**Edelweiss Applied Science and Technology**
*ISSN: 2576-8484*
*Vol. 8, No. 6, 6480-6497*
*2024*
*Publisher: Learning Gate*
*DOI: 10.55214/25768484.v8i6.3402*


# Machine learning for phonological analysis: A case study in gender prediction

Yasser A. Al Tamimi[1], Lotfi Tadj[2*]
[1]English Department, Alfaisal University, Riyadh 11533, Saudi Arabia; yaltamimi@alfaisal.edu (Y.A.A.T.).
[2]Department of Industrial Engineering, Alfaisal University, Riyadh 11533, Saudi Arabia; ltadj@alfaisal.edu (L.T.).

**Abstract:** Al Tamimi and Smith (2023) use a conventional phonological framework to investigate gender differentiation in a corpus of 656 Saudi Arabian first names. Their findings suggest that no single phonological feature—such as the number of phonemes, syllable structure (open vs. closed), stress patterns, or the voicing of initial and final consonants—can definitively determine gender. However, a combination of these features can collectively facilitate accurate gender identification. Expanding on this premise, the current study integrates phonological analysis with machine learning, employing both supervised techniques (e.g., Naïve Bayes) and unsupervised methods (e.g., k-Means Clustering) to explore whether machine learning can effectively predict gender based on these phonological characteristics. Specifically, this study compares the performance of classification methods—Gradient Boosting Machine (GBM), Random Forest, and k-Nearest Neighbors (k-NN)—against clustering methods, including hierarchical clustering and DBSCAN. The methodology involves a detailed analysis of model performance metrics, such as accuracy, F1 scores, and clustering indices, to comprehensively evaluate the accuracy and effectiveness of each approach in gender classification. The results indicate that classification methods significantly outperform clustering approaches, with the GBM model demonstrating particularly high accuracy and balanced performance across genders. In contrast, clustering methods struggled, particularly in classifying male names, due to their reliance on similarity-based grouping rather than explicit class labeling. These findings suggest that while clustering methods may be helpful to for exploratory data analysis, they are inadequate for precise gender classification. The study's implications highlight the critical importance of selecting appropriate methodologies for classification tasks, demonstrating the superiority of classification models in gender prediction.

***Keywords:*** *Gender identification, K-means clustering, Machine learning, Naive bayes, Phonological features.*

## 1. Introduction

As a fundamental component of linguistics, phonology—the study of sound patterns in language—primarily investigates the role of sounds within linguistic systems to create syllables, words, phrases, and longer utterances. The phonological structure encompasses a range of segmental features, including vowels and consonants; suprasegmental features, such as stress, intonation, and juncture; and extra-segmental features, such as syllable structure, phonotactics, prosodic hierarchy, and temporal aspects (Anderson et al., 2008). Traditionally grounded in descriptive and theoretical frameworks, phonological analysis has evolved to intersect with interdisciplinary fields, including cognitive science, sociology, psychology, neuroscience, and computer science, among other disciplines. Recent advancements in computational methods, particularly machine learning, have transformed phonological research by enabling the analysis of extensive linguistic datasets, revealing complex patterns and relationships that were previously overlooked. Integrating these computational approaches enhances analytical precision and fosters interdisciplinary collaboration, making each researcher feel included and part of a larger scientific community. As phonology progresses, combining traditional methods with modern techniques


* Correspondence: ltadj@alfaisal.edu

promises to yield new theoretical developments and a more nuanced understanding of language structure and use.

Machine learning, a branch of artificial intelligence, allows computers to learn from data and generate predictions. Its integration into phonology has transformed many linguistic tasks, such as speech recognition, phonotactic analysis, accent detection, and language acquisition research. Machine learning models deal effectively with linguistic data, as phonological features like syllable count, stress patterns, and vowel-consonant sequences can be quantified and modeled mathematically. By encoding these features as data points, machine learning algorithms can reveal hidden patterns and relationships that might elude traditional linguistic analysis.

One prominent application is in automatic speech recognition (ASR), where machine learning models such as neural networks and Hidden Markov Models (HMMs) are used to transcribe spoken language by recognizing phonemes, the smallest distinctive sound units, in a speech stream. Similarly, in dialect and accent recognition, machine learning techniques analyze phonological variation to classify speech based on regional or social factors. These computational approaches offer greater precision and scalability than traditional linguistic analysis methods.

Recent advancements in machine learning, such as deep learning and sequence-based models like LSTM (Long Short-Term Memory) networks, have further improved the accuracy and capability of phonological analysis. These models are particularly adept at processing sequential data, rendering them well-suited for phonemic and syllabic analyses, where the sequence of sounds is crucial. Such developments have led to more sophisticated applications like real-time speech translation and more accurate speech synthesis systems.

In addition to its essential role in speech technologies, machine learning has been instrumental in advancing the study of phonotactics, which involves the rules governing permissible sound sequences within a language. By employing unsupervised learning algorithms, researchers can uncover phonotactic constraints that may be difficult to identify through manual analysis, thereby enriching our understanding of the underlying structure of language. Furthermore, machine learning is pivotal in diagnosing phonological disorders, as models analyze speech patterns to detect deviations from normal phonological processes, facilitating early diagnosis and targeted intervention.

While machine learning provides significant advantages in processing linguistic data, it has challenges. Phonological data is inherently complex, featuring subtle sound patterns that might be difficult to capture in a computational model. Additionally, the risk of overfitting, especially with small or imbalanced datasets, poses another challenge for researchers in this field. Addressing these challenges is crucial for effectively applying machine learning methods in phonological analysis.

Despite such challenges, the role of machine learning in phonological analysis continues to expand, and its potential to reveal intricate patterns and offer more precise insights is becoming increasingly evident. However, traditional analytical approaches are fundamental in building a foundational understanding of linguistic structures and feeding machine learning research. A case in point is the study by Al Tamimi and Smith (2023), which employs a conventional phonological framework to investigate gender differentiation in Saudi first names. Specifically, their analysis of a dataset of 237 male and 419 female names used several phonological variables, including the number of phonemes, syllable structure (open vs. closed), stress patterns, and the voicing of initial and final consonants. Their findings reveal that female names tend to have fewer phonemes, often start with open syllables, and frequently stress the second syllable, while male names are more likely to end with voiced consonants and stress the third syllable. The study concludes that "although no single phonological variable can definitively determine gender, a combination of these phonological features allows for accurate identification of the gender of Saudi names" (p. 449).

This paper investigates the claim made by Al Tamimi and Smith (2023), which posits that "the combination of the variables in question could conceivably account for the ability to identify the gender of a Saudi name correctly." Building on this premise, the current research seeks to explore whether machine learning techniques can accurately predict the gender of an individual based on the

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

phonological features of their name, as specified above. By applying supervised machine learning techniques like Naive Bayes, we aim to predict the gender of individuals based on these features.

Additionally, we utilize unsupervised learning techniques like k-Means Clustering to explore natural groupings in the data, examining whether phonological features alone can cluster names into meaningful gender categories without prior labeling. This approach not only contributes to the field of phonology by highlighting the role of sound patterns in social identity but also demonstrates the potential of machine learning in uncovering latent structures within linguistic data.

This study integrates phonology with machine learning, contributing to both fields by demonstrating the effectiveness of computational models in predicting social variables and revealing underlying phonological structures. By bridging the gap between linguistic theory and computational techniques, we aim to provide new insights into the relationship between sound patterns and gender while also showcasing the broader applicability of machine learning to linguistic analysis.

## 2. Literature Review

The intersection of machine learning and phonological analysis has ushered in a new era of linguistic exploration, enabling researchers to uncover intricate patterns embedded in language sounds. A prime example of this synergy is gender prediction, where machine learning algorithms leverage phonological features to classify names or words based on their perceived gender associations. This literature review delves into recent advancements in this interdisciplinary field, highlighting the crucial roles of phonological features, sound symbolism, computational linguistic methodologies, and the persistent challenges that shape this domain.

### *2.1. Phonological Features and Gender Classification*

Empirical research consistently highlights the predictive power of phonological cues in gender classification. Al Tamimi and Smith (2023) delve into the phonological characteristics of Saudi Arabian anthroponyms, uncovering gender-specific patterns in their sound structures. Cho (2024) showcases the effectiveness of deep neural networks in classifying the gender of Korean personal names, even surpassing human judgments in some instances. Additionally, studies by Fredrickson, A. (2007) solidify the connection between specific phonological features, such as syllable count, vowel quality, consonant features, and gender. These findings collectively demonstrate the substantial influence of phonology on shaping gendered language patterns. Expanding the scope to diverse linguistic contexts, Jia and Zhao (2019) focus on gender prediction based on Chinese names, while Tripathi and Faruqui (2011) explore similar patterns in Indian names, underscoring the cross-linguistic relevance of phonological cues in gender classification.

### *2.2. Sound Symbolism and Gender*

Sound symbolism, the concept that sounds possess inherent meanings, traces its origins to Plato's Cratylus, which discusses the natural connections between words and meanings (Plato, 360 BCE). Otto Jespersen (1922) observed that high-pitched vowels often suggest smallness, while Roman Jakobson (1965) explored the phonological aspects of this phenomenon. John Ohala (1994) examined the biological and perceptual influences on sound-meaning relationships. Together, these scholars provide a foundational understanding of sound symbolism across languages.

Considering its implications, the intricate relationship between sound symbolism and gender perception remains a captivating research area. Chun Hau Ngai et al. (2024) explore sound symbolism in Japanese names using machine learning, uncovering intriguing sound-gender associations and suggesting a cross-linguistic dimension to this phenomenon. Kilpatrick (2023) highlights the role of sound symbolism in automatic emotion recognition and sentiment analysis, demonstrating its implications beyond gender prediction. Furthermore, Lee et al. (2024) examines the emotional responses of Korean and Chinese women to Hangul phonemes, showcasing the subtle interplay between phonology, gender, and emotional perception. Lastly, Moorthy et al. (2018) evaluate the influence of

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

sound symbolism on brand name gender perception, offering valuable insights into its practical applications.

*2.3. Computational Linguistics Approaches*

Advancements in computational linguistics, particularly the emergence of deep learning models, have transformed phonological analysis. The capacity of these models to extract intricate features and patterns from large datasets has driven the development of increasingly accurate and nuanced gender prediction models. Manning (2015) provides a comprehensive overview of the synergy between computational linguistics and deep learning, emphasizing its significant impact on various natural language processing (NLP) tasks. Moreover, research by Cao and Daumé III (2021) examines the potential for bias in coreference resolution systems, advocating for the integration of gender inclusivity throughout the machine learning lifecycle.

*2.4. Challenges and Advances*

Despite remarkable progress, the field continues to encounter challenges such as potential bias in machine learning models and the need for interpretability. Addressing these concerns is paramount for ensuring responsible and ethical AI development. Concurrently, ongoing research into data collection, annotation techniques, and the pursuit of explainable AI offers promising solutions (Arrieta et al., 2020).

The evolving landscape of machine learning and phonological analysis offers profound potential for revealing intricate connections between language, sound, and gender. By advancing our understanding of gendered language patterns and facilitating the creation of fairer, more transparent AI systems, this interdisciplinary field is well-positioned for transformative breakthroughs.

In conclusion, integrating machine learning with phonological analysis has opened new avenues for investigating the complex relationship between language and gender. The studies reviewed underscore the predictive power of phonological features, the nuanced role of sound symbolism, and the transformative potential of computational linguistic methods in gender prediction tasks. Although challenges persist, ongoing advancements in machine learning and ethical AI development offer promising pathways for future research. As this interdisciplinary field progresses, we can anticipate a deeper understanding of the intricate connections between language, sound, and social identity.

## 3. Dataset

The data considered in this paper consists of the first names of 656 Saudis, 237 males, and 419 females, drawn from the registrar of a private university in Riyadh, Saudi Arabia.

In this case study, the dependent variable is the qualitative 'gender' with its two categories: "Female" and "Male." The independent variables are the following:

- $x_1$: Phonemes (Quantitative variable)
- $x_2$: Syllables (Quantitative variable)
- $x_3$: Syllable structure (or syllable pattern or syllable configuration). This variable describes different combinations of open and closed syllables. It is a qualitative variable with 22 categories, such as closed, closed+closed, and open+open+open+closed.
- $x_4$: Stress position (or stress pattern). This variable indicates the position of the stress in a word. It is a qualitative variable with five categories: first, second, third, fourth, and monosyllabic.
- $x_5$: Word-initial sound (or initial segment type, or Word-initial consonants vs. vowels). This variable indicates the type of sound (vowel or consonant) at the beginning of the word. It is a qualitative variable with 41 categories, such as cons = b, cons = d, or vowel = a.
- $x_6$: Manner of articulation of initial consonants. This refers to how the consonant at the beginning of the word is articulated. This qualitative variable has 11 categories, such as (glottal) stop and (glottal) fricative.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

- $x_7$: Manner of articulation of final consonants. Similar to x6, this refers to how the final consonant is articulated. This is a qualitative variable with 13 categories, such as (glottal) stop and (glottal) fricative,
- $x_8$: Voicing of initial consonants. This variable refers to whether the initial consonant is voiced or voiceless. This qualitative variable has three categories: voiced, voiced (being vowel), and voiceless.
- $x_9$: Voicing of final consonants. Similar to x8, this refers to the voicing of the final consonant. This qualitative variable has three categories: voiced, voiced (being vowel), and voiceless.

## 4. Machine Learning Approach

The machine learning approaches used in this paper are classification, a supervised learning technique, and clustering, an unsupervised learning technique.

Classification leverages attributes to assign predefined class labels to data points. Its goal is to create a mapping from input features to class labels using labeled training data, allowing the model to predict class labels for previously unseen cases. In our study, classification aids in identifying patterns or characteristics that differentiate male names from female names using labeled data. This enables the development of a model capable of predicting the gender of new names based on the features learned from the training data.

In contrast, clustering involves organizing similar data points into clusters based on shared traits or attributes. The main objective of clustering is to uncover hidden patterns or structures in the data without prior knowledge of class labels. In this case, names are grouped according to their similarities without relying on labeled data. This approach can reveal natural groupings or patterns that may not be immediately apparent. For instance, clustering could reveal subgroups within male or female names that share specific phonetic or structural features.

By integrating supervised and unsupervised learning techniques, the advantages of each method are harnessed to achieve a more thorough analysis of the first names in the study. Employing both approaches enables the identification of patterns that may be missed when using only one method, ensuring that the insights are both comprehensive and accurate.

### *4.1. Classification*

The objective is to determine whether it is possible to predict a subject's gender based on all the available variables. Since the 'gender' variable is qualitative, this represents a classification problem. Various machine learning algorithms can be applied to address this type of challenge.

In machine learning, it is customary to split the data into training, validation, and test sets. The training set is used to fit the model parameters. The validation set is used to tune the model hyperparameters. The test set is used to evaluate the model performance.

The confusion matrix assesses how well a classification model fits a dataset. The confusion matrix summarizes the performance of a classification algorithm as follows:

**Table 1.**

| | **Event** | **No-event** |
|---|---|---|
| Event | True positive (TP) | False positive (FP) |
| No-event | False negative (FN) | True negative (TN) |

- "True Positive" are the correctly predicted event values.
- "False Positive" are the incorrectly predicted event values.
- "True Negative" are the correctly predicted no-event values.
- "False Negative" are incorrectly predicted no-event values.

Using the confusion matrix, the following metrics are usually calculated:

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

- Sensitivity (also known as "recall" or the "true positive rate"): The probability that the model predicts a positive outcome for an observation when the outcome is indeed positive.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Specificity (also known as the "true negative rate"): The probability that the model predicts a negative outcome for an observation when the outcome is indeed negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- The precision measures the accuracy of the positive predictions. It answers the question: Of all instances that the model predicted as positive, how many were actually positive?

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

- The accuracy is the percentage of total correct classifications. It measures the proportion of correct predictions out of all predictions made by the model.

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

We will use the following classification algorithms: logistic regression, k-nearest neighbors, neural networks, support vector machine, decision trees, random forests, and gradient boosting machine.

*4.1.1. Logistic Regression*

Logistic regression is a statistical and machine-learning method used for binary classification tasks. The aim is to predict one of two possible outcomes. Despite its name, logistic regression functions as a classification algorithm rather than a regression technique.

Logistic regression is widely used in various applications, including medical diagnosis (e.g., predicting the presence of a disease), spam detection (e.g., classifying emails as spam or not spam), marketing (e.g., predicting whether a customer will buy a product), and credit scoring (e.g., assessing the risk of loan default).

In this study, 70% of the data is used for training, while the remaining 30% is used for testing. Logistic regression produces the following confusion matrix is obtained:

**Table 2.**

| | **Reference** | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 48 | 28 |
| Male | 17 | 94 |

The technique's Accuracy was 0.7594. This means that the proportion of correctly predicted instances (True positives and true negatives) out of the total number of instances is 75.94%, which is relatively high.

*4.1.2. K-Nearest Neighbors (k-NN)*

k-Nearest Neighbors (k-NN) is a straightforward, non-parametric, instance-based learning algorithm used for both classification and regression in machine learning. The core principle of k-NN is that similar data points tend to have similar labels or target values. It uses a distance metric to assess the similarity between data points, with standard metrics including Euclidean distance, Manhattan distance, and Minkowski distance. The "k" in k-NN refers to the number of nearest neighbors considered when making a prediction.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

Among applications of k-NN, we cite pattern recognition (handwriting, fingerprint, and facial recognition), recommendation systems (suggesting products or content based on user similarity), medical diagnosis (predicting diseases based on patient similarity), and anomaly detection (identifying unusual patterns that do not conform to expected behavior).
Using the 3-NN classifier, the following confusion matrix is obtained:

**Table 3.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 106 | 25 |
| Male | 19 | 47 |

The accuracy is found to be Accuracy = 0.7766, which indicates that approximately 77.66% of the predictions were correct. This is slightly better than the logistic regression classifier.

*4.1.3. Neural Networks*

A neural network classifier is a machine learning model inspired by the structure and function of the human brain, developed to identify patterns and categorize data. Neural networks comprise layers of interconnected nodes (neurons) that process input data, learn from it, and generate predictions.

Applications of neural networks include image classification (recognizing objects in images), natural language processing (text classification, sentiment analysis, language translation), speech recognition (converting spoken language into text), medical diagnosis (identifying diseases from medical images or patient data), and finance (fraud detection, stock price prediction).
The neural network classifier produced the following confusion matrix:

**Table 4.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 100 | 25 |
| Male | 24 | 48 |

The classifier's accuracy is 0.7513, which is slightly lower than that of the logistic regression and k-NN classifiers.

*4.1.4. Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a robust supervised learning algorithm employed for both classification and regression tasks. It excels in high-dimensional spaces and can effectively manage both linear and non-linear data. SVM utilizes hyperplanes as decision boundaries to separate different classes within the feature space.

SVM applications include text classification (spam detection, sentiment analysis), image classification (object recognition in images), bioinformatics (protein classification, cancer detection), and financial analysis (Credit scoring, stock price prediction).
The confusion matrix obtained by the SVM classifier is as follows:

**Table 5.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 111 | 38 |
| Male | 12 | 34 |

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

The accuracy of SVM is Accuracy = 0.7436, which is the lowest of the classifiers' accuracies considered so far.

*4.1.5. Naïve Bayes*

Naive Bayes is a group of straightforward yet effective probabilistic classifiers that apply Bayes' theorem with strong (naive) independence assumptions between features. Despite its simplicity, Naive Bayes often performs remarkably well on complex real-world problems, particularly when dealing with large datasets.

Real-world applications of the naïve Bayes classifier have been reported in text classification (spam detection, sentiment analysis, document categorization), medical diagnosis (predicting diseases based on symptoms), recommendation systems (predicting user preferences), and market basket analysis (Understanding customer purchase patterns).

Application of the naïve Bayes classifier resulted in the following confusion matrix:

**Table 6.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 94 | 29 |
| Male | 31 | 43 |

The accuracy obtained Accuracy = 0.6954 is lower than that of the previous classifiers but still very close.

*4.1.6. Decision Tree*

A decision tree classifier is a supervised learning algorithm used for both classification and regression tasks. It employs a tree-like structure to represent decisions and their potential outcomes. The tree is built by recursively dividing the training data into subsets according to the feature values that provide the most significant information gain or the greatest reduction in impurity.

Applications of decision trees include classification (spam detection, medical diagnosis, loan approval), regression (predicting house prices, forecasting sales), and feature engineering (used in ensemble methods like random forests and gradient boosting).

The naïve Bayes classifier yielded the following confusion matrix:

**Table 7.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 111 | 14 |
| Male | 32 | 40 |

The accuracy of the classifier is Accuracy = 0.7665.

*4.1.7. Random Forest*

Random Forest is an ensemble learning technique used mainly for classification and regression tasks. It constructs multiple decision trees during training and combines their results to achieve more accurate and stable predictions. This approach, known as "bagging" (Bootstrap Aggregating), enhances the model's performance and robustness by leveraging the collective strength of multiple trees.

Applications of random forests include classification (spam detection, sentiment analysis, disease diagnosis), regression (predicting house prices, stock market forecasting, sales prediction), feature selection (identifying important features in large datasets), and anomaly detection (detecting outliers and unusual patterns in data).

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

The random forest classifier gave the following confusion matrix:

**Table 8.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 93 | 21 |
| Male | 19 | 44 |

The accuracy of the classifier is Accuracy = 0.7740.

*4.1.8. Gradient Boosting Machine (GBM)*

Gradient Boosting Machine (GBM) is an ensemble machine learning technique utilized for both classification and regression tasks. It develops a model incrementally by sequentially adding predictors, often decision trees, where each new predictor corrects the errors of the previous ones. GBM merges the strengths of several weak models (typically shallow decision trees) to form a robust predictive model.

GBM applications include regression (predicting house prices, stock prices, and other continuous outcomes), classification (spam detection, fraud detection, customer churn prediction), and ranking (search engine result ranking, recommendation systems).

The confusion matrix obtained by the GBM classifier is:

**Table 9.**

| | Reference | |
|---|---|---|
| **Prediction** | **Female** | **Male** |
| Female | 111 | 22 |
| Male | 14 | 50 |

The obtained accuracy of the classifier Accuracy = 0.8173 is the highest accuracy registered among all the classifiers considered.

*4.1.9. Comparison of Classifications Techniques*

A comparison of the results of the classification techniques used in this paper is presented in the table below:

**Table 10.**

| Method | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 score |
|---|---|---|---|---|---|---|
| Logistic regression | 0.7594 | 0.7385 | 0.7705 | 0.6316 | 0.7385 | 0.6809 |
| k-NN | 0.7766 | 0.8480 | 0.6528 | 0.8092 | 0.8480 | 0.8281 |
| Neural network | 0.7513 | 0.8065 | 0.6575 | 0.8000 | 0.8065 | 0.8032 |
| SVM | 0.7436 | 0.9024 | 0.4722 | 0.7450 | 0.9024 | 0.8162 |
| Naïve bayes | 0.6954 | 0.7520 | 0.5972 | 0.7642 | 0.7520 | 0.7581 |
| Decision tree | 0.7665 | 0.7762 | 0.7407 | 0.8880 | 0.7762 | 0.8284 |
| Random forest | 0.7740 | 0.8304 | 0.6769 | 0.8158 | 0.8304 | 0.8230 |
| GBM | 0.8173 | 0.8346 | 0.7813 | 0.8880 | 0.8346 | 0.8605 |

These metrics provide insights into the performance of each model. For example:

- Best Overall Performance: GBM achieves the highest accuracy (81.73%) and a good balance between precision and recall (F1 score of 86.05%).

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

- High Sensitivity: SVM and k-NN have high sensitivity (ability to find positive cases), but SVM sacrifices specificity. This indicates that SVM is good at identifying true positives but struggles with true negatives.
- Balanced Models: Naïve Bayes, Decision Tree, and Random Forest show balanced performance across metrics.

We will examine the BGM classifier more closely since it achieves the highest prediction accuracy. First, Figure 1 is a visual representation of the confusion matrix.

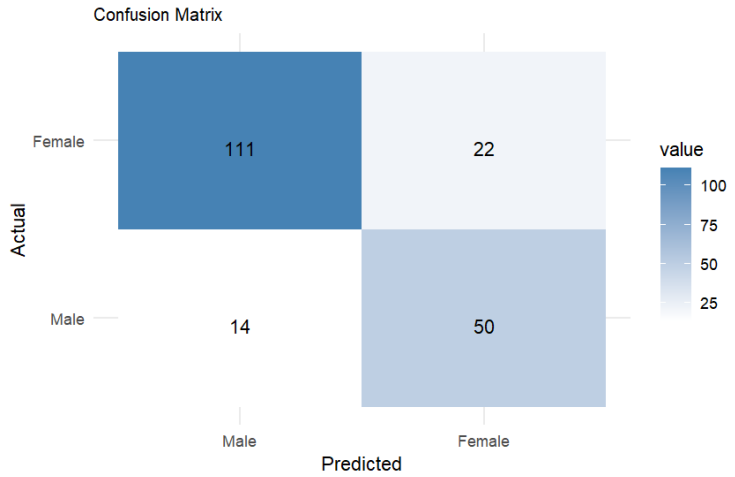


**Figure 1.**
BGM confusion matrix.

Second, an easy way to visualize the two metrics (sensitivity and specificity) is by creating a Receiver Operating Characteristic (ROC) curve, which is a plot that displays the sensitivity and specificity of a classification model, see Figure 2.

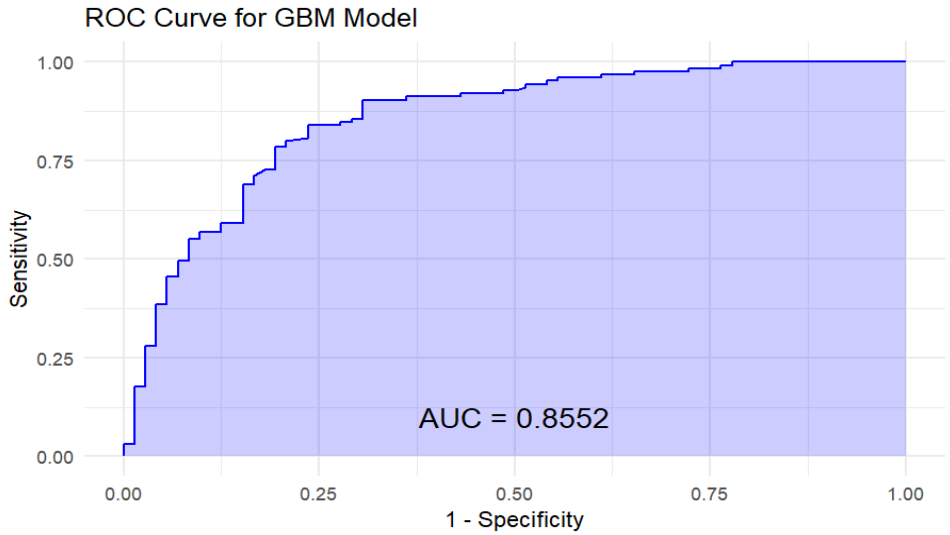


**Figure 2.**
BGM ROC Curve.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

The ROC curve displays the percentage of true positives predicted by the model as the prediction probability cutoff is lowered from 1 to 0. The higher the AUC (area under the curve), the more accurately our model can predict outcomes. The GBM classifier achieves an AUC of 0.8552.

Third, in GBM, the importance of variables (also known as feature importance or influence) can be calculated based on the improvement in the model's performance that each variable brings. Different methods can be used to measure variable importance in GBM. The variable importance in GBM models reflects how much each variable contributes to reducing the model's loss function, normalized to provide a relative measure of importance. In our case, the importance of the model variables is depicted in Figure 3.

**Table 11.**

| Variable | Importance |
|---|---|
| x5 | 52.2263 |
| x3 | 18.5586 |
| x7 | 13.5678 |
| x9 | 12.4236 |
| x4 | 2.3400 |
| x6 | 0.3819 |
| x1 | 0.2782 |
| x2 | 0.2237 |
| x8 | 0.0000 |

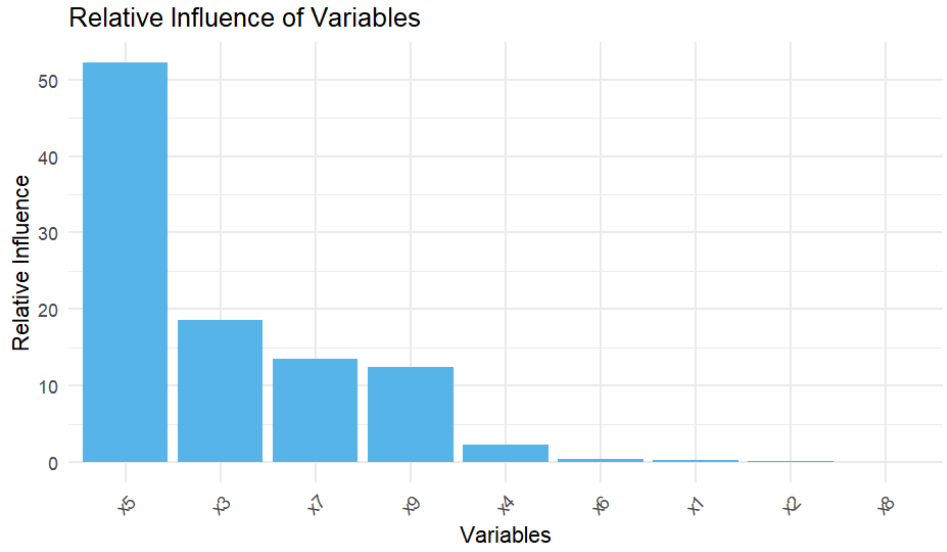


**Figure 3.**
Variables importance.

These values indicate the relative importance of each predictor variable in the model. It seems like x5 (word-initial sound), x3 (syllable structure), x7 (manner of articulation of final consonants), and x9 (voicing of final consonants) are the most important variables in predicting the outcome, while x8 (Voicing of initial consonants) has zero importance. This information can help us understand which variables contribute most to our model's predictions.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

*4.2. Clustering*

Approaching the problem using clustering instead of classification, the goal now shifts from predicting the specific label (in this case, "gender") to discovering natural groupings or patterns within the data without using predefined gender labels.

To evaluate how well the clusters match the actual gender labels, we can analyze the confusion matrix and calculate metrics like accuracy, precision, recall, and F1 score. Here is how you can interpret the results.

Note that the gender variable is excluded from the clustering process. Clustering aims to group data based on inherent patterns without considering any labels. Including gender in the clustering might influence the results, leading to biased clusters. Instead, we exclude gender during the clustering process and use it later to evaluate how well the clusters align with the actual genders.

We will use the following clustering algorithms: k-means, hierarchical clustering, and DBSCAN.

*4.2.1. k-Means*

k-means clustering is a well-known and extensively used unsupervised machine learning algorithm that divides a dataset into a predetermined number of clusters, represented by "k." The main objective of k-means clustering is to group data points so that those within the same cluster are more similar to each other than to those in different clusters.

k-means clustering has a wide range of applications across various fields due to its versatility and simplicity. Some notable examples include customer segmentation, image compression, document clustering, genomics, and anomaly detection.

Using the 2-means clustering technique yields the following confusion matrix and clusters:

**Table 12.**

| | Gender | |
|---|---|---|
| **Cluster** | **Female** | **Male** |
| 1 | 401 | 225 |
| 2 | 18 | 12 |

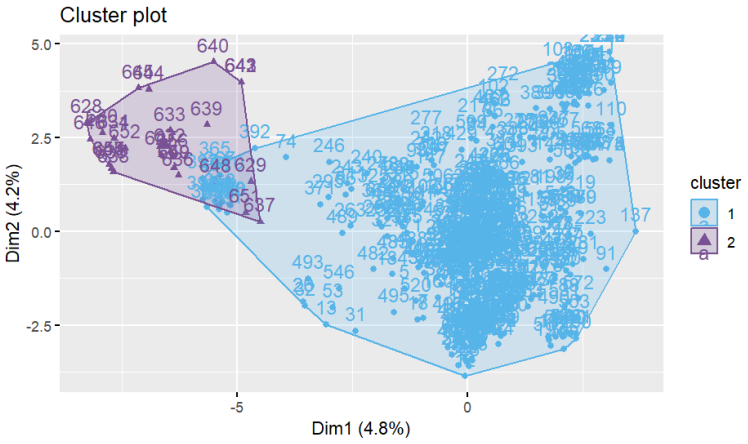


**Figure 4.**
k-means clusters.

Cluster 1 excels at identifying females, with a high recall rate of 95.69% and moderate precision of 64.09%. The F1 score of 76.60% suggests a fair balance between precision and recall. In contrast,

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

Cluster 2 performs poorly in identifying males, with very low precision (40%) and recall (5.06%), resulting in a low F1 score of 8.93%.

The clustering model exhibits a strong gender bias, favoring females. Cluster 1 effectively captures most females but also misclassifies many males. Meanwhile, Cluster 2, intended to represent males, shows poor performance. This suggests that the clustering model may not effectively differentiate between genders in this dataset and might need further adjustment or an alternative approach.

*4.2.2.* Hierarchical Clustering

Hierarchical clustering is an unsupervised machine-learning method used to organize data points into clusters based on their similarities. It generates a hierarchical or tree-like diagram known as a dendrogram, which illustrates the nested grouping of data points. The similarity between clusters is assessed using different linkage criteria.

Hierarchical clustering has a wide range of applications across various domains, such as bioinformatics, market research, document clustering, social network analysis, and ecology.

Using the hierarchical clustering technique yields the following confusion matrix and clusters:

**Table 13.**

| | Gender | |
|---|---|---|
| **Cluster** | **Female** | **Male** |
| 1 | 398 | 200 |
| 2 | 21 | 37 |

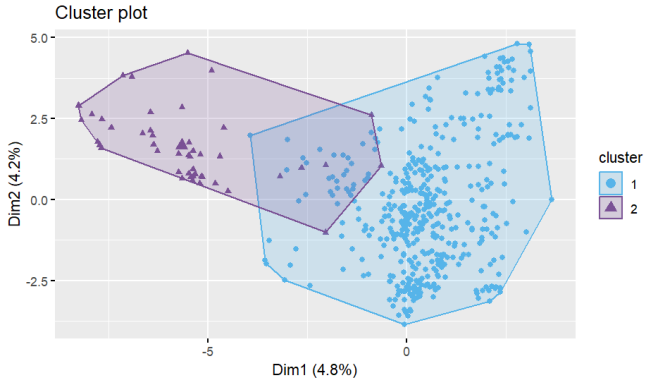


**Figure 5.**
Hierarchical clusters.

Cluster 1 is more proficient at identifying females, demonstrating high precision (95.1%) and moderate recall (66.6%). The F1 score of 79.0% indicates a strong balance between precision and recall for females. Conversely, Cluster 2 performs poorly in identifying males, with very low precision (15.6%) and recall (15.6%), leading to a low F1 score of 15.6%.

Hierarchical clustering exhibits a notable bias towards identifying females. Cluster 1 successfully captures most females but also misclassifies a substantial number of males. Cluster 2, intended to represent males, performs poorly, suggesting that the model may not be effective in distinguishing between genders in this dataset. This implies the need for further refinement of the clustering approach or exploring alternative methods to enhance gender classification.

*4.2.3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely used unsupervised machine learning algorithm designed for clustering tasks, especially with datasets that

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

include noise and varying densities. It clusters data points based on their density, identifying regions with high point density as clusters and considering points in sparser areas as noise or outliers.

One of DBSCAN's key advantages is its ability to find clusters of arbitrary shapes, making it particularly effective for complex datasets where clusters are not well-defined or are non-spherical. It is also robust to noise, as it inherently differentiates between dense clusters and sparse, noisy regions.

DBSCAN is widely used in applications such as geographic information systems (GIS), anomaly detection, astronomy, image processing, and social network analysis.

Using the DBSCAN clustering technique yields the following confusion matrix and clusters:

**Table 14.**

| | Gender | |
|---|---|---|
| **Cluster** | **Female** | **Male** |
| 0 | 381 | 232 |
| 1 | 5 | 1 |
| 2 | 5 | 0 |
| 3 | 5 | 0 |
| 4 | 5 | 1 |
| 5 | 9 | 1 |
| 6 | 4 | 2 |
| 7 | 5 | 0 |

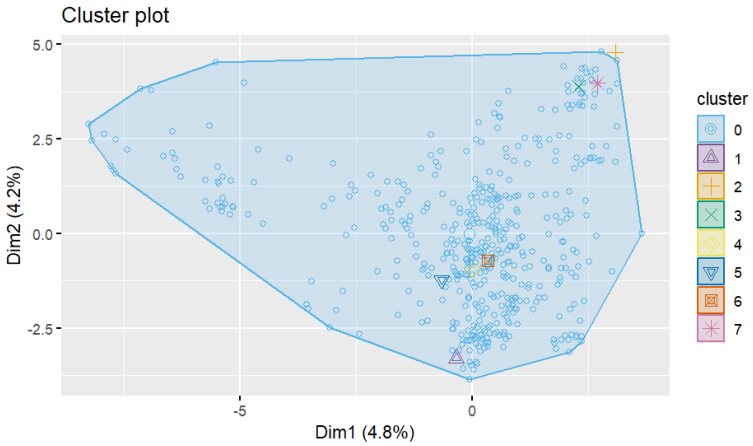


**Figure 6.**
DBSCAN clusters.

Cluster 0 predominantly identifies females, achieving high recall (90.93%) but moderate precision (62.17%). The F1 score of 73.74% reflects a reasonable balance between precision and recall. Clusters 1, 4, and 6 are intended to identify males but perform poorly, with extremely low precision, recall, and F1 scores. Clusters 1 and 4 each have a precision of 16.67%, while Cluster 6 has a slightly higher precision of 33.33%. However, their recall values are very low (approximately 0.42% for Clusters 1 and 4, and 0.84% for Cluster 6), resulting in very low F1 scores.

The DBSCAN clustering model, similar to previous methods, shows a strong bias toward identifying females, with Cluster 0 successfully capturing most females. However, it fails to accurately identify males, as the clusters designated for males (Clusters 1, 4, and 6) exhibit extremely low precision, recall, and F1 scores. This suggests that DBSCAN, given the current parameters, may not be effective in distinguishing between genders in this dataset and may require further adjustment or a different approach for improved results.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

*4.2.4. Comparison of Clustering Techniques*

In clustering, metrics for males and females are usually reported separately because the algorithm cannot access the true labels during the clustering process. Instead, it organizes data based on similarities, and we assess how well these clusters correspond to the actual gender labels afterward. Reporting metrics separately provides a clearer view of how effectively the clustering identifies each gender.

A comparison of the results of the clustering techniques used in this paper is presented in the table below:

**Table 15.**

| | | Female | | | Male | | |
|---|---|---|---|---|---|---|---|
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1 score** | **Precision** | **Recall** | **F1 score** |
| k-means | 0.6296 | 0.6409 | 0.9569 | 0.7660 | 0.4000 | 0.0506 | 0.0893 |
| Hierarchical | 0.6631 | 0.6656 | 0.9499 | 0.7827 | 0.1561 | 0.1561 | 0.1561 |
| DBSCAN | 0.5823 | 0.6215 | 0.9093 | 0.7384 | 0.2222 | 0.0168 | 0.0313 |

These metrics provide insights into the performance of each model. For example:

- Hierarchical clustering generally performs the best across all metrics, especially in terms of accuracy and F1 score for females. However, like the other methods, it struggles with identifying males, suggesting that the clustering algorithm is more biased towards capturing females accurately.
- k-Means shows a reasonable balance but has the highest precision for males at the cost of very low recall, leading to a poor F1 score for males.
- DBSCAN underperforms in general, with the lowest accuracy and a very low recall for males, though it still maintains a decent recall for females.

Hierarchical clustering appears to be the most effective method for this dataset, especially for accurately identifying females. However, none of the methods excel in identifying males, suggesting that the data may need further refinement or that an alternative clustering approach might be necessary to enhance male identification.

Given that hierarchical clustering performed best among the methods evaluated, a more thorough assessment was conducted by calculating additional clustering metrics: the Silhouette Score, Within-Cluster Sum of Squares (WCSS), and Davies-Bouldin Index (DBI). These metrics provide a more comprehensive evaluation of the clustering model's performance, helping to determine whether hierarchical clustering is the optimal choice or if another method might offer better results.

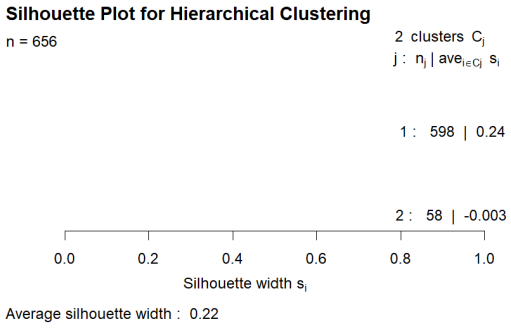


**Figure 7.**
Silhouette plot.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

The silhouette plot (Figure 7) illustrates the performance of the hierarchical clustering algorithm, with each bar representing a sample. The silhouette width, ranging from -1 to 1, reflects how well each sample fits within its cluster. Cluster 1, which contains 598 samples, has an average silhouette width of 0.24, indicating relatively good clustering. In contrast, Cluster 2, with 58 samples, has a very low silhouette width of approximately -0.003, suggesting poor clustering, potential boundary issues, or misclassification. The overall average silhouette width of 0.22 indicates moderate clustering quality, with notable overlap between clusters, especially in Cluster 2.

The silhouette plot suggests that hierarchical clustering has moderately separated the data, but further refinement of parameters (such as distance metric and linkage method) or the number of clusters might improve the results. Given the low silhouette score for Cluster 2, exploring alternative clustering methods may also be beneficial.

The Silhouette Score, with an average of 0.216, reveals that the clusters are not well-defined and show considerable overlap.

The Within-Cluster Sum of Squares (WCSS), which measures cluster compactness, stands at 67,863.81—a relatively high value—indicating that the clusters are not tightly packed. This supports the low silhouette score and suggests suboptimal clustering.

The Davies-Bouldin Index (DBI), which assesses cluster separation, is 3.17, indicating poor separation and reinforcing the notion that the clusters are not well-defined.

Therefore, these metrics indicate that hierarchical clustering may not be the most suitable method for this dataset, as both cluster cohesion and separation are inadequate. Further exploration of alternative methods or adjustments to clustering parameters may lead to improved results.

## 5. Comparative Analysis Classification vs Clustering

Classification methods significantly outperform clustering methods in terms of accuracy and balanced performance across genders. For example, the GBM model demonstrates high accuracy and F1 scores for both genders, highlighting its effectiveness in classifying data points with clear labels. In contrast, clustering methods struggle with accurately identifying males because they are not designed to optimize for specific class labels. As a result, clustering methods generally perform lower, as they focus on finding natural groupings in the data without regard to labels.

Clearly, classification is the more suitable approach for gender classification in this dataset. The models used are specifically trained to differentiate between classes, resulting in higher accuracy and more balanced performance across genders. While clustering is valuable for exploratory data analysis, it is less effective for this task, particularly because it does not directly optimize for accurate identification of males and females.

Given the results, using GBM or other robust classification models such as Random Forest or k-NN is recommended for gender classification in this dataset. These methods offer the best accuracy, precision, and F1 scores, providing strong performance for both male and female classifications. Clustering can complement classification for exploratory purposes or to gain insights into the data structure but should not be relied upon as the primary method for gender classification.

## 6. Conclusion

In this study, we assessed various machine learning techniques for gender classification by comparing the effectiveness of classification and clustering methods on a dataset of first names. Our analysis reveals that classification methods—specifically Gradient Boosting Machine (GBM), Random Forest, and k-Nearest Neighbors (k-NN)—outperform clustering methods significantly in terms of accuracy and balanced performance across genders. Notably, the GBM model exhibited high accuracy and strong F1 scores for both male and female names, demonstrating its proficiency with clearly labeled data.

In contrast, clustering methods such as hierarchical clustering and DBSCAN faced significant difficulties in accurately identifying males. These challenges arise from their design, which emphasizes

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*

grouping based on similarity without optimizing for specific class labels. Metrics such as the silhouette plot, Silhouette Score, Within-Cluster Sum of Squares (WCSS), and Davies-Bouldin Index (DBI) collectively showed that clustering methods struggled to effectively distinguish between genders, particularly for males.

Our findings suggest that classification techniques are more appropriate for gender classification in this scenario, as they are designed to differentiate between distinct classes, leading to better performance. Although clustering methods can be valuable for exploratory data analysis and understanding data structures, they are not recommended as the primary approach for gender classification.

In conclusion, robust classification models like GBM, Random Forest, or k-NN are recommended for gender classification tasks involving this dataset to achieve accurate and balanced results. While clustering methods may be useful for initial exploration, they should not replace classification techniques where precise gender identification is essential. Future research could focus on refining clustering parameters or exploring alternative approaches to enhance gender classification performance and address the limitations identified in this study.

## Funding:
This research did not receive any specific funding from any source.



## References
[1] Anderson, J., Ewen, C., & Staun, J. (2008). Phonological structure: Segmental, suprasegmental and extrasegmental. Cambridge University Press.
[2] https://doi.org/10.1017/S0954394508000272
[3] Al Tamimi, Y. and Smith, M. (2023). Phonological Features of Saudi Arabian Anthroponyms, Arab World English Journal, Vol. 14, No. 1, 486-501.
[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[5] Cao, Y. T., & Daumé III, H. (2021). Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. Computational Linguistics, 47(3), 615-661.
[6] Cho, H. (2024). Gender classification of Korean personal names: Deep neural networks versus human judgments. Lingua, 303, 103703.
[7] Chun Hau Ngai, A. J., Kilpatrick, A. J., & Ćwiek, A. (2024). Sound symbolism in Japanese names: Machine learning approaches to gender classification. PLOS ONE.
[8] Fredrickson, A. (2007). Phonological cues to gender in sex-typed and unisex names. Available at: www.swarthmore.edu/SocSci/Linguistics/Papers/2007/fredrickson_annie.pdf.
[9] Gupta, R., Chaspari, T., Kim, J., Kumar, N., Bone, D., & Narayanan, S. (2016). Pathological speech processing: State-of-the-art, current challenges, and future directions. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 6470-6474.
[10] Higgins, D., & Sadock, J. M. (2003). A Machine Learning Approach to Modeling Scope Preferences. Computational Linguistics, 29(1), 73-96.
[11] Jakobson, R. (1965). Quest for the Essence of Language. Diogenes, 13(51), 21–37.
[12] Jespersen, O. (1922). *Language: Its Nature, Development, and Origin*. London: George Allen & Unwin.
[13] Jia, J., & Zhao, Q. (2019). Gender Prediction Based on Chinese Name. In: Tang, J., Kan, M.Y., Zhao, D., Li, S., & Zan, H. (eds) Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science, 11839. Springer, Cham.
[14] Kilpatrick, A. J. (2023). Sound Symbolism in Automatic Emotion Recognition and Sentiment Analysis. Proceedings Cognitive AI 2023, Bari, Italy.
[15] Kilpatrick, A. J., Ćwiek, A., & Kawahara, S. (2023). Random forests, sound symbolism and Pokémon evolution. PLOS ONE.

[16] Lee, M-S., Lee, G-E., Lee, S. H., & Lee, J-H. (2024). Emotional responses of Korean and Chinese women to Hangul phonemes to the gender of an artificial intelligence voice. Frontiers in Psychology, 15.
[17] Manning, C. D. (2015). Computational Linguistics and Deep Learning. Computational Linguistics, 41(4), 701-707.
[18] Moorthy, S., Pogacar, R., Khan, S., & Xu, Y. (2018). Is Nike female? Exploring the role of sound symbolism in predicting brand name gender. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 1128-1132.
[19] Ohala, J.J. (1994). *The Frequency Code underlies the sound-symbolic use of voice pitch. In Sound Symbolism* (pp. 325-347). Cambridge University Press.
[20] Piotrowska, M., Korvel, G., Kostek, B., Ciszewski, T., & Czyzewski, A. (2019). Machine Learning-Based Analysis of English Lateral Allophones. International Journal of Applied Mathematics and Computer Science, 29(2), 393-405.
[21] Plato. Cratylus. 360 BCE
[22] Raahul, A., Sapthagiri, R., Pankaj, K., & Vijayarajan, V. (2017). Voice-based gender classification using machine learning. IOP Conference Series: Materials Science and Engineering, 263(4), 042083.
[23] Shahin, M., Zafar, U., & Ahmed, B. (2020). The Automatic Detection of Speech Disorders in Children: Challenges, Opportunities, and Preliminary Results. IEEE Journal of Selected Topics in Signal Processing, 14(2), 400-412.
[24] Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022). Machine learning for stuttering identification: Review, challenges and future directions. Neurocomputing, 514, 385-402.
[25] Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4), 521-544.
[26] Tripathi, A., & Faruqui, M. (2011). Gender prediction of Indian names. IEEE Technology Students' Symposium, Kharagpur, India, 137-141.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6480-6497, 2024*
*DOI: 10.55214/25768484.v8i6.3402*