have adopted various MLAs. This includes the application of SVM and DNN-based algorithms to assess the recognition of four tones in Mizo, a Kuki-Chin language spoken in Mizoram, India (Gogoi et al. 2020). This research explores the f_0 perturbations and reveals a notable challenge in distinguishing the high and low tones. Interestingly, a comparison between 1D-CNN and DNNs suggests that the former offers superior tone recognition (Gogoi et al. 2021). Further investigations employ SVM, NB, and Boosted Aggregation to probe the stressed and unstressed syllables linked to contrasting tones in Mizo (Ramdinmawii and Nath 2022). In another recent study, MLAs, including KNN, SLR, RF, and SVM, are harnessed to uncover the pivotal acoustic components driving the identification of 3-way tonal contrasts in Dharmashala Tibetan, a language spoken in the Indian subcontinent. Researchers have also combined different acoustic features (f_0 height, duration, and intensity) with MLAs and investigated different features in various languages, such as Mandarin (Chang et al. 1990; Li et al. 2006; Wang et al. 2008), Cantonese (Lee et al. 1995, 2002; Peng and Wang 2005), English (Levow 2005), and Yukuna (Lemus-Serrano et al. 2021), to name a few.

A striking realization emerges in the dynamic realm of (experimental) linguistic research integrating MLAs to yield more conclusive results. The comprehensive grasp of these methodologies, scattered across disconnected studies, often perplexes us when choosing a suitable MLA, such as exploring the intricate tonal contrasts in a given language. As we venture into this direction, our paper becomes a guiding light, illuminating the path for those embarking on the journey of data-driven intelligence. In this context, our study of Chokri tonal contrasts takes on a novel significance in multiple dimensions. Firstly, it unveils the intricacies of complex five-way tonal contrasts that include four level tones—extra high (EH), high (H), mid (M), low (L), and a contour tone— mid-rising (MR). Secondly, we pioneer the utilization of f_0 directionality as a crucial feature, augmenting our ability to identify essential traits (f_0 height, directionality, and their fusion) for MLA implementation. Furthermore, our work offers an unprecedented and comprehensive comparison of six

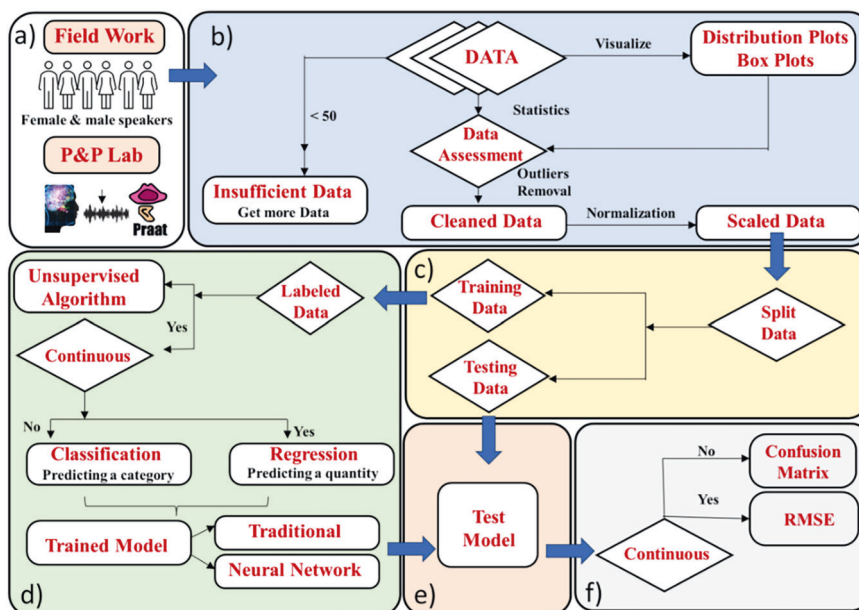


Fig. 1 Flowchart for Implementing Machine Learning Algorithms (MLAs). **a** Generation of Data: The process initiates with the collection and organization of the dataset. **b** Data preprocessing: Data refinement occurs through outlier removal and normalization, ensuring data quality. **c** Splitting into Training and Testing Datasets: The dataset is partitioned into distinct subsets for training and testing purposes. **d** MLAs Implementation: Different MLAs encompassing traditional methods and neural networks are applied to the training dataset. **e** Execution on Testing Datasets: The chosen MLAs are executed on the independent testing datasets. **f** Performance Evaluation: Results are evaluated to gauge the algorithm’s performance.

traditional MLAs (LR, DT, RF, SVM, KNN, and NB) against the backdrop of ANN. This illuminates a spectrum of methodologies, unveiling their strengths and subtleties. A primary goal of this research is to address the context itself. Chokri is an under-documented and endangered language. Through our exploration, we unravel its tonal intricacies and embark on a broader quest to understand less explored languages such as Chokri, laying the foundation for a potential corpus and nurturing its preservation.

The conventional implementation process of MLAs follows a structured path, outlined in Fig. 1a–f. This includes the implementation of crucial stages, each contributing to the accuracy and effectiveness of the outcome. The initial step involves data gathering and organization, commencing with the dataset preparation (Fig. 1a). Subsequently, the dataset undergoes preprocessing, entailing outlier removal and normalization to ensure data quality (Fig. 1b). In Fig. 1c, the dataset is partitioned into training and testing subsets, setting the stage for model evaluation. In this study, supervised MLAs take the spotlight. It includes traditional methods (LR, DT, RF, SVM, KNN, and NB) alongside the neural network (ANN), all leveraged for tonal classification in the Chokri language. The following phase entails model selection, considering whether the task involves category prediction (classification) or quantity estimation (regression). Cross-validation is then applied to gauge model performance (Fig. 1d). Moving forward, the chosen MLAs are applied to the testing datasets (Fig. 1e), paving the way for result evaluation. The ensuing step involves the construction of accuracy parameter tables (Fig. 1f), aggregating the output values and culminating in a conclusive insight. It’s imperative to highlight that the dataset preparation forms the basic of this process, and the selection of attributes (features) and the volume of instances in the final dataset predominantly influence MLA performance.

Methods

Experiments: production of data

Material and data recording procedure. The dataset for the production experiment contains eight monosyllabic toneme pairs

with five-way meaning contrasts. Participants were instructed to produce a priming sentence (containing the target word) first that would trigger the lexical meaning, followed by the target word in a fixed carrier sentence of ‘Repeat X again,’ where X is the target word, and in isolation. The sentences were given in Chokri and English. Data was recorded using a linear portable recorder (Tascam- DR-100MKII) connected with a unidirectional microphone (Shure SM10A-CN). The distance between the mouths of the participants and the mouth was approximately 25 mm to ensure a minor noise intervention and turbulence-free speech data. The speech data were recorded with a sampling frequency of 44.1 kHz and 32-bit resolution. A total of 15,400 tokens (8 toneme sets × 5 way meaning contrasts × 7 subjects × 5 repetitions × 11 time points) are analyzed in this study.

Participants. Seven native speakers (five females and two males), aged between 19 and 39 years, from the Thipüzu village in the Phek district of Nagaland, India, participated in the production experiment. None of the participants had any reported language disability or hearing impairment. All the participants speak Chokri as their first language. They also speak English and Nagamese (a lingua franca spoken in Nagaland, India) fluently. All the participants were asked to provide informed consent before the production experiment and were paid for participation.

Data annotation and acoustic measures. Post-recording, the target words were manually annotated in Praat (Boersma and Weenink 2012). Multiple tiers were created to mark each target syllable’s f0, duration, and intensity. The tier on f0 was marked following the visible f0 track in Praat; duration and intensity, on the other hand, were marked for the whole syllable. All the acoustic measurements were generated using VoiceSauce (Shue et al. 2010) for each token.

Data analysis. In the data analysis phase, we begin by processing the raw experimental data to prepare it for further analysis using machine learning algorithms (MLAs). Additionally, we describe

the various MLAs utilized in this study and their underlying principles.

Z-score calculation. The raw f_0 values are converted to Z scores to neutralize the inter-speaker and intra-speaker variability. The Z score (Adank et al. 2004) is derived by subtracting the overall average of the fundamental frequency (\bar{f}_0) across the contrastive tones from the raw f_0 value of each tone. This difference is then divided by the standard deviation (σ_{f_0}) of the measured f_0 values across all tone types. The formula to calculate the Z score is expressed as: $Z \text{ score} = \frac{f_{0i} - \bar{f}_0}{\sigma_{f_0}}$, $\bar{f}_0 = \frac{1}{n} \sum_{i=1}^n f_0(t_i)$, and $\sigma_{f_0} = \sqrt{\frac{\sum_{i=1}^n (f_{0i} - \bar{f}_0)^2}{n-1}}$ where t, t_1, t_2, \dots, t_i , and n represent any tone, aggregate measure of all tones (including repetitions), and total number (including repetitions), respectively, where f_{0i} represents the raw f_0 value of a specific i th tone. \bar{f}_0 denotes the overall average of f_0 for all tones. σ_{f_0} is the standard deviation of the measured f_0 values. The average \bar{f}_0 is computed as $\frac{1}{n} \sum_{i=1}^n f_0(t_i)$, where n represents the total number of tones, and t_1, t_2, \dots, t_i are individual tones.

Machine Learning Algorithms (MLAs). The study analyzes the data using diverse machine learning algorithms (MLAs). Here's an overview of the process:

Dataset and Division: The dataset, denoted as D , comprises pairs of input features (x_i) and corresponding output labels (y_i). The dataset is randomly divided into two sets: D_{train} and D_{test} . The supervised algorithm aims to learn the mapping from input (X) to output (Y), and it seeks to predict $y \in Y$ for any given input $x \in X$ present in the D_{train} set. The algorithm's performance is evaluated on the D_{test} set.

Data Manipulation: Data manipulation uses Pandas (ver. 0.24.2) (McKinney et al. 2011) and Numpy (ver. 1.16.4) (Harris et al. 2020) in Python (Raschka 2015). These libraries enable efficient data handling and transformation tasks.

Data Splitting: In Python, the sci-kit-learn (Pedregosa et al. 2011) library's `train_test_split()` function divides the data into training and testing datasets. A test size of 0.3 is chosen, meaning that 30% of the original data is reserved for testing, leaving the remaining 70% for training. While a 70:30 ratio is commonly used, alternatives like 60:40 to 90:10 can also be considered. The choice of split ratio affects training accuracy, but careful consideration is needed to prevent overfitting. The `random_state` parameter ensures reproducibility by producing the same results across different runs. A value of zero is used to mitigate randomness during data splitting. This study employs a repeated measures design, ensuring that multiple recordings from the same participants are considered for both training and testing purposes in such a way that the tokens of a particular subject that appear in the training set, do not occur in the testing test. The implementation of `GroupShuffleSplit` from the `sklearn.model_selection` ensures this process. The raw data is processed, transformed, and divided for subsequent analysis using different MLAs by executing the above steps. The data is now ready to undergo the training and testing phases of the machine learning algorithms.

a) **K-Nearest Neighbors (KNN):** K-Nearest Neighbors (KNN) is a classification algorithm that relies on the similarity between data points to make predictions. The choice of distance metric is crucial as it determines how the algorithm measures the similarity between data points. One common choice is the Minkowski distance, which results in the Euclidean distance formula when the parameter is set to 2. The Euclidean distance formula is given by: $D(i) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}$ where x_{1i} and x_{2i} are the values of

the specific column i for the first and second rows of data, and n is the total number of columns. Using this formula, the KNN algorithm starts by calculating the distance between two rows (data points) in the dataset. Conceptually, this calculation draws a straight line between the two data points in a multi-dimensional space. KNN predicts the class of a future data point in the testing dataset by considering the classes of its nearest neighbors. In this case, the five nearest neighbors are considered. For implementation, the `KNeighborsClassifier(n_neighbors = 5, weights = 'distance', p = 2, metric = 'minkowski')` from the `sklearn.neighbors` module is employed. The parameter `n_neighbors` specifies the number of neighbors to consider, `weights = 'distance'` indicates that closer neighbors have more influence, `p = 2` denotes the Minkowski distance with Euclidean distance, and `metric = 'minkowski'` signifies the choice of metric. This algorithm analyzes and classifies data points based on their similarity.

b) **Naive Bayes (NB):** The Naive Bayes algorithm is a probabilistic classification technique that utilizes Bayes' Theorem to calculate the probability of a data point belonging to a specific class, given our prior knowledge. Bayes' Theorem is expressed as $P(c|x) = \frac{P(x|c) \times P(c)}{P(x)}$, where $P(c|x)$ represents the probability of class c (target) given the predictor x (attributes), $P(x|c)$ is the probability of the predictor x given the class c , $P(c)$ is the prior probability of the class, and $P(x)$ is the prior probability of the predictor. Naive Bayes calculates these probabilities for each possible class and then assigns the data point to the class with the highest probability. This analysis uses the Gaussian Naive Bayes implementation, denoted as `gnb = GaussianNB()` from the `sklearn.naive_bayes` module. This classifier assumes that the likelihood follows a Gaussian (normal) distribution. Another variant, `BernoulliNB()`, was also considered but disregarded due to achieving an overall accuracy of 56%, which was unsatisfactory for the subsequent analysis.

c) **Decision Tree (DT):** The DT algorithm constructs a tree-like structure by recursively breaking the dataset into smaller subsets through binary splits based on feature values. This splitting process involves selecting the feature that provides the best separation between classes at each step. The criterion often used for making these splits is the concept of entropy. Entropy, in this context, measures the impurity or disorder in a dataset. A lower entropy implies that the dataset is more homogeneous with respect to the target classes. The entropy for a node Q_m is calculated using the formula: $P_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$, where k is the number of possible classes, $m = \text{no of nodes}$, and $Q_m = \text{data at the particular node } m$. Then entropy is computed by $E(Q_m) = -\sum_k P_{mk} (\log P_{mk})$. Here, P_{mk} is the proportion of instances in node Q_m that belong to class k . The goal of building the decision tree is to minimize entropy. At each step, the algorithm selects the feature that results in the greatest reduction in entropy when the dataset is split based on that feature. This process is repeated for every tree branch until certain stopping criteria are met, such as reaching a maximum depth or having a node with a minimum number of samples. The Decision Tree classifier builds a tree with decision nodes (internal nodes) and leaf nodes (terminal nodes). In this analysis, the `DecisionTreeClassifier` from the `sklearn.tree` module is employed with the criterion set to 'entropy' and `max_depth` set to `None` (allowing the tree to grow until all leaves are pure or until the minimum samples per leaf are reached). This classifier uses entropy to guide the decision-making process during tree construction. DT is simple to understand, visualize, and interpret. However, they can be prone to overfitting if not properly controlled through parameters like maximum depth or minimum samples per leaf.

d) **Random Forest (RF):** Random Forest often performs better than a single Decision Tree, providing higher accuracy and better

generalization to new data. The choice of the number of trees (`n_estimators`) is a hyperparameter that can be optimized. Increasing the number of trees usually improves performance up to a point, after which it might plateau or even lead to overfitting. It's robust, easy to use, and handles high-dimensional data well. However, due to the ensemble nature, it might be computationally more expensive than a single Decision Tree. The number of trees in the forest (`n_estimators`) varies from 10, 50, and 100, and we see the overall accuracy has increased when the number of trees is the highest. The entropy criterion computes the Shannon entropy of the possible classes to decide which feature to split on at each step in building the tree, similar to DT. We have used the `classifier = RandomForestClassifier(n_estimators = 100, criterion = 'entropy')` imported from `sklearn.ensemble`.

e) Multinomial Logistic Regression (LR): It is an extension of logistic regression, which solves a multi-class problem. It models the relationship between the predictors and probabilities of each class and predicts the class with the highest probability among all the classes. Let us consider $P(y = j|z^i) = \frac{e^{z_j^i}}{\sum_{k=0}^k z_k^i}$, where j is the class of the input (i), ranging from 0 to k , where k is the number of possible classes. The term, $\sum_{j=0}^k z_k^i$ normalizes the distribution such that $P(y = j|z^i) = 1$. The net input vector is z such that $z = w_1x_1 + \dots + w_mx_m + b = \sum_{l=1}^m w_lx_l + b$, where x is the feature vector of training dataset, w is the weight vector, and b is the bias unit. We have used the `classifier = LogisticRegression(multi_class = 'multinomial', solver = 'newton-cg')` imported from `sklearn.linear_model`.

f) Support Vector Machine (SVM): Support Vector Machine (SVM) is a robust and widely used machine learning algorithm for both classification (Support Vector Classification - SVC) and regression (Support Vector Regression - SVR) tasks. It works by finding the hyperplane that best separates different classes or fits the data points in the regression case. Different parameters are checked to have the best estimator for the training datasets—linear, polynomial of degree 3, and Gaussian radial basis function (RBF) with γ of [1, 0.001, 0.0001] and c of [1, 10, 100, 1000]. A kernel transforms the training dataset so that a non-linear decision surface can transform into a linear equation in a high-dimension space. The optimized kernel is given by linear scale. The net input vector is defined as z such that $z = \sum_{l=1}^m w_lx_l + b$, where w is the weight vector, x is the feature vector of the training dataset, and b is the bias unit. LR and SVM without any kernel provide almost similar performance, but the SVM is tuned better depending on the parameters. We have used the `svm_model = GridSearchCV(SVC(), params_grid, cv = 5)` imported from `sklearn.svm`.

g) Artificial Neural Network (ANN): An Artificial Neural Network (ANN) is a versatile and powerful machine learning model inspired by the human brain's neural structure. ANNs consist of interconnected layers, each performing specific transformations on the input data. The key components in building an ANN include Convolution Layer: Filters or kernels convolve across the data to capture local features like edges, corners, and textures. The output is a feature map that represents the presence of these features. Pooling Layer: After each convolutional layer, pooling layers reduce the dimensions of the feature maps while retaining the most important information. Dropout Layer: It is a regularization technique that prevents overfitting. Flattening Layer: It reshapes the multi-dimensional output from the convolutional and pooling layers into a one-dimensional vector. Fully Connected Layer: Fully connected (dense) layers connect every neuron from the previous layer to every neuron in the current layer. They enable the network to learn complex relationships between features. Proper tuning of

these hyperparameters can enhance the network's ability to extract relevant features and generalize well to new data. Epochs, Batch Size, and Learning Rate: The number of epochs determines how often the entire dataset is used to train the network. Batch size specifies the number of samples used in each iteration during training. These parameters must be carefully chosen, often through trial and error, to balance convergence speed and avoid overshooting. We have used Keras to build the ANN. There are three convolution layers, with each one including `keras.layers.Conv2D(32, kernel_size=(3, 3), activation='relu')`. We also used max pooling for each convolution layer using `MaxPooling2D(pool_size=(2, 2))`. The dropout layer is used as `keras.layers.Dropout` with a rate of 0.2. Then, the layer is flattened using `keras.layers.Flatten()` to get it ready for the dense layer. The function for the out layer consists of `keras.layers.Dense(number_of_classes, activation='softmax')`, where the number_of_classes = number of tones in this study. Everything is compiled using `model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])`.

Machine learning algorithms (MLAs) are evaluated using a set of parameters designed to assess their effectiveness in multi-class identification scenarios.

a) Confusion Matrix: It is a pivotal tool for evaluating the performance of a classifier. It compares the actual and predicted values and takes the form of an $N \times N$ matrix, where N represents the number of output classes. The matrix is constructed based on four key components: False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP). FN for a class is calculated by summing the values in the corresponding row except for the TP value. FP for a class is determined by summing the values in the corresponding column except for the TP value. TN for a class is computed by summing values across all columns and rows except those corresponding to the class in question. TP represents cases where the actual and predicted values match. The confusion matrix is generated using the `confusion_matrix` function from the `sklearn.metrics` module.

b) ROC Curve and AUC Values: Although receiver operating characteristic (ROC) curves and area under the curve (AUC) values are conventionally associated with binary classification, they can be adapted for multi-class scenarios using a one vs. rest strategy. This strategy trains the datasets to classify instances as belonging to a specific class. The `OneVsRestClassifier()` is employed, wherein each MLA is integrated. Alternatively, the one vs. one strategy, which employs a separate classifier for each combination of two or more classes, can be used. The ROC curve is a probability curve plotting the True Positive (TP) rate against the False Positive (FP) rate at various threshold levels, effectively distinguishing between the actual 'signal' and 'noise'. The AUC quantifies a classifier's ability to distinguish between classes; higher AUC values indicate better performance in distinguishing positive and negative classes. An AUC value of 1 signifies perfect classification. The ROC curve and AUC values are computed using functions like `roc_curve`, `auc`, and `roc_auc_score` from the `sklearn.metrics` module.

c) Accuracy, Precision, Recall, and F-1 Score with Micro- and Macro-weighted Averages: For multi-class identification challenges, calculating an overall F-1 score isn't straightforward. Instead, the F-1 score is computed for each class using the one vs. rest strategy. Micro- and macro-weighted averages are then used to determine the overall F1-score. Precision is given by the formula $\frac{TP}{TP+FP}$, Specificity by $\frac{TN}{TN+FP}$, Sensitivity and Recall by $\frac{TP}{TP+FN}$, and the F-1 Score by $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. Micro-averaging assigns equal weight to each instance or prediction, while macro-averaging calculates the arithmetic mean of scores across different classes. Evaluation metrics such as accuracy, precision, recall, and

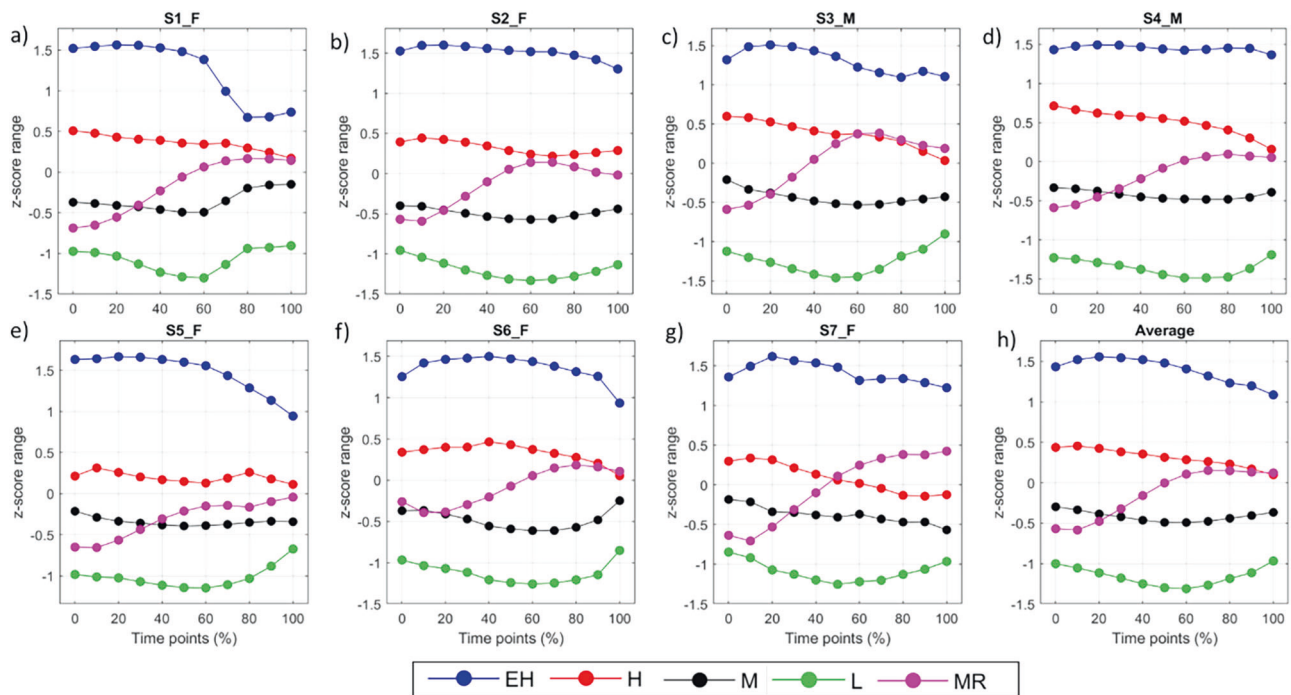


Fig. 2 Z-score values of f0 directionality for different speakers. a S1_F, b S2_F, c S3_M, d S4_M, e S5_F, f S6_F, and g S7_F, and the (h) speaker-wise average are shown. M indicates the gender of the speakers- M for males and F for females. The f0 directionality entails the division of the f0 track into eleven equidistant time points, ranging from the onset (0%) to the offset (100%) of the f0 trajectory. Different tones include four level tones- extra high (EH, in blue color), high (H, in red color), mid (M, in black color), low (L, in green color), and a contour tone (MR, in purple color).

F-1 score are calculated using functions like *accuracy_score* and *classification_report* from the *sklearn.metrics* module.

Results

In our investigation of tonal contrasts, we delve into the intricacies of f0 fluctuations, explicitly focusing on f0 directionality and height, duration, and intensity. However, duration and intensity appeared to be non-significant contributors in identifying tonal contrasts in Chokri. The computation of f0 directionality entails the division of the f0 track into eleven equidistant time points, ranging from the onset (0%) to the offset (100%) of the f0 trajectory. To trigger the meaning of the target words, a priming sentence was first recorded, followed by the target word occurring in a fixed sentence frame (where the target words occur in the middle position of the sentence) and in isolation. In our analysis, we have used the values of the tokens uttered in fixed sentence frames and in isolation to maintain uniformity. Five female and two male speakers, aged between 19 and 39 years, participated in the production experiment and naturally produced the scripted sentences. The dataset comprises eight monosyllabic toneme sets with 5-way meaning contrasts (8 toneme pairs \times 5-way lexical contrasts = 40 lexical items). The toneme pairs were randomly distributed and displayed on a monitor screen. Each speaker produced the whole set five times with an interval of 30 minutes between each repetition. Finally, the data is annotated using the Praat program (Boersma and Weenink 2012) (see Fig. 1a). Guided by the workflow outlined in Fig. 1b, the collected data undergoes preprocessing and refinement. This preparatory stage involves the removal of outliers and subsequent scaling via the Z-score normalization technique. This normalization method is a potent tool to address intra-speaker and inter-speaker variability. The transformation of these values through normalization lays the foundation for our subsequent in-depth analysis.

Visual inspection of different tones based on f0 directionality.

The normalized f0 tracks measured at 11 equidistant time points are averaged across all the tonemes and repetitions for each speaker individually and plotted in a line chart for visual inspection. The normalized data of individual speakers [Fig. 2a–g] and the mean z-score values of the individual toneme averaged across all the speakers' data represent a uniform trend [Fig. 2h]. The trend of the f0 direction confirms four level tones, viz., extra high (EH, in blue color), high (H, in red color), mid (M, in black color), low (L, in green color), and a contour tone (MR, in purple color). A noteworthy aspect worth highlighting is the depiction of speaker-wise raw f0 directionality, presented in Supplementary Fig. S1 of the supplementary section. Supplementary Figure S1 shows the f0 range disparities between females (140–300 Hz) and males (90–200 Hz). The disparities in the f0 range are due to the differences in the vocal tract of male and female speakers. The size and the grid of the grown-up adult male's vocal tract are usually bigger than their female counterparts, producing significantly lower f0. Therefore, our further investigation segregates male and female data for in-depth exploration. Additionally, sample sound files representing a toneme set ([pu] series) with five-way meaning contrast produced by one male and one female speaker are provided in SF1–SF10 of the supplementary section.

Multi-class identification of tonal contrasts using

Traditional MLAs. The significant acoustic components, viz., f0 height and direction, are included as the feature vector in this study. The data is divided into training and testing sets in the 70:30 ratios (Fig. 1c). Based on the training on the 70% data (Fig. 1d), the testing dataset is used to evaluate the performance of each MLA (Fig. 1e–f).

Evaluation of traditional MLAs based on the confusion matrix. The performance of the six traditional MLAs are highlighted

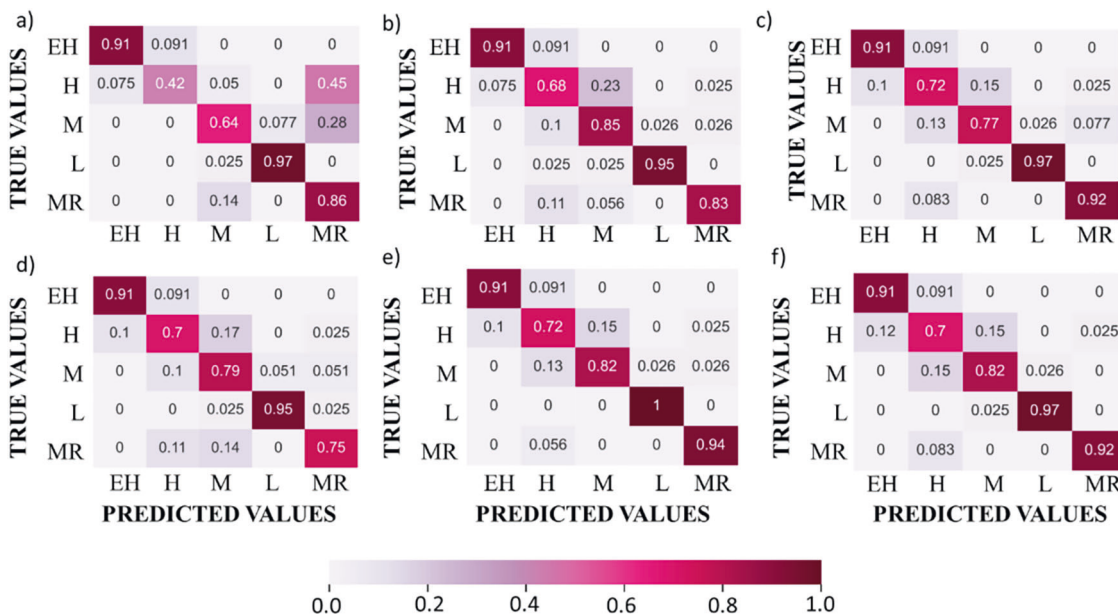


Fig. 3 Normalized confusion matrix of the female data showcasing the performances of different MLAs in classifying the five tones (EH, H, M, L, and MR) present in Chokri. a Decision Tree (DT), **b** K-Nearest Neighbors (KNN), **c** Logistic Regression (LR), **d** Naive Bayes (NB), **e** Random Forest (RF), and **f** Support Vector Machine (SVM). The color bar exhibits shades from light (=0) to dark (=1).

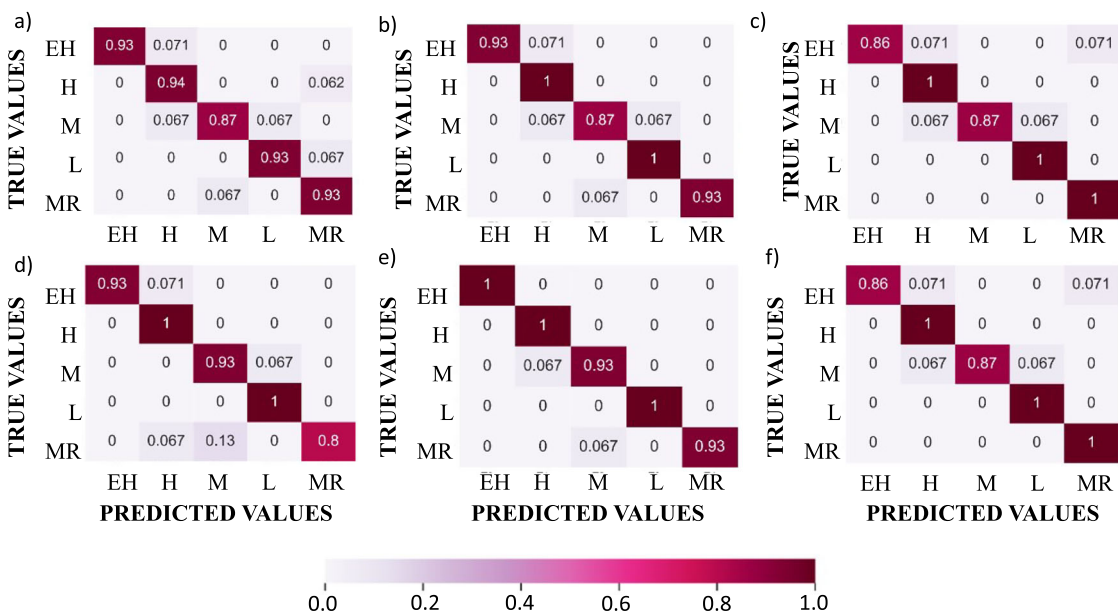


Fig. 4 Normalized confusion matrix of the male data showcasing the performances of different MLAs in classifying the five tones (EH, H, M, L, and MR) present in Chokri. a Decision Tree (DT), **b** K-Nearest Neighbors (KNN), **c** Logistic Regression (LR), **d** Naive Bayes (NB), **e** Random Forest (RF), and **f** Support Vector Machine (SVM). The color bar exhibits shades from light (=0) to dark (=1).

using a confusion matrix, as shown in our flowchart of Fig. 1f. Distinct insights emerge when breaking down the data by gender (viz., female and male), illustrated in Figs. 3a–f (female) and 4a–f (male). Each value in the confusion matrices’ rows has been normalized for a holistic comparison. A notable pattern emerges in the color variations, reflecting the dichotomy between diagonal and off-diagonal positions. The color bar (0 to 1) visually encapsulates our analysis, with diagonal entries indicating precise tone predictions, reaching 100% accuracy for individual MLAs.

The evaluation of confusion matrices of female data consistently reveals that off-diagonal values rarely surpass the 10% threshold (Fig. 3a–f). Analyzing the predictions for each tonal class provides intriguing insights. The EH tone achieves a commendable 91% accuracy, with minor divergences into the H tone (9%). Similar success is observed for the L tone (97% accuracy) and MR tone (86% accuracy). However, the M and the H tones pose challenges, with correct predictions at 64% and 42%, respectively. Observing diagonal values across MLAs, each MLA’s accuracy percentages for tonal contrasts (EH, H, M, L, and MR) fall within specific ranges

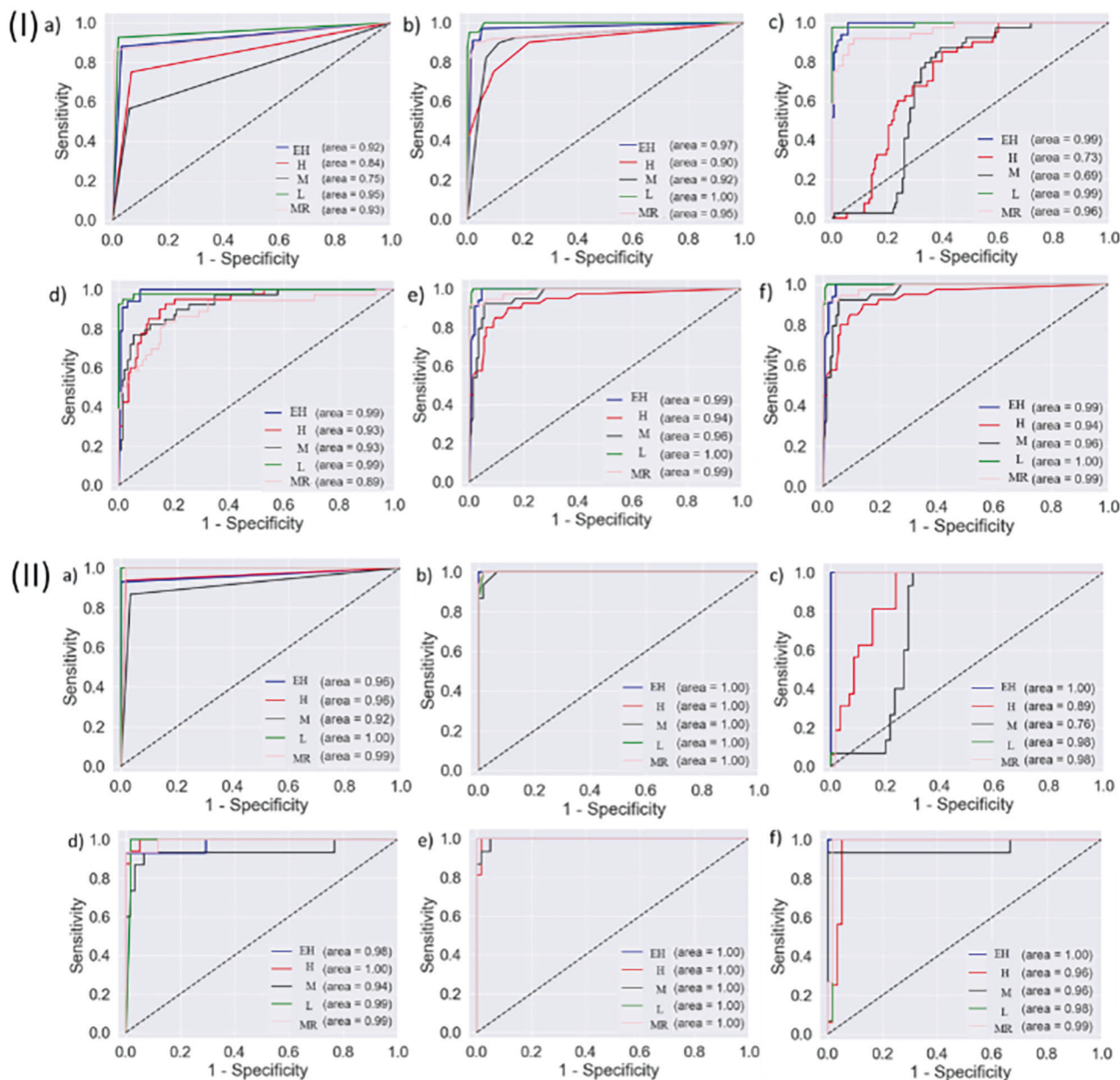


Fig. 5 The Receiver Operating Characteristic (ROC) curves show the implementation of different MLAs for classifying the contrastive tones (EH, H, M, L, and MR). I [a-f] represent the curve generated on the female speakers for different MLAs, and II [a-f] show the curve generated on the male speakers data for different MLAs. The different MLAs include **a** Decision Tree (DT), **b** K-Nearest Neighbors (KNN), **c** Logistic Regression (LR), **d** Naive Bayes (NB), **e** Random Forest (RF), and **f** Support Vector Machine (SVM).

(Fig. 3a-f). Different MLAs, viz., DT, KNN, LR, NB, RF, and SVM, achieve the accuracy range of identifying tonal contrasts (viz., EH, H, M, L, and MR) in the range of (91%, 42%, 64%, 97%, and 86%), (91%, 68%, 85%, 95%, and 83%), (91%, 72%, 77%, 97%, and 92%), (91%, 70%, 79%, 95%, and 75%), (91%, 72%, 82%, 1%, and 94%), and (91%, 70%, 82%, 97%, and 92%), for each tone type respectively. Notably, the EH tone consistently garners 91%, and the L tone surpasses 94% accuracy. The H and the M tones, however, show uncertainty across MLAs. On the flip side, the MR tone maintains an average of 89% accuracy, except for NB at 75%.

Figure 4a-f presents the normalized confusion matrix for male speakers. DT achieves 93% accuracy for the EH tone, with a 7% spill into the H tone. Similar trends emerge for the MR and L tones, standing at 93% accuracy (Fig. 4a). A striking divergence appears in predicting M and H tones when comparing female and male speakers' data, indicating a possible gender influence on

MLA performance. All MLAs achieve higher accuracy in identifying contrastive tones for male speakers. Some MLAs reach 100% accuracy without classification errors. The accuracy achieved by individual MLA in classifying the contrastive tones (EH, H, M, L, and MR) is as follows: DT = (91%, 94%, 87%, 93%, and 93%), KNN = (93%, 100%, 87%, 100%, and 93%), LR = (86%, 100%, 87%, 100%, and 100%), NB = (93%, 100%, 93%, 100%, and 80%), RF = (100%, 100%, 93%, 100%, and 93%), and SVM = (86%, 100%, 87%, 100%, and 100%). A general trend shows that H and L tones are consistently predicted with 100% accuracy, except for DT (93-94%), and MR tone is predicted with 93% accuracy by DT, KNN, and RF. The lowest accuracy (80%) is for NB in classifying the MR tone, with dual-sided misclassification- 7% with the H tone and 13% with the M tone. DT, KNN, LR, and SVM achieve 87% for the M tone, while NB and RF improve it by 6-93%.

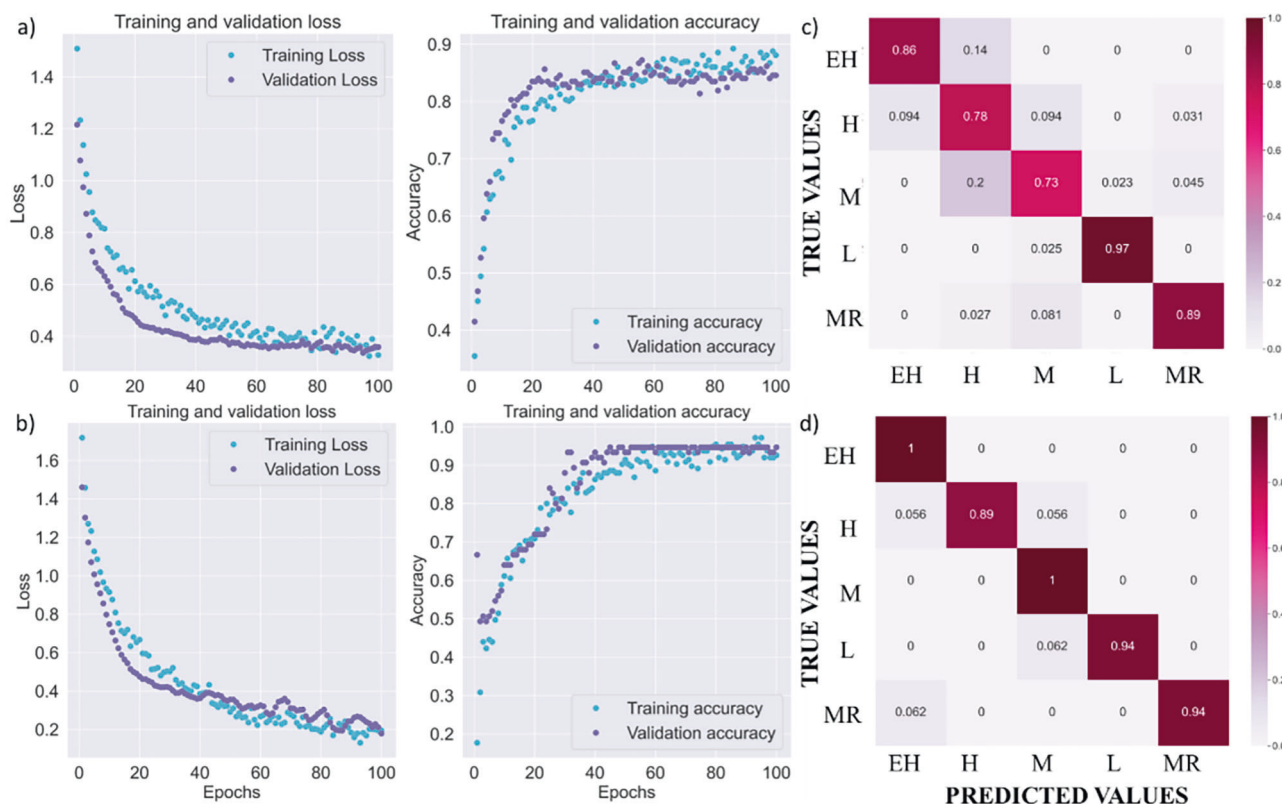


Fig. 6 Plots showcasing the training, validation, and confusion matrix of artificial neural network (ANN). The model’s loss and accuracy curves are validated using the training image set of all five tonal classes produced by the female speakers in (a) and male speakers in (b). The confusion matrix of the testing dataset for different tonal classes produced by the female and male speakers are shown in (c, d), respectively.

Evaluation of traditional MLAs based on ROC curve analysis.

Figure 5(I-II)a–f presents the Receiver Operating Characteristic (ROC) curve, evaluating the performance of classification methods (DT, KNN, LR, NB, RF, and SVM) in identifying contrastive tones produced by the female [Fig. 5(I)a–f] and male speakers [Fig. 5(II)a–f]. The true positive rate (Sensitivity) is on the vertical axis, showcasing the correct identification of positive cases. The horizontal axis reflects the false positive rate (1-Specificity), indicating the misclassification of negative cases as positive. An ideal MLA would have a curve from the bottom left to the top left, resulting in an Area Under the Curve (AUC) value of 1, signifying perfect accuracy.

In female speakers’ data, RF and SVM exhibit high AUC values for all tones (EH, H, M, L, and MR) at 0.99, 0.94, 0.96, 1.00, and 0.99, respectively, indicating excellent performance. However, DT and LR show weaker results, especially for the H tone (AUC = 0.84 and 0.73) and the M tone (AUC = 0.75 and 0.69) [Fig. 5(I)a–f].

Transitioning to male speakers, RF and KNN excel with AUC values of 1 for all tones. DT, NB, and SVM maintain respectable performance (AUC > 0.92), while LR falls short, particularly for the M tone (AUC = 0.76) [Fig. 5(II)a–f]. Comparing RF and SVM for male speakers, RF’s AUC values (all 1.00) outperform SVM (AUC = 0.96–1.00). Notably, DT’s performance improves for male speakers, while LR remains subpar for both groups.

The ROC curves visually guide us on the effectiveness of each method in distinguishing contrastive tones, while the AUC values quantitatively measure accuracy. The insights gained from these aid in selecting appropriate methods for tonal identification in specific gender groups.

Multi-class identification of tonal contrasts using neural network-based MLAs. Following conventional practices for traditional

MLAs, the ANN undergoes training on the dataset. A well-fitting algorithm should accurately conform during validation, with the validation set constituting 10% of the data unseen during training. The number of epochs denotes how often the ANN iterates through the training set, adjusting parameters based on observed errors and the optimization function.

The ANN algorithm exhibits a robust fit for both training and validation datasets, unaffected by gender differences (females in Fig. 6a and males in Fig. 6b). The loss curve consistently reduces signal noise during training without unexpected spikes. Accuracy saturates with increasing epochs, indicating a balanced training model. No signs of underfitting or overfitting emerge; both curves converge harmoniously.

Examining the confusion matrix reveals the ANN’s performance for each tonal class (EH, H, M, L, and MR) among females (Fig. 6c) and males (Fig. 6d). Tone L achieves 97% accuracy, with a 3% misclassification as tone M. EH, H, M, and MR achieve accurate identification rates of 86%, 78%, 73%, and 89%, respectively. Among females (Fig. 6c), tones H and M exhibit slightly lower accuracy (73–78%). In males (Fig. 6d), EH and M achieve 100% accuracy, while L and MR showcase 94%. Misclassifications include 6% of L confused with M and 6% of MR misclassified as EH. Overall, male speakers outperform females in ANN performance, mirroring trends in traditional MLAs.

Comparison of the performance of all MLAs based on different features and F1-scores. We have calculated the aggregate F1-scores in Fig. 7a–b for female and male speakers to evaluate each MLA’s effectiveness in the multi-class identification of various tones. All seven classifiers (DT, KNN, LR, NB, RF, SVM, and ANN) underwent extensive evaluation, with detailed metrics in

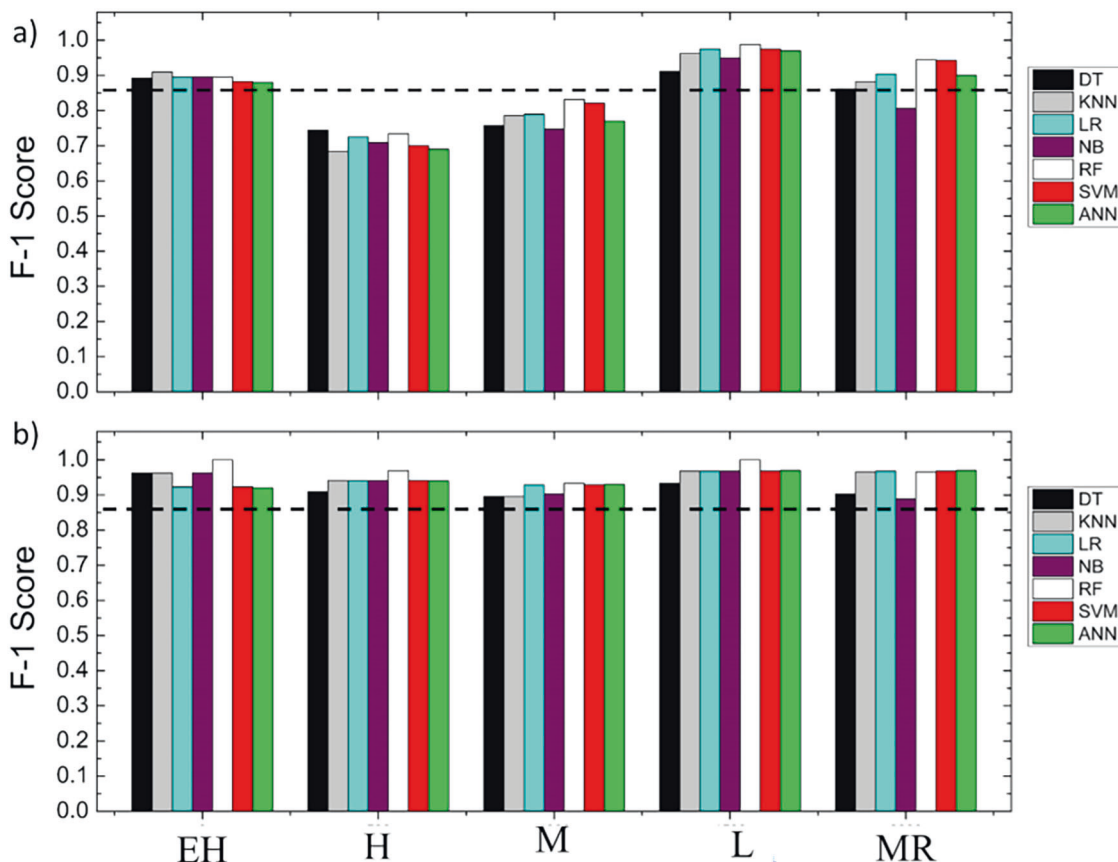


Fig. 7 F1-score to estimate the performance of different MLAs for classifying contrastive tones (EH, H, M, L, and MR). Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) for **a** Females, and **b** Males. The dashed line shows an F1-score of 0.85.

Table 1 Evaluation of different MLAs for classifying different tones (EH, H, M, L, and MR).

MLAs	F_f0	M_f0	F_f0 directionality	M_f0 directionality
DT	0.83076	0.920315	0.798672	0.906667
KNN	0.841651	0.946500	0.823503	0.946500
LR	0.855796	0.945903	0.833356	0.945903
NB	0.819437	0.932509	0.814570	0.932509
RF	0.876814	0.973305	0.871904	0.973305
SVM	0.862024	0.945903	0.856361	0.945903
ANN	0.844666	0.957733	0.847646	0.957733

Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and ANN, for different features and gender combinations in terms of average F1-score: a) Females with f0 height+f0 directionality as features (F_f0), b) Males with f0 height + f0 directionality as features (M_f0), c) Females with f0 directionality as a feature (F_f0 directionality), and d) Males with f0 direction as a feature (M_f0 directionality).

Supplementary Table ST1 in the Supplementary Section. The metrics, including accuracy, precision, recall, and F1-score, exhibit consistency across MLAs. The F1-score, a harmonic average of precision and recall, holds particular significance. A high F1 score indicates elevated precision and recall, while a low score suggests both are diminished.

Figure 7a, b reveals a trend: all MLAs, traditional and neural networks alike, demonstrate admirable performance with male speaker data, achieving F1-scores surpassing 0.85. Conversely, F1-scores decline to approximately 0.7 for female data, especially in identifying tones H and M. Notably, traditional MLAs like NB

exhibit an F1-score under 0.8 when pinpointing tone MR. RF stands out with an F1-score of ~0.88 for females and an even more impressive ~0.97 for males, surpassing other MLAs (DT, KNN, LR, NB, SVM, and ANN). NB performs least for females (~0.81), while DT shows the lowest F1-score for male speaker data at 0.92. Delving deeper, DT and NB hover around 0.82–0.83 for female speaker data, while ANN achieves F1-scores of 0.84 and 0.95 for females and males, respectively.

Our investigation explores the impact of f0 height, f0 direction, or their combined influence on MLA performance. Two feature vectors are considered: (a) f0 directionality alone, excluding f0 height, and (b) a vector incorporating both f0 height and f0 directionality. Table 1 presents MLA performance in terms of the crucial F1-score metric, with detailed data in Supplementary Table ST2 for all seven classifiers (DT, KNN, LR, NB, RF, SVM, and ANN), encompassing accuracy, precision, recall, and F1-score for both female and male speakers using these feature vectors.

A notable observation emerges: for male speaker data, F1-scores by KNN, LR, NB, RF, SVM, and ANN remain consistent, regardless of considering f0 height (scores: 0.9465, 0.9459, 0.9325, 0.9733, 0.9459, and 0.9577, respectively). However, DT displays a 2% F1-score increase when both f0 height and f0 direction are considered. Conversely, female speaker data shows an overall enhancement in MLA performance. When comparing f0 height + f0 direction to solely f0 direction, F1-scores improve for all MLAs. For instance, DT scores progress from 0.79 to 0.83, KNN from 0.84 to 0.82, LR from 0.85 to 0.83, NB from 0.8194 to 0.8145, RF from 0.8768 to 0.8719, SVM from 0.86 to 0.85, and

ANN from 0.8476 to 0.8446. This underscores the significance of considering both f_0 height and directionality, particularly for female speaker data.

Discussion

Performance measures and evaluation of different MLAs.

Evaluating the performance of machine learning algorithms (MLAs) for tonal classification requires a comprehensive assessment using various metrics that include confusion matrix, AUC measure, ROC curve, accuracy, precision, recall, and F1-score. It's noteworthy that, on average, male speaker data consistently performs better than female data across all MLAs, whether traditional or network-based MLAs are implemented in this study (see Supplementary Table ST2 in the supplementary section). Although classification accuracy is commonly employed for assessing MLA performance, it might not be suitable when class distribution is imbalanced and different errors carry varying costs. The F1 score balances precision and recall and provides a more informative evaluation (see Table 1).

An interesting trend emerges in the area under the ROC curve (AUC). KNN and RF perform similarly for male data (see Fig. 5(II)b, e), while RF surpasses KNN for female data (see Fig. 5(I)b, e). The F1-score averages around 84% for females and 94% for males with KNN, whereas RF achieves a 3% improvement, reaching 87% for females and 97% for males (see Table 1). This suggests that RF outperforms KNN and is a more suitable choice. However, MLAs' performance evaluation goes beyond a single metric. LR, for instance, yields an F1-score of 85% for females and 95% for males, showcasing impressive results (see Table 1). Nevertheless, closer examination reveals LR's challenges in classifying the H and the M tones, indicating its limitations. Similarly, DT exhibits F1-scores of 83% and 92% for females and males, respectively. Despite these seemingly good scores, the confusion matrix uncovers DT's struggle to capture tones H and M, warranting its exclusion (see Figs. 3a and 4a). Another case emerges with Naive Bayes (NB), which should be omitted due to its 75% accuracy in classifying tone the MR, significantly lower than the 89% average accuracy across MLAs for female speakers (see Figs. 3d and 4d). The RF and SVM comparison presents a nuanced scenario. While both exhibit high AUC values for female speaker data, RF surpasses SVM in male speaker data, indicating RF's superiority (see Figs. 3e-f and 4e-f).

It is crucial to emphasize that judging an MLA's performance requires considering all relevant metrics rather than relying solely on the average F1-score. This holistic perspective makes RF the preferred choice among the traditional MLAs; nevertheless, compared to neural network-based algorithms (ANN) [see Fig. 6c, d], both prove equally adept at classifying the five contrastive tones in Chokri. ANN achieves an average F1-score of 84–87% for females and 95–97% for males, highlighting its competence (see Table 1).

Evaluation of feature importance by implementing different MLAs.

Since we have established RF as the best-performing MLA in this study, our focus is narrowed to exploring features specifically relevant to RF. Table 1 provides insights into the variation in MLAs' performance levels based on the average F1-score. Notably, including f_0 height as a feature yields an overall enhancement in RF's performance for female speaker data, while there's no notable impact on male speakers. For example, when comparing female speaker data with and without f_0 height as a feature, Accuracy, Precision, Recall, and F1-score exhibit improvements of 0.532%, 0.322%, 0.532%, and 0.491%, respectively (see Supplementary Table ST2 in the Supplementary Section). This observation leads to the conclusion that

incorporating f_0 height as a feature may be advantageous for enhancing the performance of RF, specifically for female speakers. However, this augmentation does not significantly influence male speakers' ability to discern different tones in the Chokri language. This insight sheds light on the nuanced relationship between features and MLA performance, emphasizing the importance of tailoring features to specific contexts and characteristics of the data.

This study highlights a captivating finding—contrary to the prevailing notion that the neural network techniques exemplified by Artificial Neural Networks (ANN) do not necessarily outperform traditional MLAs in all scenarios. Instead, the approach's efficacy hinges on various factors, including dataset quality, quantity, class complexity, and feature representation. While neural networks might demonstrate robust performance in many instances, their superiority is not guaranteed. This investigation underscores that tried-and-true methods, such as Random Forest (RF), can effectively discern complex tonal distinctions. Furthermore, this study underscores a gender-specific nuance in the feature composition. Combining f_0 height and f_0 directionality is a pivotal feature for female speakers, enhancing tonal contrast discernment. Interestingly, relying solely on f_0 direction for male speakers appears sufficient to achieve the same task. It is worth noting that each of the seven MLAs exhibits commendable performance in classifying the five tonal contrasts in Chokri. However, the key takeaway is the importance of comparisons and selection of MLAs for various investigations. This insight transcends the tonal classification, serving as a generalized framework for evaluating MLAs across diverse problem domains like phoneme classification and image detection.

In tonal classification, this paper establishes that MLAs can achieve an accuracy range of 84–87% for female speakers and 95–97% for male speakers. This adds significance to the study, especially considering that the size and shape of the vocal cords of a grown-up male adult are relatively bigger and wider, leading to a relatively lower f_0 compared to their female counterparts. This study, however, does not draw conclusive evidence if this factor (the relative differences amongst the contrastive tones being less in terms of f_0) is an advantage for the MLAs to detect the intricate tonal contrasts exhibited in Chokri. Its status as a less documented and endangered language adds another layer of significance to this work. By unraveling the complexity of the tonal contrasts in Chokri, the study also provides a noble technique for examining the tonal structure of a given language. It is to be noted that the present work is the first comprehensive study based on production experiments to establish the complex tonal structure in this language. This work will help other researchers who aim to explore other linguistic aspects of Chokri and safeguard its longevity. This research demonstrates how technology can be harnessed to explore complex linguistic nuances, such as the multi-class tonal contrasts in a language.

Data availability

The data that support the findings of this study are available with the corresponding author (Amalsh Gope, email id: amaleshtezu@gmail.com) and can be made available upon reasonable request. The representative sound files are given in the Supplementary Section. The details include: (a) SF1_s3_M_pu_bridge_sen, where SF1 = soundfile 1, s3 = subject 3, M = Male, [pu] = 'pu' series, English equivalent = bridge, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a male speaker, (b) SF2_s3_M_pu_fall_sen, where SF2 = soundfile 2, s3 = subject 3, M = Male, [pu] = 'pu' series, English equivalent = fall, sen = fixed sentence frame, and

the target word occurs in the sentence medial position, uttered by a male speaker, (c) SF3_s3_M_pu_fat_sen, where SF3 = soundfile 3, s3 = subject 3, M = Male, [pu] = 'pu' series, English equivalent = fat, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a male speaker, (d) SF4_s3_M_pu_hit_sen, where SF4 = soundfile 4, s3 = subject 3, M = Male, [pu] = 'pu' series, English equivalent = hit, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a male speaker, (e) SF5_s3_M_pu_take_sen, where SF5 = soundfile 5, s3 = subject 3, M = Male, [pu] = 'pu' series, English equivalent = take, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a male speaker, (f) SF6_s2_F_pu_bridge_sen, where SF6 = soundfile 6, s2 = subject 2, F = Female, [pu] = 'pu' series, English equivalent = bridge, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a female speaker, (g) SF7_s2_F_pu_fall_sen SF7 = soundfile 7, s2 = subject 2, F = Female, [pu] = 'pu' series, English equivalent = fall, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a female speaker, (h) SF8_s2_F_pu_fat_sen, where SF8 = soundfile 8, s2 = subject 2, F = Female, [pu] = 'pu' series, English equivalent = fat, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a female speaker, (i) SF9_s2_F_pu_hit_sen, where SF9 = soundfile 9, s2 = subject 2, F = Female, [pu] = 'pu' series, English equivalent = hit, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a female speaker, (j) SF10_s2_F_pu_take_sen, where SF10 = soundfile 10, s2 = subject 2, F = Female, [pu] = 'pu' series, English equivalent = take, sen = fixed sentence frame, and the target word occurs in the sentence medial position, uttered by a female speaker.

Received: 5 January 2024; Accepted: 25 April 2024;
Published online: 10 May 2024

References

- Adank P, Smits R, Van Hout R (2004) A comparison of vowel normalization procedures for language variation research. *J Acoust Soc Am* 116:3099–3107
- Boehmke B, Greenwell BM (2019) *Hands-on machine learning with R*. CRC Press. Boca Raton, Florida, USA
- Boersma P, Weenink D (2012) *Praat: Doing phonetics by computer (version 5.3.82)[computer software]*. Institute of Phonetic Sciences, Amsterdam
- Brownlee J (2016a) *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery. Australia
- Brownlee J (2016b) *Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end*. Machine Learning Mastery. Australia
- Chang P-C, Sun S-W, Chen S-H (1990) Mandarin tone recognition by multi-layer perceptron. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, New Mexico, USA, p 517–520
- Gogoi P, Dey A, Lalhminghlui W, Sarmah P, Prasanna SR (2020) Mahadeva Lexical tone recognition in mizo using acoustic-prosodic features. In *Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association (ELRA), Marseille, France, p 6458–6461*
- Gogoi P, Kalita S, Lalhminghlui W, Sarmah P, Prasanna SRM (2021) Learning mizo tones from f0 contours using 1d-cnn. In *Speech and Computer: 23rd International Conference, SPECOM 2021*. Springer, St. Petersburg, Russia, pp, 214–225
- Gogoi T, Tetseo S, Gope A (2023) The phonetics of downtrends in chokri. In *Proceedings of the 20th International Congress of Phonetic Sciences, Prague 2023*. GUARANT International Spol., Prague, Czech Republic, pp 1628–1632
- Gope A (2021) *The phonetics of tone and voice quality interactions in sylheti*. Languages 6:154
- Gope A (2016) *Phonetics and phonology of sylheti tonogenesis*. PhD thesis, IIT Guwahati
- Gope A, Mahanta S (2014) Lexical tones in sylheti. In *Fourth International Symposium on Tonal Aspects of Languages, Nijmegen, International Speech Communication Association (ISCA), the Netherlands, pp 10–14*
- Harris CR et al. (2020) *Array programming with numpy*. Nature 585:357–362
- Lee T, Lau W, Wong YiuWing, Ching PC (2002) Using tone information in cantonese continuous speech recognition. *ACM Trans Asian Lang Inf Proces* 1:83–102
- Lee T, Ching PC, Chan Lai-Wan, Cheng YH, Mak B (1995) Tone recognition of isolated cantonese syllables. *IEEE Trans Speech Audio Proces* 3:204–209
- Lemus-Serrano M, Allasonnière-Tang M, Dediú D (2021) What conditions tone paradigms in yukuna: Phonological and machine learning approaches. *Glossa* 6:1–22
- Levov G-A (2005) Context in multi-lingual tone and pitch accent recognition. In *Ninth European Conference on Speech Communication and Technology*. International Speech Communication Association (ISCA), Lisbon, Portugal
- Li X et al. (2006) Mandarin chinese tone recognition with an artificial neural network. *J Otol* 1:30–34
- Liu M, Li Y, Su Y, Li H (2023) *Text complexity of chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms*. Scientific Studies of Reading, Taylor & Francis, London, United Kingdom, pp 1–21
- Maxim Svitlana K et al. (2023) Features, problems and prospects of the application of deep machine learning in linguistics. In *Bulletin of Science and Education (Series" Philology", Series" Pedagogy", Series" Sociology", Series" Culture and Art", Series" History and Archeology")*. East European Scientific Journal, Warsaw, Poland
- McKinney W et al. (2011) pandas: a foundational python library for data analysis and statistics. *Python High Performance Sci Comput* 14:1–9
- Moirra Jean Winsland Yip (2002) *Tone*. Cambridge, United Kingdom
- Pedregosa F et al. (2011) Scikit-learn: Machine learning in python. *J Mach Learn Res* 12:2825–2830
- Peng G, Wang WS-Y (2005) Tone recognition of continuous cantonese speech based on support vector machines. *Speech Commun* 45:49–62
- Ramdinmawii E, Nath S (2022) A preliminary analysis on the correlates of stress and tones in mizo. *ACM Trans Asian Low Resource Lang Inf Proces* 22:1–15
- Raschka S (2015) *Python machine learning*. Packt Publishing Ltd. Birmingham, United Kingdom
- Shue Yen-Liang, Keating P, Vicens C, Yu K (2010) *Voicesauce: A program for voice analysis*. Energy 1:H1–A1
- VanDriem G (2007) Endangered languages of south asia. In *Language diversity endangered*, Mouton de Gruyter, Berlin, Germany, p 303–341
- Wang S, Li R, Wu H (2023) Integrating machine learning with linguistic features: A universal method for extraction and normalization of temporal expressions in chinese texts. *Comput Methods Programs Biomed* 233:107474
- Wang Xiao-Dong, Hirose K, Zhang Jin-Song, Minematsu N (2008) Tone recognition of continuous mandarin speech based on tone nucleus model and neural network. *IEICE Trans Inf Syst* 91:1748–1755

Acknowledgements

This research is partially funded by National Language Translation Mission (NLTM): BHASHINI, sponsored by the Ministry of Electronics & Information Technology, Government of India for the project entitled "Speech Datasets and Models for Tibeto-Burman Languages" (SpeeD-TB), Project No: DoRD/LLT/AG/20-521/826-A. AP acknowledges the Japan Society for Promotion of Science (JSPS), KAKENHI Grant No. 23KF0104, and expresses appreciation for the JSPS International Postdoctoral Fellowship for Research in Japan (Standard) for the period 2023-25.

Author contributions

Conceptualization: AG, AP and ST. Dataset Preparation: ST. Experimental Design: ST, TG, and AG. Data Recording: ST, TG and DKB. Data annotation: ST, TG, JJ and DKB. Data curation: AP, JJ, DKB, and ST. Formal analysis: AG, AP, and ST. Investigation: AG, AP, ST, and JJ. Methodology: AG, AP, TG, and ST. Project administration: AG. Validation: AG, AP, ST, and TG. Visualization: AP, ST, JJ, DKB, and AG. Writing (original draft): ST, AP, TG, and AG. Writing (review and editing): AG, Funding acquisition: AG, and Overall Supervision: AG. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

Approval was obtained from the ethics committee of Tezpur University, comprising of the Member Secretary, Dean of Research and Development, and two internal experts, granted under order No. DoRD/TUEC/10-14 (Vol-III)/1528-4. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Informed consent

The informed consent was obtained from all participants for this study prior to the data collection. The participants were explained about the recording procedure, and a demo was shown before the actual recording started. An agreed consent form was signed by the participants. A token amount was also offered to them for participating in the recording experiment. The consent form is provided in the supplementary file.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-024-03113-2>.

Correspondence and requests for materials should be addressed to Amalesh Gope.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024