Received: 16 Mar 2024 | Revised: 02 Aug 2024 | Accepted: 28 Aug 2024 | Published: 24 Sept 2024
Volume 2(2) Pages 29-41 DOI: https://doi.org/10.56279/jicts.v2i2.89

JICTS
Journal of ICT Systems

Advancing Face Recognition Technologies: The Role of Decision Trees in Classifying Complex Image Pairs

Francis Kakula[1], Jimmy Mbelwa, Hellen Maziku

Department of Computer Science and Engineering, College of Information and Communication Technologies, University of Dar esSalaam, Dar es Salaam, Tanzania

[1]Corresponding author
Email: frakasama@gmail.com

Funding information

This work is not funded by any organization

Abstract

The advancement of face recognition technologies has been pivotal in various applications, from security systems to personalized user experiences. There are significant efforts already devoted to solving challenges of multimodality and pose variation in face recognition. Some studies focus on multimodality but pose-invariant, and other studies focus on pose variation but single modality. Despite significant progress, various face recognition algorithms do not consider both multimodality and pose variation constraints in their proposed methods. Recognizing face images presented both in a different modality and in a different pose presents serious challenges to current algorithms. This paper proposes an algorithm that combines the strengths of deep learning with decision trees to improve face recognition performance across modalities and poses in constrained and unconstrained environments. This hybrid approach leverages the representational power of deep learning and the interpretability and simplicity of decision trees. The findings indicate significant improvements over existing methodologies, particularly in challenging conditions like when multimodality and pose variation constraints are compounded together in the input face images in both constrained and unconstrained environments. The proposed algorithm not only addresses the limitations of current face recognition systems but also offers scalable, efficient solutions suitable for real-world applications.

Keywords

Deep Learning
Decision Trees
Face Recognition
Multimodality
Pose Variation

1. Introduction

Face recognition is a part of pattern recognition applied to identify or confirm individuals' identities using their faces. Face recognition systems can be used to identify individuals in photos, videos, and in real-time [1]. Face recognition involves several key processes. First, image acquisition captures facial images using devices like cameras. Next, face detection identifies and locates faces within these images. Following this, alignment standardizes the orientation and scale of the

detected faces. Feature extraction then identifies distinctive facial features. These features are subsequently compared against a database in the face matching step using algorithms to measure similarity. Finally, decision making determines identity based on these comparisons. Face recognition is one of the most successful applications in computer vision for analyzing images to obtain effective information and understanding of images [2]. Face recognition has numerous applications in image analysis, security, surveillance, law enforcement, and other areas. Face recognition across modalities refers to matching face images across different imaging domains [3]. Face recognition for varying poses refers to recognizing face images presented in different poses [4]. In the context of this study, face recognition across modalities and poses refers to recognizing face images presented both in a different modality (different sensors: visible images, near-infrared images, thermal images, and computerized facial sketches) and in a different pose. A constrained environment is one in which one can control the parameters (lighting, background, camera angle, pose, and other factors) of the input images, while an unconstrained environment is one in which one has little or no control over such parameters [5]. Recognizing human faces across modalities and poses, both in constrained and unconstrained environments, has a great potential in surveillance and security applications dealing with uncooperative subjects.

There are significant efforts already devoted to solving challenges of multimodality and pose variation in face recognition. Some studies focus on multimodality but pose-invariant [3, 6-8], and other studies focus on pose variation but single modality [9-15]. Despite recent significant advances in the field of face recognition, various face recognition algorithms do not consider both multimodality and pose variation constraints. Recognizing face images presented both in a different modality and in a different pose presents serious challenges to current algorithms. This challenge is compounded by variations in illumination, expression, gesture, and occlusion. Addressing challenges for multimodality and pose variations together as one problem and not as separate problems is extremely important for surveillance, applications dealing with uncooperative subjects, and security applications.

This paper addresses the challenges of multimodal and pose variation combination in face recognition. Researchers taking this route face several difficulties, including the resource intensive nature of gathering diverse multimodal data and the challenge of integrating information from different modalities effectively and ensuring that the model performs well across various poses. Additionally, Training complexity arises as multimodal models often require specialized architectures. Choosing appropriate metrics for evaluating the performance of multimodal and pose variation combination is also a critical and complex task.

In face recognition, various types of decision trees can be used to enhance performance and interpretability. Traditional decision trees are simple and intuitive, splitting data based on feature values but are often prone to overfitting. Random Forests (RF) mitigate this by constructing multiple trees and aggregating their predictions, making them robust and effective for high-dimensional data, which is common in face recognition tasks. Gradient Boosting Decision Trees (GBDT) build trees sequentially, with each tree correcting errors from the previous ones, leading to high predictive accuracy. Variants of GBDT like Dropouts meet Multiple Additive Regression Trees (DART) and Gradient-based One-Side Sampling (GOSS) incorporate techniques to improve performance and efficiency. DART integrates dropout mechanisms from neural networks to prevent overfitting,

whereas GOSS selects data subsets based on gradient information to enhance speed.

This paper proposes an algorithm that combines the strengths of deep learning with decision trees to improve face recognition performance across modalities and poses in constrained and unconstrained environments. This hybrid approach leverages the representational power of deep learning and the interpretability and simplicity of decision trees. The objective of this approach is to handle complex, high-dimensional data efficiently while providing clear, rule-based decisions. The benefits include enhanced interpretability, flexibility, and the potential for reduced overfitting [16, 17]. However, challenges such as increased computational complexity and careful optimization of all components must be managed. The proposed algorithm represents a promising direction in machine learning, balancing complexity and transparency in decision making processes.

Unlike traditional approaches that focus solely on deep learning or decision trees, this work is unique in its integration of deep learning with decision trees to address the challenges of multimodal and pose variation combination in face recognition, an area that has been largely unexplored due to the inherent challenges it presents. The deep learning component ensures robust feature extraction from diverse and complex image data, whereas the decision trees component provides a structured and an interpretable classification process. Analytically, this synergy is justified as it combines the high-dimensional feature representation power of neural networks with the transparent and rule-based classification of decision trees, thus reducing overfitting and enhancing generalizability. Our work bridges the gap between deep learning and interpretable models, making it a valuable contribution to the field of face recognition. Moreover, our analysis using multiple datasets from both constrained and unconstrained environments demonstrates that the proposed algorithm significantly improves recognition performance under challenging conditions such as multimodal and pose variation combination, showcasing its robustness and practicality.

The performance of the proposed algorithm is evaluated using Precision, Recall, F1 score, Accuracy, and Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve performance evaluation metrics. Precision is the number of true positives divided by the sum of true positives and false positives. If a model has a precision value of 0.5 it means when it predicts a positive case, it is correct 50% of the time. Recall is the number of true positives divided by the sum of true positives and false negatives. If a model has a recall value of 0.25 it means it correctly identifies only 25% of all positive cases. The F1 score is the harmonic mean of precision and recall taking both metrics into account. If you want to create a balanced classification model with the optimal balance of recall and precision, then try to maximize the F1 score. Accuracy is the ratio of correct predictions to total predictions made. When dealing with skewed datasets, accuracy is not the preferred performance evaluation. AUC - ROC also written as Area Under the Receiver Operating Characteristics (AUROC) is one of the most important performance evaluation metrics for classification problems. AUC is the degree or measure of separability and ROC is a probability curve. The best model has an AUC near 1, which means it has a good measure of separability. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between the positive class and the negative class.

2. Related Works

Cheema *et al*. [3] proposed a unified end-to-end Cross-Modality Discriminator Network (CMDN) for Heterogeneous Face Recognition (HFR), which

matches face images across different imaging domains (visible-thermal, visible-infrared, and visible-sketch). The CMDN uses a Deep Relational Discriminator module to learn deep feature relations and extract modality-independent embedding vectors. The CMDN is optimized using a novel Unit-Class Loss that leads to higher stability and accuracy compared with other metric-learning loss functions. The proposed method achieves a Heterogeneous Face Recognition Rank-1 accuracy of 93.6% on UND-X1, 91.6% on SFD, 97.0% on TUFTS, 99.3% on USTC-NVIE, and 99.2% on CASIA NIR-VIS 2.0 datasets.

He *et al.* [7] proposed the Wasserstein convolutional neural network (WCNN) to handle near-infrared and visual face image recognition. The WCNN trains low-level layers using visual images and splits high-level layers into three parts: near-infrared (NIR), visual (VIS), and NIR-VIS shared. The WCNN showed improved performance compared to current state-of-the-art methods when evaluated on three NIR-VIS face recognition datasets.

Ghosh *et al.* [8] proposed the Subclass Heterogeneity Aware Loss (SHEAL) to tackle cross-modality face recognition. The SHEAL function trains deep neural networks for cross-spectral and cross-resolution recognition. When tested on four databases, the SHEAL function achieved outstanding results in both homogeneous face recognition and in the challenging scenario of cross-spectral cross-resolution recognition.

Ullah *et al.* [9] proposed a deep learning-based real-time framework for recognizing human faces in CCTV images. The framework achieved recognition accuracy of over 90% with minimal computing time.

Liang [10] proposed an unrestricted face recognition algorithm for recognizing faces in unrestricted environments. The algorithm demonstrated better recognition speed and recognition rate of 98.19% on LFW (Labeled Faces of the Wild), 92.20% on MegaFace, and 99.52% on CASIA-WebFace datasets.

Ye *et al.* [6] proposed a modality-aware collaborative ensemble learning method for visible thermal person re-identification, where pedestrian images captured by different cameras (visible during the day and thermal at night) need to be matched. The proposed method uses a middle-level sharable two-stream network to handle feature-level discrepancies. In classifier level, both modality-specific and modality-sharable identity classifiers are introduced. When tested on two cross-modality datasets, the proposed method achieves Rank-1/mAP accuracy of 51.64%/50.11% on SYSU-MM01 dataset, and 72.37%/69.09% on RegDB dataset.

Mostofa *et al.* [18] proposed a method for face recognition across poses by using pose as auxiliary information. Instead of frontalization or disentangling pose information, the authors introduce a pose attention block that guides feature extraction from profile faces. When tested on both constrained and unconstrained benchmarks including Multi-PIE (Multi Pose, Illumination, Expressions) dataset, CFP (Celebrities in Frontal-Profile) dataset, and IJB-C (IARPA Janus Benchmark-C) dataset, the results indicate that the proposed method significantly improves face recognition performance, especially for profile-to-frontal face matching, compared to state-of-the-art methods.

3. Proposed Algorithm

This study used an experimental research design. The method involves a two-stage process: feature extraction using deep learning and classification using decision trees. The deep learning component extracts high-dimensional features from facial images, while the decision trees

classify these features into distinct identity classes. The algorithm consists of two parts: the Multimodality and Pose Variation Discriminator Network (MPVDN) for deep learning-based feature representation, and the Decision Trees Discriminator (DTD) for making interpretable decisions based on these representations. The MPVDN is expressed as a function F that maps input image pair X to a similarity distance vector V.

$$V = F(X ; \theta)$$

where X is the input image pair, V is the similarity distance vector, F represents the MPVDN, and θ denotes the parameters of the MPVDN.

The MPVDN functions by generating a vector V of similarity distances for a given pair of images. The computation of similarity distances between images in a pair typically employs mathematical and statistical models to evaluate and quantify the similarity metrics. These models may include, among others, Euclidean distance, Cosine similarity, Structural Similarity Index (SSIM), or more complex neural network-based configurations that extract and compare deep features of the images. The choice of the method for computing similarity distances depends on the specific requirements of the application, including the level of sensitivity needed to detect differences, the computational resources available, and the nature of the images themselves (e.g., size, quality, and domain-specific characteristics). The outcome of this process is quantitative measures of similarity.

The study employed neural network-based configurations to computing the vector V of similarity distances for each image pair. The computational environment was equipped with T4 GPU and 13.6 gigabytes of Random Access Memory (RAM). There is a notable variance in processing times across different neural network-based configurations. The processing times vary not just with the neural network-based configurations but also with the dataset being used. This indicates that the dataset characteristics (such as image resolution, diversity of face orientations, and lighting conditions) can significantly influence processing speed. The dimensionality of vector V is directly proportional to the number of neural network-based configurations applied. The study systematically varied the dimensionality of this vector. The underlying hypothesis was that increasing the diversity and number of these neural network-based configurations would correlate with enhanced face recognition performance. The similarity distance vector V was expanded to explore its impact on the algorithm's performance. The most extensive configuration tested employed 105 distinct neural network-based configurations. This specific configuration yielded a similarity distance vector with a dimensionality of 105.

The neural network-based configurations used in generating the vector V of similarity distances for each pair of images were chosen from academic and commercial studies whose structures and pre-trained weights have been shared publicly or are available from the open-source community. The decision to use neural network-based configurations, specifically those whose structures and pre-trained weights are publicly shared or available from the open-source community, reflects a commitment to leveraging cutting-edge technology as well as facilitating reproducibility and validation of the study's findings. This open-source approach not only enhances the study's credibility but also contributes to the broader academic and technical communities by building upon publicly available resources.

In the context of computing similarity distances between images, the preliminary step involves the generation of image pairs. Image pairing can be systematically executed through various methodologies, including, but not limited to,

manual selection, algorithmic matching based on metadata, or employing machine learning models that understand and categorize images based on their visual content. The study generated image pairs through algorithmic matching based on metadata. The computational environment was equipped with 54.8 gigabytes of RAM, 225.89 gigabytes of disk storage, and an Intel(R) Xeon(R) Central Processing Unit (CPU) operating at 2.20 GHz.

The DTD is expressed as a function G that maps the similarity distance vector V to a decision Y.

$$Y = G\,(V\,;\,R)$$

where V is the similarity distance vector, Y is the decision output, G represents the DTD, and R denotes the decision rules derived during the training of the decision tree.

The input to DTD is the similarity distance vector V. This vector is obtained from MPVDN, where $V = [v_1, v_2, ..., v_i, ..., v_{105}]$ represents the computed similarity distances for a given pair of images. Each component v_i of V is a scalar value representing the similarity score resulting from one neural network-based configuration.

During the training phase, the DTD was exposed to a diverse dataset S that includes face images across different modalities (visible, near-infrared, thermal, and computerized facial sketches) and poses. This exposure enables the DTD to derive robust decision rules. This process is represented as optimizing a loss function L, where L = Loss (Y, Y_{true}), with Y_{true} being the true classification labels for the training samples. The optimization is subject to constraints C that reflect the specific requirements of the deployment environment, hence $L(S;\theta) \rightarrow \min_{\theta \in C}$, where θ denotes the parameters of the DTD and $\rightarrow \min_{\theta \in C}$ means that we want to find the parameters (θ) within our constraints (C) that will give us the lowest possible score on the loss function L. The adaptation of the DTD to specific environmental constraints is essential for its effective deployment.

The study leveraged the computational resources available to process and analyze a substantial dataset, resulting in the generation of 121,300 105-dimensional vectors, which serve as representations of the original image pairs in high-dimensional spaces.. This dataset was meticulously compiled from a random homogeneously distributed sample of image pairs, totaling 121,300, with an equal split between 60,650 positives (image pairs belonging to the same identity) and 60,650 negatives (image pairs belonging to different identities). The sources of these image pairs were carefully selected to include both constrained and unconstrained environments, utilizing images from the Extended Yale Face Database B and the Tufts Face Database for constrained settings, alongside the VGGFace2 dataset and the Labeled Faces in the Wild Database for unconstrained environments (Table 1).

This approach ensures a balanced representation of conditions in the sample. A homogeneous sample, by definition in this context, refers to a dataset that maintains an equal number of positive and negative instances, which is crucial for eliminating bias in machine learning, particularly in the domain of face recognition where balanced datasets can significantly impact the accuracy and fairness of the outcomes.

Table 1. The distribution of the generated 121,300 105-dimensional vectors.

Datasets		**Positives**	**Negatives**	**Total**
Constrained environments	Extended Yale B	12,750	12,750	**25,500**
	Tufts	4,850	4,850	**9,700**
Unconstrained environments	VGGFace2	16,350	16,350	**32,700**
	LFW	26,700	26,700	**53,400**
Total		**60,650**	**60,650**	**121,300**

The random homogeneous sample was split into 89% (107,954 image pairs) homogeneous training set and 11% (13,346 image pairs) homogeneous testing set. To prevent overfitting, the training set was further split into a 90% (97,154 image pairs) homogeneous training set and a 10% (10,800 image pairs) homogeneous validation set. By using a validation set, the DTD's performance can be assessed on independent data, allowing for early detection of overfitting and adjusting the DTD's parameters as needed. All data splits were random to ensure that the training, validation, and testing sets were representative of the underlying population.

However, the ratio of the training and validation sets can have an impact on the DTD's performance. If the validation set is too small, it may not provide a representative sample of the data and may not accurately reflect the DTD's performance on new, unseen data. On the other hand, if the validation set is too large, it may leave too little data for training, which could lead to underfitting, where the DTD fails to capture the underlying patterns in the data. Ultimately, the ideal ratio of training and validation sets will depend on the size and nature of the data, as well as the specific requirements of the task at hand. In the case of this study, the specific ratio of 90% training and 10% validation data was chosen to balance the need for a large enough training set to learn from, with the need for a representative validation set to prevent overfitting.

The Optuna framework was used to tune the following hyperparameters in DTD: feature fraction (0.5979122703220067), number of leaves (209), bagging fraction (0.9253614353373035), bagging frequency (12), regularization factors (lambda_l1: 1.2665055161513616 and lambda_l2: 5.458331618514055e-06), minimum child samples (43), and seed (4924485674646214656). Optuna is an open-source framework that can be used to optimize hyperparameters using the Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian optimization method balancing exploration and exploitation.

The proposed algorithm is expressed as a composite function H:

$$H(X; \theta, R) = G(F(X; \theta); R)$$

where X is the original input image pair, H represents the proposed algorithm, F and G represent the MPVDN and DTD, respectively, and θ and R denote the parameters of the MPVDN and the decision rules of the DTD, respectively.

Flowchart of the Proposed Algorithm:

1. Input Image Pair (X)
   - Two images are fed into the algorithm as input.
2. MPVDN Processing ($F(X; \theta)$)
   - The images are passed through the MPVDN, which computes a similarity distance vector V where $V = [v_1, v_2, ..., v_i, ..., v_{105}]$ represents the computed similarity distances for a given pair of images. Each component v_i of V is a scalar value representing the similarity score resulting from one neural network-based configuration.
   - MPVDN employs mathematical and statistical models (Euclidean distance, Cosine similarity, Neural network-based configurations) to compute similarity distances between images in a pair.
   - The output of the MPVDN is a similarity distance vector $V = [v_1, v_2, ..., v_i, ..., v_{105}]$ that represents the relationship between the input images.
3. DTD Processing ($G(V; R)$)
   - The similarity distance vector $V = [v_1, v_2, ..., v_i, ..., v_{105}]$ is input into the DTD, which applies decision rules R to classify the image pair.

- The decision rules *R* are derived from the training of the decision tree and determine how the classification is performed.
- The final output is the classification decision *Y*, indicating whether the images match or not.

4. Experimental Results

The results of the experiments are promising, as the algorithm achieved state-of-the-art performance in terms of Precision, Recall, F1 score, Accuracy, and AUC–ROC curve to variations in modalities and poses. The empirical analysis reveals a positive correlation between the dimensionality of the similarity distance vector and the performance metrics. Specifically, as the dimensionality increases, notable improvements are observed across several performance metrics (Table 2).

While the overall trend is positive, the incremental gains in performance metrics become less pronounced as dimensionality reaches higher levels (notably beyond 70 dimensions). This suggests the presence of a point of diminishing returns, where additional dimensions add less value to the algorithm's predictive capabilities, potentially increasing computational costs without significant performance improvement. The AUC consistently increases with dimensionality, indicating that the algorithm's ability to distinguish between classes improves as more dimensions are added. This pattern underscores the importance of dimensional complexity in enhancing algorithm discriminative power.

Table 2. Correlation between the dimensionality of the similarity distance vector (Dim of Vector *V*) and the performance metrics.

Dim of Vector *V*	Precision	Recall	F1 score	Accuracy	AUC
1	90.25%	81.03%	85.39%	86.14%	91.16%
3	89.69%	85.76%	87.68%	87.95%	94.18%
5	90.63%	87.14%	88.85%	89.07%	95.38%
15	94.41%	91.38%	92.87%	92.99%	98.07%
35	95.48%	93.15%	94.30%	94.37%	98.83%
70	95.41%	93.06%	94.22%	94.29%	98.82%
105	95.66 %	94.77 %	95.21 %	95.24 %	99.19 %

Both precision and recall improve in tandem as dimensionality increases. This balanced improvement is crucial, as it indicates that the algorithm is not only returning more relevant results (precision) but is also increasingly capable of identifying a higher proportion of actual positives (recall) without disproportionately favoring one metric over the other. There is a strong positive correlation between dimensionality and the F1 score, suggesting that the harmonic mean of precision and recall benefits significantly from increased dimensional complexity. This correlation is particularly relevant in scenarios where a balance between precision and recall is essential for algorithm performance. While both accuracy and AUC improve with increased dimensionality, the growth rate of AUC is particularly noteworthy. This implies that dimensionality has a more pronounced effect on the algorithm's ability to rank predictions effectively across different thresholds, a critical aspect of performance in many applications.

A very high positive correlation exists between recall and the F1 score, indicating that the ability to correctly identify true positives is a strong driver of the overall performance balance (F1 score). The correlation between the F1 score and accuracy is also very high, indicating that higher F1 scores typically lead to higher accuracy. The correlation between precision and accuracy is positive, suggesting a moderate relationship where higher precision may lead to higher accuracy.

The DTD was evaluated using various configurations to determine the optimal configuration. The configurations tested included GBDT, DART, GOSS, and RF. The results of these experiments indicated that the GBDT configuration consistently outperformed the other configurations.

GBDT demonstrated superior performance in terms of F1 score, which measures the balance between precision and recall, Accuracy, which reflects the overall correctness of the predictions, and AUC, which assesses the algorithm's ability to distinguish between positive and negative classes. While DART, GOSS, and RF also provided competitive results, they did not match the overall performance of GBDT. These findings underscore the suitability of GBDT as the optimal configuration for the DTD in the context of face recognition (Table 3).

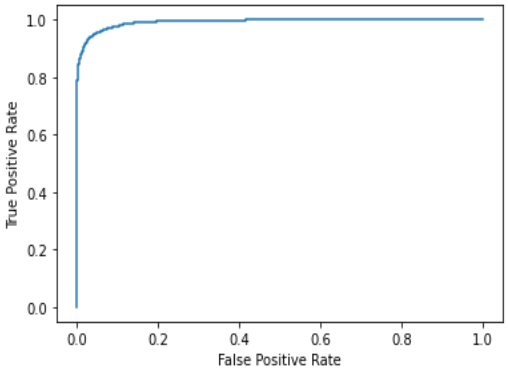
The experimental results highlight the pivotal role of decision trees, especially GBDT, in advancing face recognition technologies. GBDT and its variants have shown superior performance in classifying complex image pairs, making them crucial for face recognition across different modalities and poses in both constrained and unconstrained environments.

Table 4 shows the performance of the algorithm based on the MPVDN's configuration that yielded a similarity distance vector with a dimensionality of 105 for each pair of images. With 6,386 true negatives (TN) and 6,324 true positives (TP), the algorithm effectively distinguishes between non-matching and matching pairs of images. The relatively low number of false positives (FP=287) and false negatives (FN=349) further underscore the algorithm's robustness in making classifications. The precision (95.66%) and recall (94.77%) metrics are both higher, indicating a strong correlation between the algorithm's ability to correctly identify TP and minimize FN.

Table 3. Performance of the algorithm based on different DTD configurations.

	Precision	Recall	F1 score	Accuracy	AUC
GBDT	95.66%	94.77%	95.21%	95.24%	99.19%
DART	95.74%	94.64%	95.18%	95.21%	99.18%
GOSS	95.59%	94.55%	95.07%	95.09%	99.17%
RF	94.75%	92.52%	93.62%	93.70%	98.63%

Table 4. Performance of the algorithm based on similarity distance vector, *V*, with a dimensionality of 105.

Confusion Matrix	$\begin{bmatrix} 6386 & 287 \\ 349 & 6324 \end{bmatrix}$
Precision	95.66 %
Recall	94.77 %
F1 score	95.21 %
Accuracy	95.24 %
AUC	99.19 %
ROC Curve	

This balance is crucial in applications where both FP and FN have significant consequences, such as security or identity verification. The close values suggest the algorithm is well-tuned to maintain a balance between these two aspects, which is not always easy to achieve.

The F1 score (95.21%) serves as a bridge between precision and recall, providing a single metric that encapsulates the algorithm's overall performance in terms of its precision-recall balance. The high F1 score indicates that the algorithm does not heavily favor precision over recall or vice versa, which is a common challenge in classification tasks. This suggests a well-optimized approach to handling the trade-offs between identifying all relevant instances and ensuring the relevancy of identified instances.

The accuracy of 95.24% indicates the algorithm's reliability across all predictions. This high level of accuracy ensures that the algorithm is dependable in varied scenarios, affirming its utility in real-world applications.

The AUC (99.19%) is particularly noteworthy. This is one of the most important performance evaluation metrics for classification problems. There is a 99.19% chance that the algorithm will be able to distinguish between positive class and negative class. It suggests that the algorithm has a promising ability to discriminate positive class from negative class. In practical terms, this means the algorithm can accurately rank pairs of images by their likelihood of being a match with few errors, approaching the ideal scenario. The AUC being significantly high indicates that the algorithm's performance is robust across various decision thresholds, which implies that its predictive capabilities are not confined to a specific operating point. This is particularly important in practical scenarios where decision thresholds might need to be adjusted based on specific requirements.

The ROC curve is close to the top left corner of the plot. This position indicates excellent performance, where the algorithm maximizes the TP rate while minimizing the FP rate.

The algorithm demonstrated high performance on a wider range of datasets that simulate real-world conditions more closely with AUC values consistently above 95%. In constrained environments, the Extended Yale B dataset yielded an AUC of 97.56%, and the Tufts dataset followed closely with an AUC of 95.78%. These results indicate the algorithm's effectiveness in constrained settings. In unconstrained environments, the algorithm's performance was even more impressive. The VGGFace2 dataset resulted in an AUC of 98.81%, while the LFW dataset achieved near-perfect performance with an AUC of 99.98%. These datasets represent real-world scenarios with variations in lighting, pose, and background, highlighting the algorithm's robustness in handling diverse and challenging conditions (Table 5).

Table 5. Performance (AUC) of the algorithm on a wider range of datasets that simulate real-world conditions more closely.

Datasets		AUC
Constrained environments	Extended Yale B	97.56 %
	Tufts	95.78 %
Unconstrained environments	VGGFace2	98.81 %
	LFW	99.98 %

5. Practical Applications and Integration

The proposed algorithm has significant potential applications in various domains such as law enforcement, surveillance, security, and image analysis. In law enforcement, the proposed algorithm can be used to match computerized facial sketches with visible, near-infrared, or thermal images. This could be useful in identifying suspects based on sketches provided by witnesses or victims, which is crucial for identifying suspects when only a sketch is available. In surveillance and security applications, the proposed algorithm can be used to recognize faces in surveillance systems dealing with uncooperative subjects or in security systems to enhance real-time monitoring by identifying individuals in crowded and dynamic environments, both during the day (visible) and at night (near-infrared and thermal). In the field of image analysis, the algorithm’s ability to recognize faces across different modalities and poses in both constrained and unconstrained environments can be used to analyze images and extract valuable information in applications such as social media analysis, where understanding the content of images is crucial.

However, integrating the proposed algorithm into existing systems may present an increased computational complexity challenge that could be mitigated by optimizing the algorithm for efficiency, for example, by leveraging hardware acceleration or developing modular and interoperable components that can easily integrate with different systems.

6. Conclusion

This study has shown that by combining the representational power of deep learning and the interpretability and simplicity of decision trees, we can significantly enhance the performance and robustness of face recognition across modalities and poses in constrained and unconstrained environments. The algorithm's design handles complex, high-dimensional data efficiently while providing clear, rule-based decisions. The findings indicate significant improvements over existing methodologies, particularly in challenging conditions like when facial images are presented in a different modality and pose. The proposed algorithm not only addresses the limitations of current face recognition systems but also offers scalable, efficient solutions suitable for real-world applications.

Despite the inherent complexities and the increased computational demands, the outcomes of this study not only contribute valuable insights into the field of computer vision but also pave the way for more secure and reliable biometric recognition and surveillance systems in the future. The proposed algorithm represents a promising direction in machine learning, balancing complexity and transparency in decision-making processes.

Future work will focus on refining the proposed algorithm to further improve its effectiveness and efficiency, exploring the integration of additional modalities, such as 3D images, to further enhance face recognition performance under varied environmental conditions, and optimizing the proposed algorithm for deployment on devices with limited computational resources.

REFERENCES

[1] M. A. Yaman, A. Subasi, and F. Rattay, *Comparison of random subspace and voting ensemble machine learning methods for face recognition*, Symmetry (Basel), vol. 10, no. 11, Nov. 2018.

[2] Y. Feng, X. An, and S. Li, *Research on Face Recognition Based on Ensemble Learning*, in Chinese Control Conference, CCC, 2018.

[3] U. Cheema, M. Ahmad, D. Han, and S. Moon, *Heterogeneous visible-thermal and visible-infrared face recognition using cross-modality discriminator network and unit-class loss*, *Comput Intell Neurosci*, vol. 2022, pp. 1–15, 2022.

[4] M. O. Oloyede, G. P. Hancke, and H. C. Myburgh, *A review on face recognition systems: recent approaches and challenges*, Multimed Tools Appl, vol. 79, no. 37–38, pp. 27891–27922, 2020.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, *Labeled faces in the wild: a database for studying face recognition in unconstrained environments*. [Online]. Available: http://vis-www.cs.umass.edu/lfw/.

[6] M. Ye, X. Lan, Q. Leng, and J. Shen, *Cross-modality person re-identification via modality-aware collaborative ensemble learning*, IEEE Transactions on Image Processing, vol. 29, pp. 9387–9399, 2020.

[7] R. He, X. Wu, Z. Sun, and T. Tan, *Wasserstein CNN: Learning invariant features for NIR-VIS face recognition*, IEEE Trans Pattern Anal Mach Intell, vol. 41, no. 7, pp. 1761–1773, 2019.

[8] S. Ghosh, R. Singh, and M. Vatsa, *Subclass heterogeneity aware loss for cross-spectral cross-resolution face recognition*, IEEE Trans Biom Behav Identity Sci, vol. 2, no. 3, pp. 245–256, 2020.

[9] R. Ullah et al., *A Real-Time Framework for Human Face Detection and Recognition in CCTV Images*, Math Probl Eng, vol. 2022, pp. 1–12, 2022.

[10] Z. Liang, *Unrestricted Face recognition algorithm based on transfer learning on self-pickup cabinet*, Math Probl Eng, vol. 2021, 2021.

[11] M. M. Y. Zhang, K. Shang, and H. Wu, *Learning deep discriminative face features by customized weighted constraint*, Neurocomputing, vol. 332, pp. 71–79, 2019.

[12] Y. Qian, W. Deng, and J. Hu, *Unsupervised Face normalization with extreme pose and expression in the wild*. [Online]. Available: https://github.com/mx54039q/fnm.

[13] C. Li, M. Liang, W. Song, and K. Xiao, *A Multi-scale parallel convolutional neural network based intelligent human identification using face information*, Journal of Information Processing Systems, vol. 14, no. 6, pp. 1494–1507, 2018.

[14] Q. Zhou et al., *Face recognition via fast dense correspondence*, Multimed Tools Appl, vol. 77, no. 9, pp. 10501–10519, 2018.

[15] G. S. Hsu, A. M. Ambikapathi, S. L. Chung, and H. C. Shie, *Robust cross-pose face recognition using landmark oriented depth warping*, J Vis Commun Image Represent, vol. 53, pp. 273–280, 2018.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Chapman and Hall/CRC, 1984. https://doi.org/10.1201/9781315139470.

[17] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, *Ensemble selection from libraries of models*, Twenty-First International Conference on Machine Learning, pp. 137–144, 2004.

[18] M. Mostofa, M. S. E. Saadabadi, S. R. Malakshan, and N. M. Nasrabadi, *Pose Attention-Guided Profile-to-Frontal Face Recognition*, 2022, [Online]. Available: http://arxiv.org/abs/2209.07001.