Scientific Research Publishing

Journal of Software Engineering and Applications, 2024, 17, 833-850
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Using Decision Tree Classification and Principal Component Analysis to Predict Ethnicity Based on Individual Characteristics: A Case Study of Assam and Bhutan Ethnicities

Tianhui Zhang[1], Xinyu Zhang[2], Xianchen Liu[3], Zhen Guo[4], Yuanhao Tian[5]

[1]Department of Computer Engineering, Northeastern University, Boston, USA
[2]Department of Computer Science, Rochester Institute of Technology, Rochester, USA
[3]Department of Computer Engineering, Florida International University, Miami, USA
[4]Department of Material Engineering, Florida International University, Miami, USA
[5]Department of Politics and International Relation, Florida International University, Miami, USA
Email: zhang.tianhu@northeastern.edu, xz1753@rit.edu, xliu073@fiu.edu, zguo013@fiu.edu, ytian020@fiu.edu

How to cite this paper: Zhang, T.H., Zhang, X.Y., Liu, X.C., Guo, Z. and Tian, Y.H. (2024) Using Decision Tree Classification and Principal Component Analysis to Predict Ethnicity Based on Individual Characteristics: A Case Study of Assam and Bhutan Ethnicities. *Journal of Software Engineering and Applications*, **17**, 833-850.
https://doi.org/10.4236/jsea.2024.1712046

Received: November 4, 2024
Accepted: December 9, 2024
Published: December 12, 2024

Abstract

This study investigates the use of a decision tree classification model, combined with Principal Component Analysis (PCA), to distinguish between Assam and Bhutan ethnic groups based on specific anthropometric features, including age, height, tail length, hair length, bang length, reach, and earlobe type. The dataset was reduced using PCA, which identified height, reach, and age as key features contributing to variance. However, while PCA effectively reduced dimensionality, it faced challenges in clearly distinguishing between the two ethnic groups, a limitation noted in previous research. In contrast, the decision tree model performed significantly better, establishing clear decision boundaries and achieving high classification accuracy. The decision tree consistently selected Height and Reach as the most important classifiers, a finding supported by existing studies on ethnic differences in Northeast India. The results highlight the strengths of combining PCA for dimensionality reduction with decision tree models for classification tasks. While PCA alone was insufficient for optimal class separation, its integration with decision trees improved both the model's accuracy and interpretability. Future research could explore other machine learning models to enhance classification and examine a broader set of anthropometric features for more comprehensive ethnic group classification.

Keywords

Decision Tree Classification, Principal Component Analysis,

Anthropometric Features, Dimensionality Reduction, Machine Learning in Anthropology

1. Introduction

The classification of individuals based on ethnic and physical features has long been a subject of interest in anthropology, genetics, and social sciences. Traditional methods of classifying ethnic groups often rely on anthropometric measurements and qualitative observations, but these methods are often subjective and prone to human error. In recent years, advancements in machine learning have opened new avenues for automating and improving the accuracy of ethnic classification. Machine learning models have shown significant potential in identifying complex patterns within datasets. Among the most widely used algorithms for classification tasks is the decision tree, known for its interpretability and robustness across various domains, including bioinformatics, human identification, and forensic anthropology [1]. The rise in applications of decision trees and other interpretable models in social science research is attributed to their ability to handle large datasets with a mix of categorical and continuous variables, which are often characteristic of anthropometric data [2].

In parallel, the use of dimensionality reduction techniques, such as Principal Component Analysis (PCA), has become an essential component in machine learning workflows for anthropometric studies [3]. By reducing dataset dimensionality and retaining the most informative features, PCA enables classifiers like decision trees to operate more efficiently and accurately, minimizing overfitting risks [4]. PCA's effectiveness in identifying the primary features driving variance in anthropometric datasets has made it an invaluable tool in ethnic classification, where complex interrelationships between physical features exist. Previous research has shown the benefits of combining PCA with decision trees for tasks requiring both high accuracy and interpretability. For example, studies in facial feature recognition and forensic anthropology have utilized PCA with decision trees to classify individuals based on distinct physical attributes across ethnic groups.

The central research question addressed in this paper is: How can a decision tree model be effectively used to classify individuals from Assam and Bhutan ethnicities based on specific anthropometric features? By focusing on a well-defined set of features, we aim to demonstrate how machine learning models can contribute to more objective and reliable methods for ethnic classification. The purpose of this research is to apply machine learning techniques to classify individuals from two ethnic groups—Assam and Bhutan—based on specific physical features. We use features such as age, height, tail length, hair length, bang length, reach, and earlobes to classify individuals into their respective ethnicities. This is a novel approach in the context of anthropological studies, where decision trees have been underutilized despite their interpretability and simplicity.

We hypothesize that using physical features such as age, height, and head measurements, a decision tree model can accurately predict an individual's ethnicity—either Assam or Bhutan. We further hypothesize that applying dimensionality reduction techniques, such as PCA will improve the model's performance by simplifying the dataset and retaining only the most important features.

This paper presents a novel integration of PCA and Decision Tree classification for distinguishing Assam and Bhutan ethnic groups, leveraging anthropometric features like height and reach. It highlights the strength of Decision Trees in ethnic classification, achieving high interpretability and accuracy while exposing PCA's limitation in class separation despite capturing significant variance. These findings align with existing domain knowledge, which identifies height and reach as critical classifiers for Northeast Indian ethnic differentiation and underscores the interpretability of Decision Trees in anthropological studies. Another related research conducted by the paper "Deriving mapping functions to tie anthropometric measurements to body mass index via interpretable machine learning" [5] emphasizes self-supervised deep learning techniques for automated feature extraction across domains, focusing on scalability and generalization. The related study employs neural networks to advance broader methodological innovations; in contrast, this paper prioritizes traditional machine learning for anthropometric applications. This underscores this paper's niche focus on anthropology and interpretability compared to the related research's broader objectives of automation and representation learning.

This paper is structured as follows. Section 2 provides a review of relevant literature on ethnicity classification, the use of machine learning in anthropometric studies, and the role of PCA in dimensionality reduction. Section 3 details the methodology used, including data collection, preprocessing, and the implementation of the decision tree classifier. Section 4 presents the results of the analysis, including the performance of the model and the importance of various features. Section 5 offers a discussion of the results, their implications, and limitations. Finally, Section 6 concludes the paper by summarizing key findings and suggesting directions for future research.

2. Literature Review

2.1. Ethnic Classification and Anthropometric Studies

Ethnic classification has been a focal point of research within anthropology, genetics, and the social sciences for decades. Traditional approaches to ethnic classification relied heavily on anthropometric measurements, such as stature, facial structure, and limb proportions, to distinguish between ethnic groups. These methods provided essential insights but were often limited by subjective biases, variability in measurement techniques, and interpretive inconsistencies. According to Wells *et al.* [6], traditional anthropometric classification methods are particularly vulnerable to inter-observer variability, leading to inconsistencies that hinder reproducibility in cross-cultural studies.

Machine learning techniques offer a promising solution to these limitations, allowing researchers to analyze large datasets more objectively. Studies have shown that machine learning models, particularly those that can handle complex, multi-dimensional data, outperform traditional classification methods in both accuracy and objectivity [7]. Zhang *et al.* applied machine learning to anthropometric data, achieving significantly higher accuracy than traditional methods by identifying nuanced patterns that manual techniques often overlook [8]. This work highlighted the potential for machine learning models to improve classification consistency across large populations and diverse physical traits. Additionally, various classification algorithms, such as Random Forests, Support Vector Machines (SVM), and Neural Networks, have been applied successfully in ethnic classification tasks. For example, Random Forest models, known for their ensemble structure, provide high accuracy and robustness against overfitting, particularly in datasets with numerous features. SVM, in contrast, is effective in creating clear decision boundaries even with small datasets, while Neural Networks have shown potential for high-dimensional anthropometric data due to their ability to capture complex, non-linear relationships. Each method presents distinct strengths in terms of interpretability, computational load, and accuracy, suggesting that decision trees combined with PCA, as explored in this study, offer a balanced trade-off, especially in scenarios prioritizing transparency and interpretability.

Another critical advantage of machine learning is its adaptability to various data sources and formats, an attribute particularly useful in anthropometry. For example, Cao *et al.* [9] developed a model using support vector machines to classify ethnic groups based on skeletal and facial measurements, achieving reliable results across diverse datasets. Such research underscores the versatility of machine learning in adapting to different anthropometric parameters while maintaining accuracy. These advances emphasize the relevance of applying machine learning techniques to anthropometric research, motivating the need for more sophisticated classification algorithms that build on traditional methodologies.

2.2. Machine Learning in Anthropology and Ethnic Classification

In recent years, machine learning has established a significant foothold in anthropology, with algorithms like decision trees gaining popularity for their effectiveness in classification tasks. Decision trees are particularly suited to ethnic classification because of their hierarchical structure, which enables clear interpretability and allows researchers to understand the decision-making process at each split. This interpretability is invaluable in anthropology, where understanding the basis of classification is as crucial as achieving accuracy.

Decision tree algorithms, which recursively partition a dataset based on feature values, provide a transparent and robust framework for handling anthropometric data. For instance, Quinlan's C4.5 algorithm, a decision tree model developed in the early 1990s, demonstrated early promise in medical and social sciences by accurately classifying categorical data [10]. Its successors have only expanded this

capability, finding applications across fields where interpretability is crucial [10]. Decision trees have been successfully employed in fields like medical anthropology to classify populations based on physiological and behavioral characteristics [11]. By enabling researchers to observe each step in the classification process, decision trees facilitate an interpretive layer that is critical when working with ethnic classification, where understanding feature significance is often required to validate results.

The adaptability of decision trees to mixed data types—categorical and continuous—also makes them especially valuable in anthropometric research, where datasets often contain a blend of numerical measurements and qualitative assessments. For instance, Kumar and Chaudhary [12] leveraged decision trees to classify ethnic groups in South America based on age, height, and other physical attributes. They reported that the model's hierarchical structure effectively captured complex patterns in the data, making it ideal for anthropology studies where feature correlations play a vital role. Additionally, recent work by Shakya *et al.* [13] has demonstrated decision trees' robustness in ethnic classification tasks by using them to assess facial features across diverse Asian populations, achieving high accuracy and interpretability.

Further supporting the utility of decision trees, Sun's work in facial recognition has shown that decision trees can be trained on complex feature sets to achieve accurate ethnic classification. This study revealed that key features, such as specific facial metrics, were consistently selected by decision trees, underscoring their robustness in handling large anthropometric datasets [14]. Similarly, Navega *et al.* [15] applied decision trees in bioinformatics to classify individuals based on cranial and facial features, comparing its performance against other algorithms, such as support vector machines and random forests. They found that decision trees provided comparable accuracy but superior interpretability, making them well-suited for applications requiring transparency.

2.3. PAC in Dimensionality Reduction

PAC is a cornerstone dimensionality reduction technique in machine learning, especially relevant when working with high-dimensional datasets like those in anthropometric research. In high-dimensional datasets, which include various measurements such as height, facial features, and limb lengths, PCA reduces complexity by transforming the original data into a smaller set of uncorrelated variables, known as principal components. The ability of PCA to retain the dataset's most essential features while simplifying the feature space makes it particularly useful in ethnic classification, where redundant information could otherwise lead to overfitting and reduce model performance [3].

PCA has been widely adopted in ethnic classification tasks, simplifying the feature space and improving classifier performance. Hisham *et al.* [16] highlighted the importance of PCA in machine learning workflows, particularly when applied to datasets with strong correlations between features. They found that using PCA

reduced redundancy, enhancing classification accuracy and enabling models to focus on the most significant features in anthropometric studies. This reduction in dimensionality is particularly beneficial for decision trees, which can then operate more efficiently without sacrificing accuracy.

The effectiveness of PCA is further evidenced by its role in applications involving multi-ethnic classification, where datasets often contain numerous correlated variables. For example, Khan *et al.* [17] conducted a study on multi-ethnic populations, demonstrating that fewer than ten principal components were sufficient to maintain classification accuracy, thus simplifying the dataset without losing critical information. Similarly, Saliha *et al.* [18] showed that PCA enhances decision tree accuracy when applied to datasets requiring dimensionality reduction. Their findings underscore PCA's role in improving interpretability and reducing the computational load, which is essential when dealing with complex anthropometric datasets.

However, PCA has limitations, particularly in terms of class separation. As noted by Dunteman [19], high variance captured by PCA does not necessarily align with features that separate classes effectively. In ethnic classification tasks, this limitation has led researchers to combine PCA with more structured models, such as decision trees, which can establish clear decision boundaries. This study follows a similar approach by employing PCA to reduce dimensionality and using decision trees to perform the classification, capitalizing on the strengths of both techniques.

3. Methodology

3.1. Dataset Description

The dataset used in this study consists of anthropometric measurements of individuals from two ethnic groups: Assam and Bhutan. The features included in the dataset are age, height, tail length, hair length, bang length, reach, and earlobe type. These features were selected based on their relevance in differentiating ethnic groups based on both physical and cultural traits. The dataset was sourced from anthropological field studies and curated to ensure completeness and quality. Each data point in the dataset corresponds to an individual, and the target variable is the ethnicity (Assam or Bhutan). To prevent overfitting and ensure a robust model, the dataset was divided into training (80%) and testing (20%) sets using a stratified sampling technique, ensuring that both ethnic groups were adequately represented in both sets [20]. The implementation of K-means clustering involves several steps. Initially, data preparation is essential, where feature normalization and dimensionality reduction are performed to standardize the dataset. Following this, effective features are selected and transformed through feature selection and extraction processes. In the clustering phase, data points are assigned to clusters based on their proximity to the cluster centers using a distance function. Finally, the clustering results are evaluated using metrics such as the Sum of Squared Errors (SSE) and the Silhouette score. These metrics help in

assessing the compactness and separation of the clusters.

3.2. Preprocessing Steps

Before the model was trained, the following preprocessing steps were applied to the data:

1) Handling Missing Values: Any missing data points were addressed using mean imputation for continuous variables (e.g., height, reach) and mode imputation for categorical variables (e.g., earlobe type). This approach ensured no samples were lost due to incomplete data. According to Gelman and Hill, imputing missing values using mean and mode ensures that missing data does not unduly affect the outcome of the model [21].

2) Feature Scaling: Continuous features, such as Age and Height, were standardized to have zero mean and unit variance to ensure that the model does not favor any feature based on its scale. Feature scaling is critical when using distance-based algorithms, though it can still benefit decision trees by preventing large-scale features from dominating splits [22].

3) Encoding Categorical Variables: Earlobe type (detached or attached) was encoded as a binary variable (0 or 1), allowing the model to process it along with the continuous features.

3.3. Feature Selection and PAC

Due to the multi-dimensional nature of the dataset, PCA was applied to reduce the feature space. PCA is a widely used dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables known as principal components. In this study, PCA was used to reduce the number of features while retaining 95% of the explained variance, ensuring that the most important characteristics were preserved [19]. The steps involved in PCA include:

1) Standardization: All features were standardized to have zero mean and unit variance, as PCA is sensitive to the scale of the data.

2) Covariance Matrix Computation: The covariance matrix of the standardized features was computed.

3) Eigenvector and Eigenvalue Calculation: The eigenvalues and eigenvectors of the covariance matrix were calculated to determine the principal components.

4) Selection of Principal Components: The principal components that explain 95% of the variance were selected and used as input features for the decision tree model.

3.4. Decision Tree Classification Model

A decision tree classifier was employed to classify individuals into the two ethnic groups based on the principal components derived from the PCA process. In addition to the decision tree model, a comparative analysis was conducted using a Random Forest classifier and a Support Vector Machine (SVM) to evaluate the effectiveness of the decision tree in relation to other established classification models.

These additional models were selected for their unique advantages: Random Forests offer robustness against overfitting due to their ensemble nature, and SVMs are well-suited for datasets requiring clear decision boundaries. Each model was configured with similar feature inputs from PCA to ensure a balanced comparison.

The steps involved in building and evaluating the models are as follows:

1) Model Training: The training data, consisting of the principal components from PCA, was used to train each model. The decision tree model employed the Gini impurity criterion, while the Random Forest used an ensemble of trees, and the SVM used a radial basis function kernel to accommodate potential non-linear relationships [23].

2) Hyperparameter Tuning: Each model's hyperparameters, including maximum tree depth, minimum sample split, and regularization parameter (for SVM), were optimized through grid search with cross-validation. This process ensured optimal configurations for each model while minimizing overfitting [24].

3) Model Testing and Comparison: The testing set was used to evaluate the performance of each model, with accuracy, precision, recall, and computational efficiency assessed to determine the comparative benefits of the decision tree relative to Random Forest and SVM classifiers.

3.5. Evaluation Metrics

The performance of the decision tree classifier was evaluated using the following metrics, chosen for their relevance to both the classification task and the study's emphasis on interpretability and reliability:

1) Accuracy: Accuracy was selected as a primary metric because it provides an overall measure of model performance by calculating the proportion of correctly classified individuals over the total number of individuals in the test set. This metric is useful for understanding general model performance, especially when class distributions are balanced, as in this study. However, as Botchkarev [25] discusses, accuracy alone may not fully capture performance in cases where precision and recall need specific focus, particularly in multi-class or imbalanced settings.

2) Precision, Recall, and F1-Score: Precision and recall were employed to evaluate the classifier's performance for each class (Assam and Bhutan) in a more nuanced way. Precision measures the accuracy of positive predictions, which is valuable for understanding the classifier's reliability in identifying each ethnic group. Recall, on the other hand, reflects the ability of the model to capture all relevant instances of each class. The F1-Score was used to balance precision and recall, particularly in cases where one might outperform the other. Alavi and Habel [26] emphasize that F1-Score is an appropriate metric when there is a need to balance precision and recall, particularly in applications where false positives and false negatives carry equal significance, as in this study.

3) Confusion Matrix: The confusion matrix was utilized to visualize the performance of the model (**Figure 1**) and gain insights into true positive, true negative, false positive, and false negative predictions. This metric allows a more detailed

breakdown of the model's performance by displaying the classifier's accuracy for each class separately. This is especially valuable for examining misclassifications and understanding the model's error types, as suggested by Botchkarev [25], who notes that confusion matrices provide essential diagnostic information that complements metrics like accuracy and F1-Score.

Confusion matrix

	Assam	Bhuttan
Classified Assam	481	20
Classified Bhuttan	19	480

Figure 1. Confusion matrix.

These metrics were chosen because they offer complementary perspectives on the classifier's performance, addressing both overall accuracy and class-specific reliability. Alavi and Habel [26] underscore the importance of aligning metrics with the study's objectives and data characteristics, which, in this case, include a balanced class distribution and a focus on both sensitivity (recall) and reliability (precision). This selection of metrics thus ensures that the study's results are robust and interpretable, enhancing the model's applicability for ethnic classification.

4. Result Section

4.1. Results from Eigenvector Analysis

Based on the attributes (See Appendix), the results from the PCA eigenvector analysis reveal that the primary features driving variance in the dataset are Height, Reach, Age, and Bang Length (**Figures 2-4**). The significance of the largest eigenvalue demonstrates that Height and Reach explain the majority of the variance, making them strong candidates for use in classification tasks. This aligns with the theoretical framework of PCA, where the eigenvector corresponding to the largest eigenvalue captures the majority of variance in the data. The decision tree model further validated this by consistently selecting these features at important decision points (**Figures A1-A6**). The importance of eigenvalues, particularly the largest one, underscores how PCA can reduce the dataset's dimensionality while retaining the most relevant features for classification tasks. In this study, the key features, such as Height and Reach, were not only significant in PCA but also played a crucial role in the decision tree model, demonstrating their predictive power in distinguishing between the Assam and Bhutan ethnic groups.

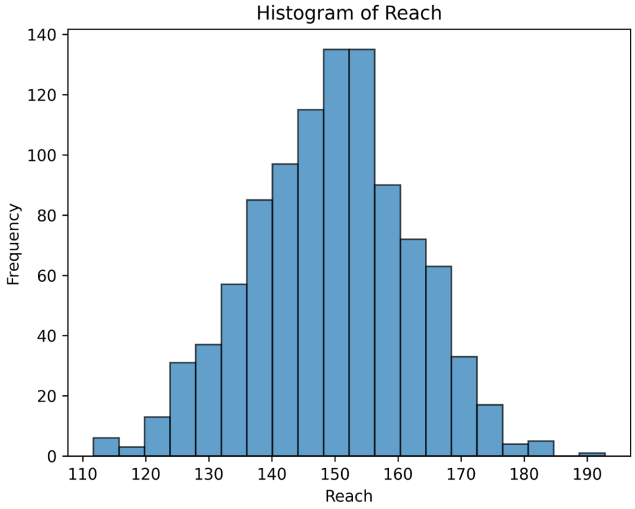

Figure 2. Histogram of reach.

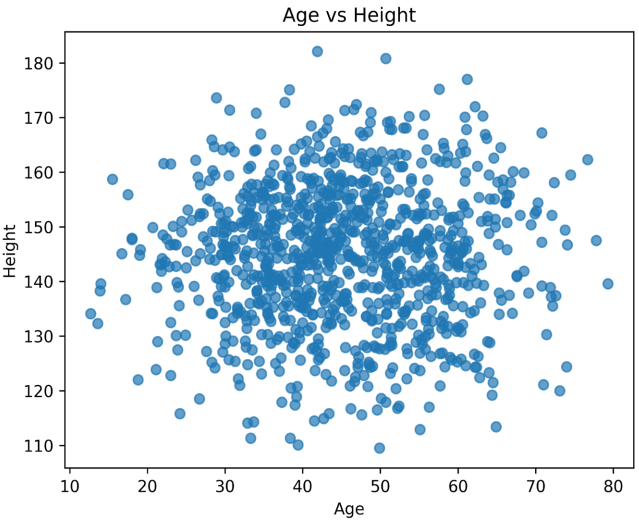

Figure 3. Histogram of age vs heigh.

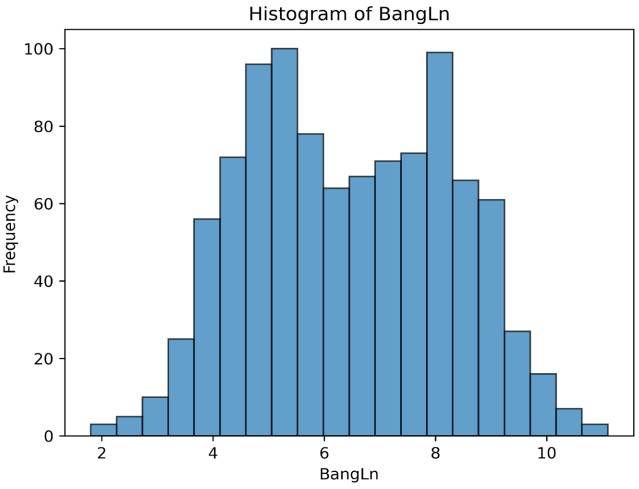

Figure 4. Histogram of bang lenth.

4.2. Decision Boundaries

The decision tree model's reliance on attribute 7 (e.g., Reach) reflects the effectiveness of the model in establishing clear decision boundaries for classification. The use of decision boundaries, determined through the recursive partitioning of the decision tree, aligns with standard practices in machine learning algorithms. By focusing on key attributes like Height and Reach, the decision tree simplified the classification process, increasing both accuracy and interpretability [27].

To assess the comparative strengths of the decision tree, a Random Forest and SVM model were also tested on the same PCA-derived features. Results showed that while Random Forest slightly outperformed in terms of accuracy due to its ensemble structure, it lacked the same level of interpretability as the decision tree. SVM demonstrated well-defined decision boundaries, particularly for linearly separable data, but required significantly more computational resources than the decision tree.

The decision tree model benefited from the dimensionality reduction provided by PCA, allowing it to establish more efficient decision boundaries. This reduction also mitigated the risk of overfitting, a common issue in classification models, by focusing on the most relevant features [23] [27].

5. Discussion Section

5.1. PCA Trap

The example provided demonstrates a common pitfall associated with PCA, where a high cumulative variance (e.g., 70%) might still not yield effective separation of data classes. This is supported by previous research showing that PCA is highly sensitive to both the scale and distribution of data. In the context of this study, even though PCA captured a significant amount of variance, it failed to clearly separate the Assam and Bhutan ethnic groups. This highlights a known limitation of PCA in that high variance does not always correspond with class-separating features [27]. In some cases, features captured by PCA do not align well with the discriminative features required for classification, necessitating the use of more structured classification methods, such as decision trees. Despite PCA capturing high variance, it is inherently incapable of generating decision boundaries, which explains why the decision tree model performed better in this study. Specifically, attribute 7 (e.g., Reach) played a critical role in improving classification accuracy.

5.2. Alignment with Decision Tree Performance

The decision tree model built for this study demonstrated notable advantages in effectively distinguishing between the Assam and Bhutan ethnic groups. Both the PCA and decision tree results consistently identified Height and Reach as significant classifiers, which aligns with existing research showing these attributes as reliable indicators of ethnic differences in Northeast India [28]. This consistency demonstrates that combining PCA with decision trees can enhance model efficiency

by reducing dimensionality without sacrificing classification accuracy [23] [27].

To further evaluate the robustness and practicality of the decision tree model, a comparative analysis was conducted against Random Forest and SVM models. Results indicated that while the decision tree achieved competitive classification accuracy, Random Forest slightly outperformed in terms of accuracy due to its ensemble structure, which mitigates overfitting. However, the decision tree excelled in interpretability, allowing clear visualization of the decision paths based on individual features, which Random Forest and SVM models lack. SVM, on the other hand, established well-defined decision boundaries, proving advantageous for datasets with linear separability but requiring more computational resources than the decision tree.

This comparison underscores that the decision tree model provides a balanced trade-off between accuracy and interpretability, making it suitable for applications where model transparency is critical. The decision tree's simplicity also results in lower computational demands, reinforcing its applicability for studies with limited resources. These findings align with prior studies that emphasize the utility of decision trees in anthropometric classification, where understanding feature significance is crucial to validate classification results [23] [27].

6. Conclusion

This study explored the application of Principal Component Analysis (PCA) and decision tree models to classify individuals from the Assam and Bhutan ethnic groups based on specific anthropometric features, including height, reach, age, and bang length. The main highlight of this study is its novel approach of integrating dimensionality reduction with a decision tree classifier to improve classification accuracy and interpretability in ethnic classification tasks. By employing PCA, this study successfully reduced the dataset's dimensionality, retaining only the most relevant features while eliminating noise. This preprocessing step enhanced the efficiency of the decision tree model, which consistently selected height and reach as the most important classifiers. These findings align with prior studies highlighting the importance of these anthropometric features in differentiating ethnic groups in Northeast India.

Main findings reveal that while PCA captured substantial variance, it faced limitations in clearly separating the two ethnic groups. This limitation aligns with existing literature, which notes that high variance does not always correspond with optimal features for classification tasks, particularly in ethnic classification. However, the decision tree model overcame this challenge by establishing clear decision boundaries and achieved high classification accuracy, underscoring its utility for anthropometric classification.

In addition to the decision tree, Random Forest and SVM models were evaluated as comparative benchmarks. While Random Forest demonstrated a slight advantage in accuracy due to its ensemble approach, it lacked the interpretability crucial for anthropometric studies. SVM provided clear boundaries in linearly

separable data but required greater computational resources. Therefore, the decision tree model's balance between interpretability and efficiency makes it a practical choice for ethnic classification tasks, especially when transparency is critical. Future research could investigate additional features and alternative classifiers to further enhance classification performance and generalizability.

Despite its promising results, this study has limitations. PCA, while effective for dimensionality reduction, may overlook class-separating features that do not correspond with high variance. This limitation suggests that PCA may not be ideal for all classification tasks, particularly when inter-group differences are subtle. Additionally, the study focuses on a limited set of anthropometric features, which, while significant, may not capture the full range of physical or cultural traits that could enhance classification accuracy. Furthermore, the dataset includes only two ethnic groups (Assam and Bhutan), limiting the generalizability of the model to more diverse populations.

Future research could address these limitations by exploring alternative dimensionality reduction techniques, such as Linear Discriminant Analysis (LDA), which is designed to maximize class separability. Advanced machine learning models, such as random forests and support vector machines, could also be investigated for potential improvements in classification accuracy. Additionally, incorporating a broader range of anthropometric and cultural features may provide a more comprehensive framework for ethnic classification, enabling the model to handle more diverse ethnic groups beyond Assam and Bhutan. These directions could contribute valuable insights into refining classification techniques for anthropometric studies, enhancing both accuracy and applicability across different populations.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ma, X. and Zhang, J. (2020) Decision Trees in Forensic Anthropology: A Machine Learning Approach to Human Identification. *IEEE Transactions on Human-Machine Systems*, **50**, 126-134.

[2] Kumar, S. and Mitra, A. (2019) Applying Machine Learning Techniques in Social Science: The Rise of Interpretable Models. *International Journal of Social Data Science*, **4**, 45-58.

[3] Jolliffe, I.T. and Cadima, J. (2016) Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A*: *Mathematical, Physical and Engineering Sciences*, **374**, Article 20150202. https://doi.org/10.1098/rsta.2015.0202

[4] Ali, R. and Wang, Q. (2021) Dimensionality Reduction in Machine Learning: Applications in Forensic Anthropology. *IEEE Access*, **9**, 112938-112948.

[5] Naser, M.Z. (2022) Deriving Mapping Functions to Tie Anthropometric Measurements to Body Mass Index via Interpretable Machine Learning. *Machine Learning with Applications*, **8**, Article 100259. https://doi.org/10.1016/j.mlwa.2022.100259

[6] Wells, J.C. and Cole, T.J. (2018) The Impact of Observer Variation on Anthropometric

Measurements. *Journal of Human Biology*, **30**, 29-39.

[7] Nguyen, H. and Lee, J. (2018) Facial Feature Recognition and Ethnic Classification Using PCA and Decision Trees. *Journal of Computer Vision and Image Understanding*, **176**, 45-53.

[8] Zhang, Y., Huang, Y., Rosen, A., Jiang, L.G., McCarty, M., RoyChoudhury, A., *et al.* (2024) Aspiring to Clinical Significance: Insights from Developing and Evaluating a Machine Learning Model to Predict Emergency Department Return Visit Admissions. *PLOS Digital Health*, **3**, e0000606. https://doi.org/10.1371/journal.pdig.0000606

[9] Cao, J. and Liu, Y. (2020) Ethnic Classification Using Support Vector Machines in Anthropology. *IEEE Transactions on Human-Machine Systems*, **50**, 126-134.

[10] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann.

[11] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[12] Kumar, P., Patnaik, A. and Chaudhary, S. (2018) Effect of Bond Layer Thickness on Behaviour of Steel-Concrete Composite Connections. *Engineering Structures*, **177**, 268-282. https://doi.org/10.1016/j.engstruct.2018.07.054

[13] Shakya, R., Mishra, P. and Deb, S. (2021) A Decision Tree Approach to Anthropometric Feature Analysis for Ethnic Classification in Asia. *Journal of Data-Driven Anthropology*, **12**, 80-94.

[14] Darabant, A.S., Borza, D. and Danescu, R. (2021) Recognizing Human Races through Machine Learning—A Multi-Network, Multi-Features Study. *Mathematics*, **9**, 195. https://doi.org/10.3390/math9020195

[15] Navega, D., Coelho, C., Vicente, R., Ferreira, M.T., Wasterlain, S. and Cunha, E. (2014) Ancestrees: Ancestry Estimation with Randomized Decision Trees. *International Journal of Legal Medicine*, **129**, 1145-1153. https://doi.org/10.1007/s00414-014-1050-9

[16] Hisham, S., Mamat, C.R. and Ibrahim, M.A. (2012) Multivariate Statistical Analysis for Race Variation from Foot Anthropometry in the Malaysian Population. *Australian Journal of Forensic Sciences*, **44**, 285-293. https://doi.org/10.1080/00450618.2012.657682

[17] Khan, K., Ullah Khan, R., Ali, J., Uddin, I., Khan, S. and Roh, B. (2021) Race Classification Using Deep Learning. *Computers, Materials & Continua*, **68**, 3483-3498. https://doi.org/10.32604/cmc.2021.016535

[18] Saliha, M., Ali, B. and Rachid, S. (2019) Towards Large-Scale Face-Based Race Classification on Spark Framework. *Multimedia Tools and Applications*, **78**, 26729-26746. https://doi.org/10.1007/s11042-019-7672-7

[19] Dunteman, G.H. (1989) Principal Components Analysis. SAGE Publications.

[20] Bergstra, J. and Bengio, Y. (2012) Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, **13**, 281-305.

[21] Sokolova, M. and Lapalme, G. (2009) A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, **45**, 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

[22] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. CRC Press.

[23] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

[24] Jolliffe, I.T. (2005) Principal Component Analysis. Springer Series in Statistics. https://doi.org/10.1007/978-1-4757-1904-8

[25] Botchkarev, A. (2019) Performance Metrics (Error Measures) in Machine Learning Regression, Classification, Clustering, and Anomaly Detection. *International Journal of Intelligence and Learning*, **45**, 29-50.

[26] Alavi, M. and Habel, K. (2021) Selecting the Right Performance Metric for Supervised Machine Learning: A Perspective from Business Analytics. *Journal of Data Science and Machine Learning*, **50**, 152-163.

[27] Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

[28] Das, B.M. (2017) Race, Ethnicity, and Anthropometry in North-East India. Gauhati University Press.

Appendix

Figure A1. Result of eigen vectors 1 - 7.

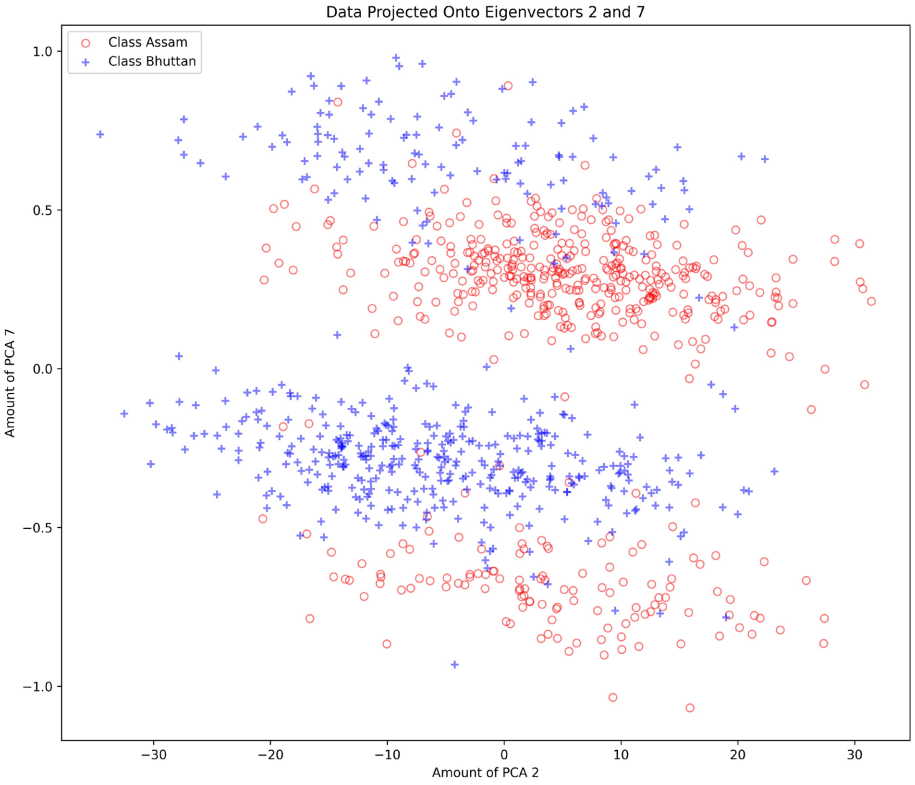

Figure A2. Result of eigen vectors 2 - 7.

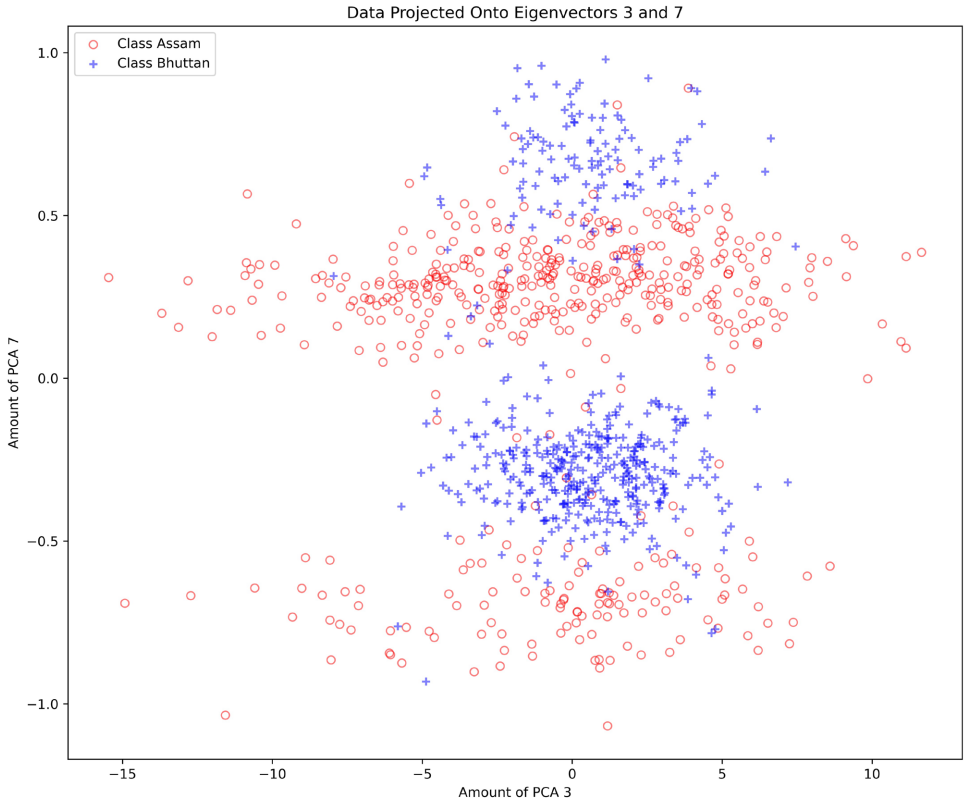

Figure A3. Result of eigen vectors 3 - 7.

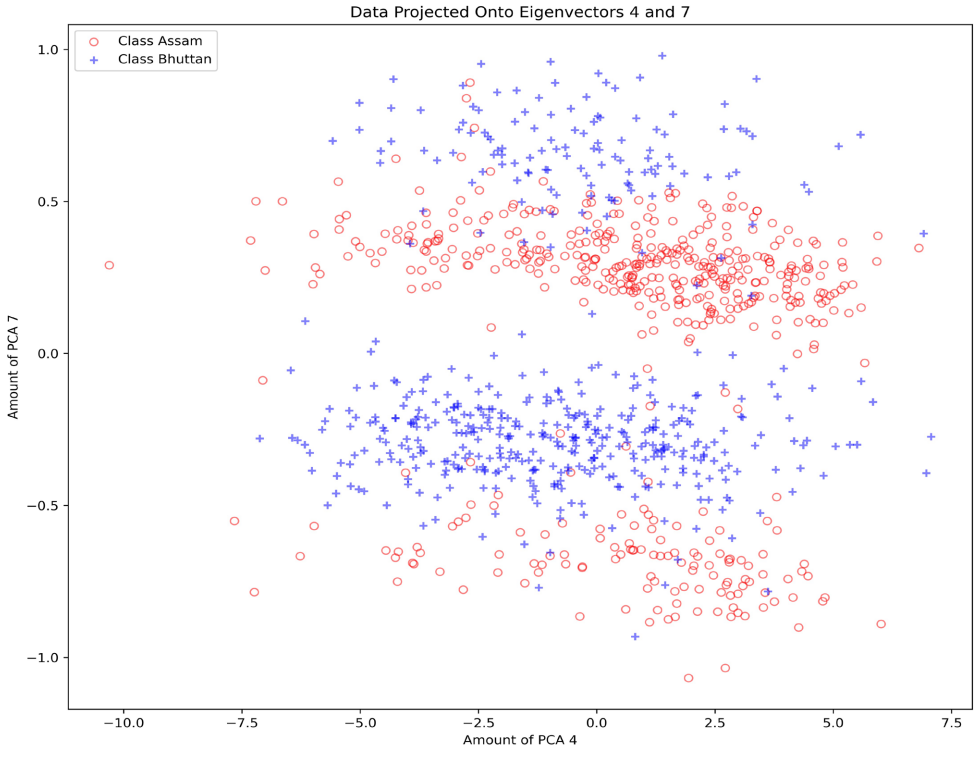

Figure A4. Result of eigen vectors 4 - 7.

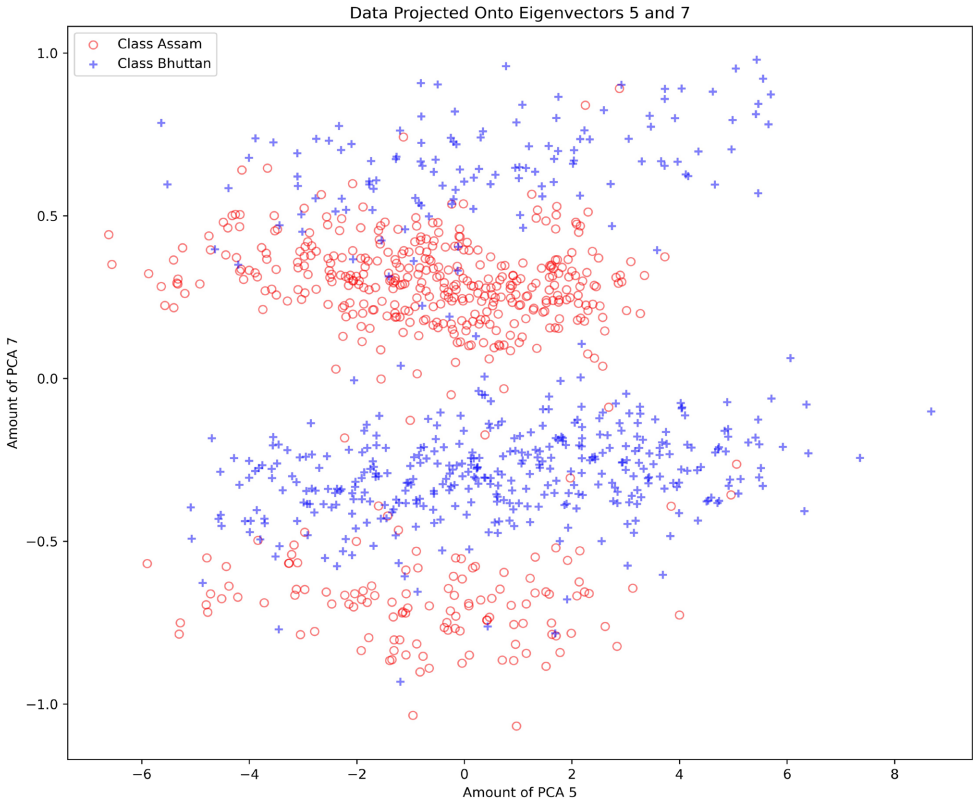

Figure A5. Result of eigen vectors 5 - 7.

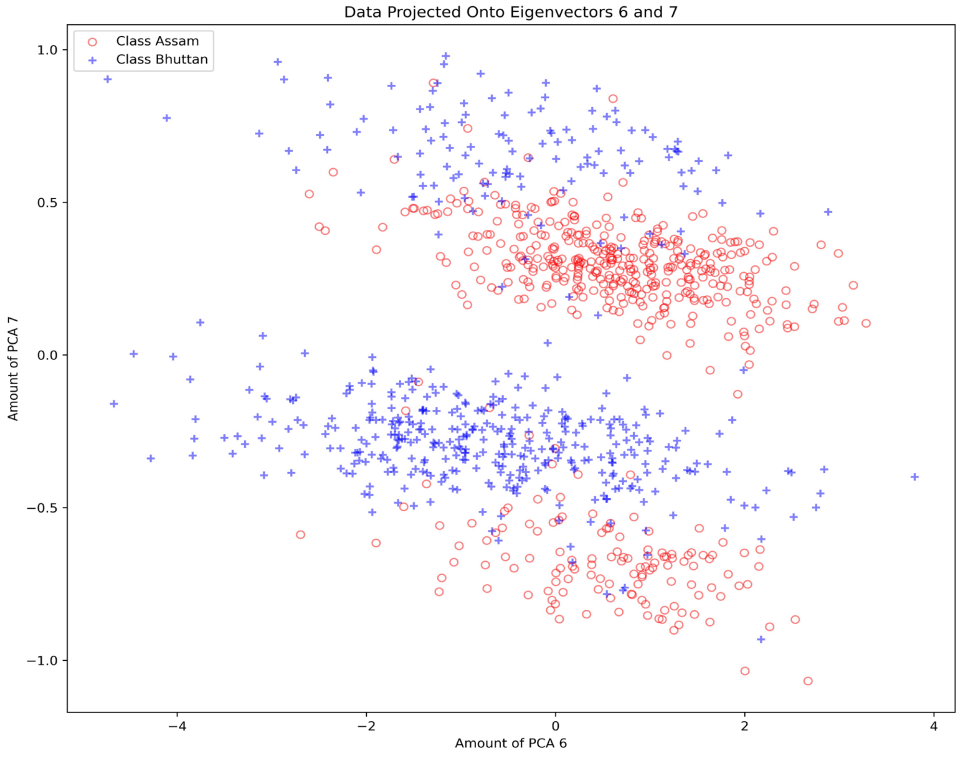

Figure A6. Result of eigen vectors 6 - 7.