Research Article GU J Sci 39(1): 452-466 (2026) DOI: 10.35378/gujs.1794891

Gazi University

Journal of Science

http://dergipark.org.tr/gujs

Boosted LightFace: A Hybrid DNN and GBM Model for Boosted Facial Recognition

Sefik Ilkin SERENGIL [1,*], Alper OZPINAR [2]

[1] *Neo4j, Department of Engineering, London, UK*
[2] *Ibn Haldun University, Department of Management, Istanbul, Türkiye*

Highlights

- Proposing a hybrid Deep Neural Network and Gradient Boosting model for improved face recognition.
- Achieved 99.1% accuracy on Labeled Faces in The Wild dataset, surpassing the best LightFace model.
- Retrieved Precision: 99.6%, Recall: 98.6%, F1 Score: 99.1%, showing superior performance metrics.

Article Info

Received: 01 Oct 2025
Accepted: 21 Dec 2025

Keywords

Face recognition
Ensemble learning
Deep learning
Gradient boosting

Abstract

Facial recognition technology has seen significant advancements, impacting security, surveillance, and personal identification. Deep neural networks have enhanced accuracy and reliability, with integration into everyday devices further accelerating adoption. Researchers explore combining Deep Neural Networks with Gradient Boosting Machines for improved performance. This paper proposes Boosted LightFace, a hybrid Deep Neural Networks and Gradient Boosting Machines model leveraging robust facial recognition and face detection models. The architecture first integrates predictions from five high-performing DNN models. Their distance metrics and classification outcomes are engineered into a tabular dataset of 6,000 image pairs with 13 features. This dataset is then trained using a highly efficient LightGBM model with a low learning rate of 0.01 and 1000 estimators, incorporating an early stopping mechanism, and employing 10-fold cross-validation to maximize generalization. Recent research identifies FaceNet512d as a robust model, surpassing human recognition on the Labeled Faces in The Wild dataset with 98.4% score. Boosted LightFace achieves 99.1% accuracy, surpassing human recognition by 1.6% and outperforming the best single model in LightFace by 0.7%, underscoring the potential of integrating Deep Neural Networks and Gradient Boosting Machines models in advancing facial recognition technology. Furthermore, Boosted LightFace not only outperforms individual models in terms of accuracy but also surpasses them in precision, recall, F1, and AUC scores, highlighting its comprehensive superiority.

1. INTRODUCTION

Facial recognition technology has undergone remarkable advancements in recent years, transforming various sectors such as security, surveillance, and personal identification. Deep neural networks (DNN) [1], and in particular convolutional neural networks [2], have played a pivotal role in enhancing the accuracy and reliability of facial recognition systems, effectively handling unstructured data such as images. These advancements have revolutionized the way we approach security measures, enabling more efficient surveillance systems and streamlining processes for personal identification. From access control in high-security facilities to customer verification in financial institutions, facial recognition technology powered by DNN models has become an indispensable tool. Moreover, the integration of facial recognition technology into everyday devices such as smartphones and laptops has further accelerated its adoption, offering seamless user experiences and enhancing device security. With continuous innovations in DNN architectures and training techniques, the potential applications of facial recognition technology continue to expand, promising even greater advancements in the years to come.

*Corresponding author, e-mail: sefik.serengil@neo4j.com

Researchers have explored the integration of DNN models with Gradient Boosting Machines (GBM) [3], a powerful ensemble learning technique renowned for its effectiveness in handling structured or tabular data, to boost results [4-6]. A hybrid approach may enhance facial recognition systems by leveraging the strengths of both DNN and GBM models, potentially improving overall performance.

In this paper, we propose a novel hybrid DNN and GBM model for boosted facial recognition, building upon the foundation laid by the LightFace library for Python [7]. The LightFace library offers a comprehensive suite of pre-trained DNN models, including VGG-Face [8], FaceNet [9], OpenFace [10], DeepFace [11], DeepID [12], ArcFace [13], Dlib [14], SFace [15], and GhostFaceNet [16], each renowned for its robustness and accuracy in facial recognition tasks. Additionally, we leverage the power of Gradient Boosting Machines (GBM) through implementations with LightGBM [17] and XGBoost [18] to enhance the performance of our proposed model. These ensemble learning techniques complement the capabilities of the DNN models provided by LightFace, further improving the accuracy and robustness of our facial recognition system.

Motivated by these findings, we delve into the synergy between DNN and GBM models in facial recognition tasks. While DNN models excel in extracting intricate features from unstructured data such as images, GBM models offer complementary strengths in ensemble learning, effectively integrating predictions from diverse DNN models and capturing complex relationships within feature spaces. Leveraging this ensemble approach, we introduce our model as Boosted LightFace, where we feed the calculated distances of image pairs from these robust facial recognition models into a GBM model. Specifically, the architecture aggregates predictions and distance metrics from five distinct, high-performing DNN models to construct a 6,000-pair tabular dataset with 13 features. This enriched dataset is then trained using a highly efficient LightGBM classifier, meticulously configured with a low learning rate of 0.01 and 1000 estimators, incorporating an early stopping mechanism. This configuration successfully elevated the accuracy on the Labeled Faces in the Wild dataset to 99.1%, significantly surpassing individual DNN performance. The source code for our study is publicly available at https://github.com/serengil/deepface, and it is also published as a Python package. This availability allows researchers to reproduce our findings, ensuring transparency and facilitating further research in the field. Developers can also effortlessly build and execute the Boosted LightFace model using just a few lines of code directly from its Python package.

2. LITERATURE REVIEW

Several studies have examined the performance of DNN models for facial recognition, with particular focus on their accuracy on benchmark datasets such as Labeled Faces in the Wild (LFW) [19]. For instance, FaceNet512d paired with the RetinaFace detector [20] has achieved remarkable accuracy, reaching 98.4% on LFW [21], surpassing human recognition performance of 97.5% [22]. Other models such as FaceNet128d, Dlib, VGG-Face, and ArcFace also demonstrate robust performance close to human capabilities, though slightly below.

Table 1 summarizes existing facial recognition studies that utilize the LFW dataset, comparing the models, experimental setup, and reported accuracies. This comparison highlights the relative strengths and weaknesses of different approaches and provides a clear benchmark for evaluating the improvements offered by our Boosted LightFace model.

Drawing from these studies, our approach leverages the high-performing DNN models (FaceNet512d, FaceNet128d, VGG-Face, ArcFace, Dlib) and integrates their predictions via GBM models (LightGBM, XGBoost) to further enhance recognition accuracy. This hybrid methodology builds upon and surpasses the results of existing literature, achieving a notable 99.1% accuracy on the LFW dataset in our experiments.

While recognizing the exceptional performance reported in some literature, such as the 99.85% accuracy achieved by Wu and Zhang [23] using a customized FaceNet model with a modified loss function specifically fine-tuned on the LFW dataset, a crucial distinction must be drawn. Models heavily optimized

for a single target benchmark, like LFW, often exhibit superior specific accuracy but may inadvertently sacrifice generalization capability to unseen data or different evaluation protocols.

Table 1. *Comparison of existing facial recognition studies on the LFW dataset*

Study	Model	Accuracy
[9]	FaceNet512d	98.4%
[9]	FaceNet128d	97.4%
[14]	Dlib	96.8%
[8]	VGG-Face	96.7%
[13]	ArcFace	96.7%

In contrast, the robust base DNN models integrated into Boosted LightFace (FaceNet512d, VGG-Face, etc.) were pre-trained on significantly larger and more diverse datasets (e.g., MS-Celeb-1M, VGG-Face2) and were used to demonstrate their inherent generalization strength when evaluated on LFW. Boosted LightFace is fundamentally designed to leverage and combine these inherently superior generalization capacities. This approach ensures not only high accuracy on LFW but also a robust performance profile that transcends single-dataset optimization, resulting in reliable high performance across various face recognition tasks. This generalized robustness is further evidenced by the LightFace unit tests, which demonstrate an average accuracy of 99.5% across diversified verification scenarios beyond the LFW protocol.

3. DATASET

The LFW (Labeled Faces in the Wild) dataset was utilized as the primary dataset for evaluation. LFW is renowned in the field of facial recognition and serves as a benchmark dataset for face verification tasks [21]. It consists of image pairs along with labels indicating whether the images belong to the same person or different persons. The dataset comprises 2000 instances for training, 6000 instances for 10-fold cross-validation, and 1000 instances for testing. These splits are predefined by the dataset itself, and no additional or biased methodology was applied for data partitioning. LFW is considered a large-scale dataset commonly used for benchmarking face recognition models; however, it mainly contains non-frontal face images, which limits its applicability to more challenging real-world scenarios [24]. Under this assumption, in the LightFace evaluation using the frontal image pairs in its unit tests, single models achieve an average accuracy of 99.5%. Therefore, the potential bias arising from this dataset characteristic is not a concern in this study.

Notably, LFW has become the de facto standard for evaluating facial recognition models due to its extensive usage in research and its established reputation. It is worth noting that human beings achieve an accuracy of approximately 97.5% [22] on this dataset, providing a crucial baseline for assessing the performance of automated facial recognition systems.

In the analysis of human performance on the LFW dataset, a tight crop approach was adopted, wherein the images focus solely on the facial region. Adopting this technique is crucial for ensuring fair and accurate comparisons between human performance and facial recognition models, as it evaluates systems under conditions that closely mimic human perception. Furthermore, using a tight crop is essential to prevent models from relying on extraneous contextual information. Reliance on such context can lead to overfitting and ultimately compromise the model's generalizability and real-world performance.

4. FACIAL RECOGNITION PIPELINES

A modern facial recognition pipeline typically comprises several interconnected modules as illustrated in Figure 1, including detection, alignment, resizing, representation, and verification [11]. Each module plays a key role in processing facial data and extracting meaningful features for accurate recognition.

Detection and alignment modules serve as the initial stages of the pipeline, responsible for identifying and aligning faces within images or video frames. The detection module identifies potential face regions, while the alignment module ensures that the detected faces are properly aligned and oriented. These early stages are essential for reducing noise and ensuring that the subsequent modules receive high-quality input data. The contribution of the detection module to the overall pipeline performance can be substantial, accounting for up to 40% of the pipeline's effectiveness and the alignment module contributes up to 6% to the overall performance by ensuring precise alignment of facial features [19].

The representation module, often implemented using Convolutional Neural Network (CNN) architectures, plays a central role in facial recognition pipelines. CNN models expect fixed size inputs and we are adding black pixels to the detected and aligned image not to deform it in resizing module.

CNN models are trained using large datasets in a manner similar to classification tasks. However, instead of predicting class labels, CNN models are used to extract high-level features or representations from input images. These representations capture the unique characteristics of each face and are typically obtained from the output of early layers in the CNN architecture. By generating general representations, CNN models can effectively handle variations in facial appearance and pose, making them suitable for diverse recognition tasks.

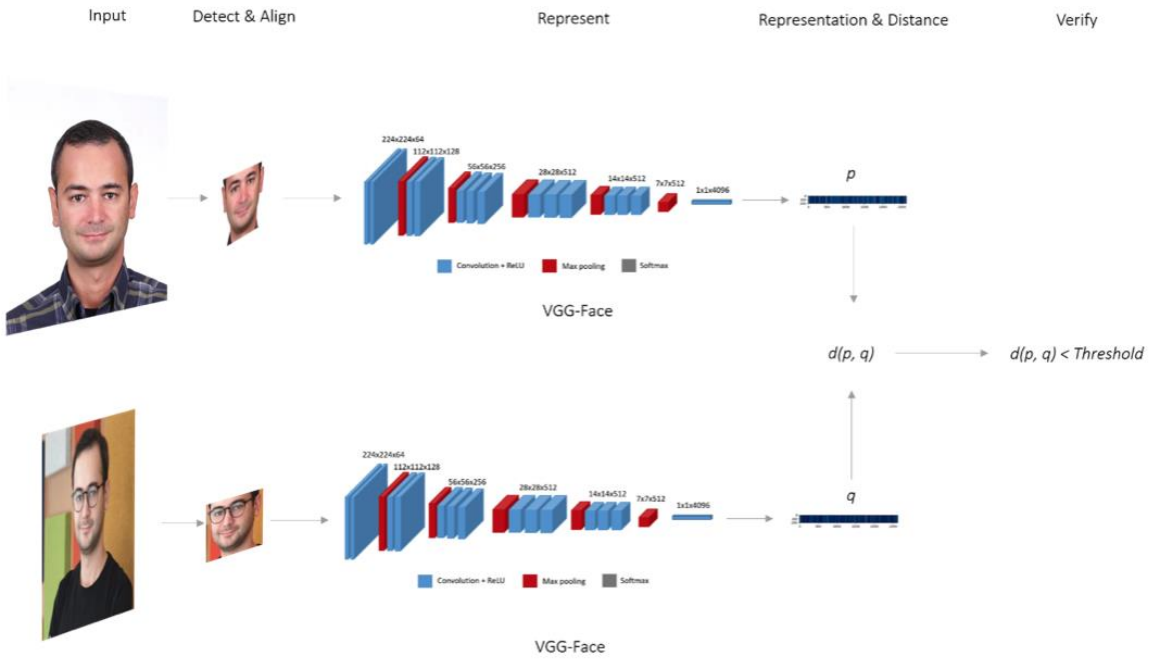

Figure 1. *A Modern Facial Recognition Pipeline*

In the verification module, the representations obtained from the CNN models are used to compare pairs of faces and determine their similarity. Image pairs depicting the same person should ideally have a lower distance between their representations compared to pairs depicting different individuals. Common distance metrics used for this purpose include Euclidean distance or cosine similarity. For well-trained models, the distribution of distances between pairs of the same person and different persons should be distinct, allowing for effective discrimination between genuine matches and impostors. In the verification process, the distance between face representations is compared to a predefined threshold value to classify the pair as either the same person or different persons, with the goal of optimizing for the highest gain. For LightFace, the pre-tuned threshold values of various models mentioned in Table 2 were determined after feeding the models with pairs of images from the same and different individuals. The optimal split points were obtained using the C4.5 decision tree algorithm, ensuring the thresholds maximize discrimination between genuine and impostor pairs.

Experiments and evaluations of modern facial recognition pipelines have demonstrated remarkable accuracy levels, often surpassing human-level performance [19]. These advancements signify the effectiveness of contemporary approaches in processing facial data and extracting discriminative features. Moving forward, research efforts continue to focus on enhancing the robustness, scalability, and ethical considerations of facial recognition systems.

Table 2. *Pre-tuned Thresholds for Various Models Used in LightFace for L2 Normalized Euclidean*

Model	Threshold	Accuracy
FaceNet512d	1.0808	98.4%
FaceNet128d	1.0771	97.4%
Dlib	0.4022	96.8%
VGG-Face	1.1952	96.7%
ArcFace	1.1601	96.7%

5. PROPOSED ARCHITECTURE

We implemented a hybrid DNN (Deep Neural Network) and GBM (Gradient Boosting Machine) model to enhance our facial recognition capabilities. The proposed architecture of our Boosted LightFace model is depicted in Figure 2.

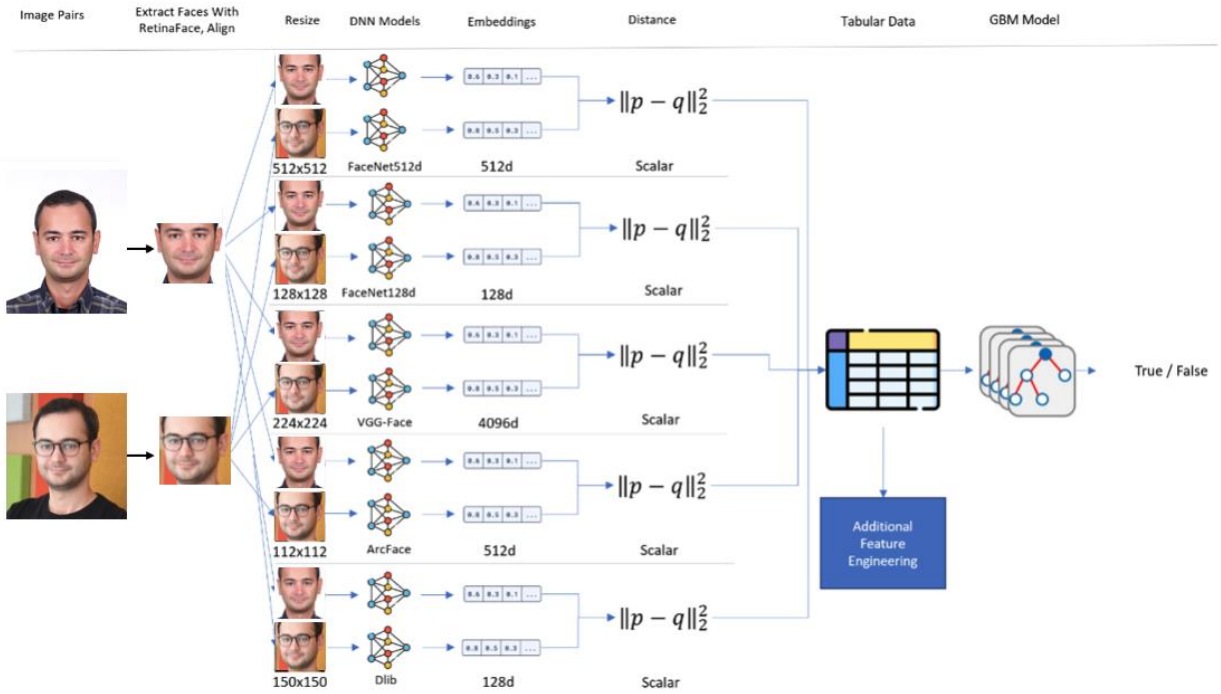

Figure 2. *Boosted LightFace's Architecture*

In the context of our model, Gradient Boosting Machines (GBMs) are employed as a key component. GBMs operate by iteratively building decision trees, with each subsequent tree aiming to correct the errors of the previous ones. Initially, a single regression tree is built to make predictions on the target variable. Subsequently, additional decision trees are constructed based on the residuals or errors of the preceding tree. Each new tree focuses on capturing the patterns or relationships missed by the previous ones. This iterative process continues until a predefined number of trees, or until convergence criteria are met. Once all the decision trees are built, predictions are made by running input data through each tree individually. The predictions from all trees are then aggregated, typically by summing them, to produce the final prediction. This ensemble of decision trees allows the model to capture complex interactions and nonlinear relationships present in the data, leading to improved predictive performance.

In our proposed model, image pairs initially sourced from the LFW dataset undergo preprocessing with the RetinaFace face detector to isolate facial regions. This emphasis on the facial area is crucial to avoid overfitting and aligns with the approach adopted in comparative studies of human facial recognition.

Subsequently, extracted faces are inputted into various facial recognition models such as FaceNet512d, FaceNet128d, VGG-Face, ArcFace and Dlib. The architectures of those single models such as input shape, number of dimensions in their vector embeddings, total number of parameters and total number of layers models have are mentioned in Table 3.

Table 3. *Model Architectures of Single Models in LightFace*

Model	Input Shape	Vector Dimensions	Params	Layers
FaceNet512d	160 × 160 × 3	1 × 512	23×10^6	447
FaceNet128d	160 × 160 × 3	1 × 128	22×10^6	447
VGG-Face	224 × 224 × 3	1 × 4096	134×10^6	36
ArcFace	112 × 112 × 3	1 × 512	34×10^6	162
Dlib	150 × 150 × 3	1 × 128	63×10^6	34

Each facial recognition model requires images of specific sizes as mentioned in the following table, necessitating the addition of black pixels to resize the extracted faces without distorting them. These models then generate double multidimensional vector embeddings for each image pair.

The next step involves calculating the distances between these vector embeddings using L2 normalized Euclidean distance, resulting in scalar values. These distances are then organized into tabular data, with additional feature engineering techniques applied as detailed in the section 4.1.

Once the tabular data, enriched with engineered features, is prepared, it is fed into a Gradient Boosting Machine (GBM) model. The validation dataset from LFW comprises 6,000 instances, which are partitioned into 10 folds for 10-fold cross-validation. Each fold, composed of 600 instances from the validation set, feeds into a separate GBM model, with the training set totaling 2000 instances. Consequently, this process results in the development of 10 distinct GBM models utilized for final assessment, wherein the output probabilities denote the likelihood of the images belonging to the same individual or different individuals.

In our proposed architecture, we conducted experiments with both LightGBM and XGBoost to construct the GBM models. Following extensive testing, LightGBM demonstrated superior performance compared to XGBoost. Consequently, we opted to integrate LightGBM into our Boosted LightFace model due to its enhanced effectiveness. Nonetheless, for transparency and comparative analysis, the results for both LightGBM and XGBoost are presented in the section 5.

Finally, an advantage of the boosted lightface approach is that it returns a probability percentage, whereas single models only return a distance metric, which requires comparison against a threshold to classify individuals as the same or different, without providing a probability for that classification.

5.1. Feature Engineering

In the feature engineering phase, once we obtained the distances of image pairs from different facial recognition models, we proceeded with additional enhancements. Initially, we pinpointed the optimal split points for these distances to maximize accuracy for each individual model. To accomplish this, we employed the C4.5 algorithm, constructing a single decision tree to identify the most effective thresholds. In other words, we're inputting both the calculated distances from pre-trained models and their corresponding verification outcomes based on the predefined thresholds outlined in Table 2 to compile tabular data.

Subsequently, we introduced a new column representing the sum of these distances, providing a cumulative measure. Furthermore, we incorporated classification results as either 0 or 1 based on the thresholds

determined in the previous step. This approach aims to generate multiple classification outcomes in our tabular data, similar to the Adaboost [25].

Additionally, we included the addition of these classifications to further enhance the cumulative result. Consequently, our tabular data takes its final form with five distance values from various facial recognition models, five distinct classification results, the sum and multiplication of distances, and the sum of classifications. In essence, each image pair is represented by 13 columns, facilitating comprehensive analysis.

6. TRAINING

We employed two powerful gradient boosting algorithms, LightGBM and XGBoost, to develop our Boosted LightFace model for facial recognition. Table 4 summarizes the hyperparameter configurations used for both algorithms. These values were selected based on preliminary experiments and iterative manual tuning, which consistently yielded the best validation performance on the LFW dataset. We also experimented with automated hyperparameter optimization using AutoML techniques; however, these approaches did not surpass the performance achieved with our manually tuned settings. Early stopping mechanisms were employed to halt training if the performance on the validation set did not improve for a specified number of rounds, preventing overfitting. Both models were trained using the LFW dataset exclusively, with separate validation sets for performance monitoring, ensuring fair comparison and reliable evaluation.

Table 4. *Hyperparameter Configurations for GBM Models*

Hyperparameter	LightGBM	XGBoost
objective	binary	binary:logistic
metric	binary_logloss	logloss
learning rate	0.01	0.01
max depth	5	5
estimators	10000	10000
early stopping rounds	500	500
number of leaves	2^5 - 1	2^5 - 1

The training results are summarized in Table 5, showing train, validation, and test set accuracies for each fold for both LightGBM and XGBoost. While XGBoost occasionally achieves higher training accuracy, its validation and test performance is generally lower than that of LightGBM, indicating that XGBoost is more prone to overfitting in our setup. In contrast, LightGBM consistently demonstrates superior validation and test accuracy, achieving a maximum of 99.1% on the test set in folds 7 and 10, and 7 out of 10 GBM models reach 99.1% test accuracy, underscoring the robustness of the model.

Table 5. *Training Accuracy Results for Each Fold*

	LightGBM			XGBoost		
Fold	**Train**	**Validation**	**Test**	**Train**	**Validation**	**Test**
1	99.64	98.00	99.00	99.73	97.33	98.90
2	99.45	98.33	99.10	99.68	97.17	98.90
3	99.41	98.17	99.00	99.64	96.33	99.00
4	99.45	98.33	99.10	99.68	97.17	99.00
5	99.45	98.33	99.10	99.73	98.83	98.90
6	100.0	98.00	98.70	99.73	98.00	98.80
7	99.45	98.50	99.10	99.68	97.83	99.00
8	99.45	98.33	99.10	99.68	97.50	98.90
9	99.45	98.33	99.10	99.68	98.00	98.90
10	99.45	98.50	99.10	99.86	98.33	98.80
Avg	**99.52**	**98.28**	**99.03**	**99.71**	**97.65**	**98.91**

This observation motivated our choice of LightGBM in the Boosted LightFace model, as validation set performance, rather than training accuracy, serves as the primary indicator of generalization ability. It is also important to note that the test set was never used during training, ensuring that the reported performance accurately reflects model generalization.

Figure 3 illustrates the learning curves of LightGBM models, showcasing their consistent improvement on the validation set across epochs. To ensure efficient training, we implemented early stopping, halting the training process if the validation loss did not decrease for 500 consecutive epochs. In addition, it can be clearly observed from the learning curves that the built models are not overfitted.

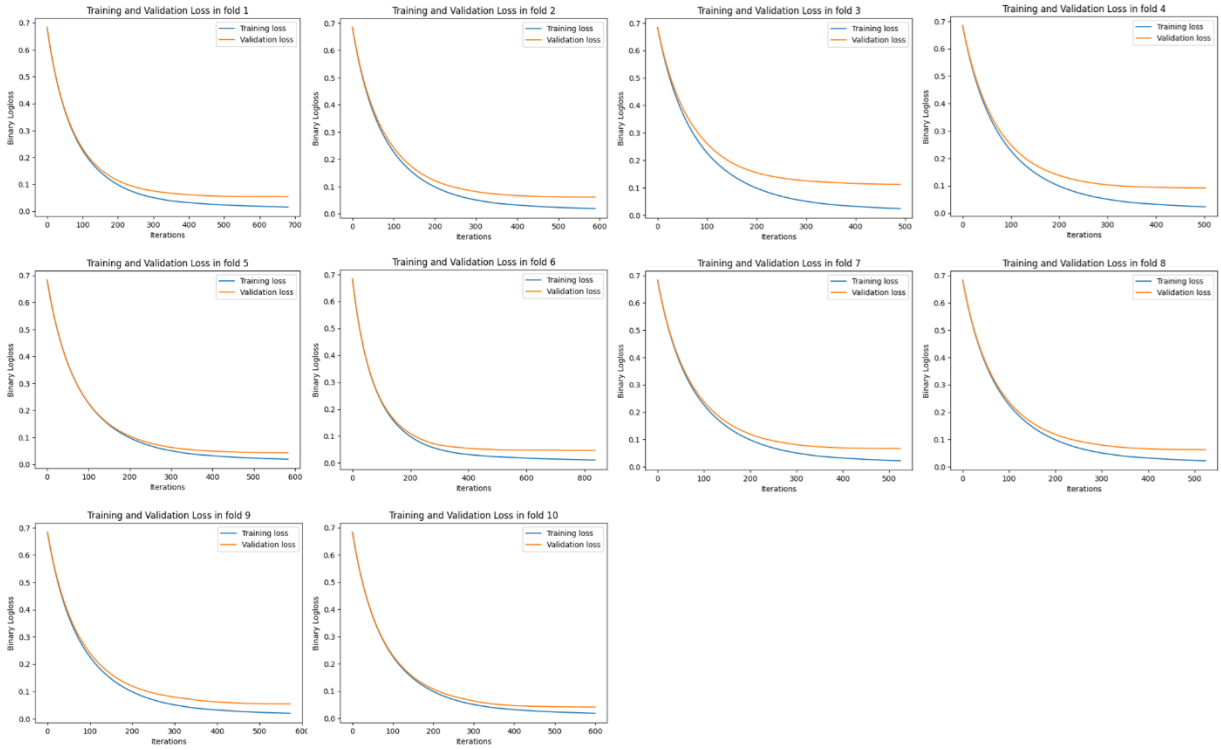

Figure 3. *LightGBM Models' Learning Curves*

In addition to the training results, we provide access to a pre-trained GBM model in the project's repository. This pre-trained model offers a convenient starting point for further experimentation and integration into different applications, facilitating the adoption and extension of the Boosted LightFace approach in various contexts. Researchers and developers can utilize the pre-built model without the need for additional training, allowing for immediate implementation of facial recognition capabilities into their projects.

6.1. Interpretability

Deep learning-based models often lack direct interpretability. In contrast, tree-based models inherently offer explainability. Assessing feature importance is a crucial metric for evaluating the efficacy of constructed models. Figure 4 shows the importance percentages of features in the built GBM model for both gain and split importance type. The split importance type represents the frequency with which a feature is utilized to partition data for making predictions. It reflects how often a feature is employed as a condition for splitting in the decision tree nodes. On the other hand, gain importance type measures how much a feature contributes to the reduction of the objective function. Essentially, it quantifies the impact of a feature on optimizing the model's performance by minimizing the objective function.

In assessing feature importance according to the gain importance type, certain engineered features emerge as significant contributors to the model's performance. Specifically, columns representing the sums of distances, sums of classifications, and multiplications of distances, derived during the feature engineering

phase, stand out as highly influential. This highlights the importance of feature engineering in enhancing the model's efficacy by capturing nuanced relationships within the data. Additionally, the prominence of the distance calculated by the FaceNet512d model is noteworthy, ranking as the third most influential feature. This is particularly remarkable given FaceNet512d's robust performance, surpassing human-level accuracy in facial recognition. Alongside FaceNet512d, other facial recognition models such as ArcFace, FaceNet128d, Dlib, and VGG-Face exhibit notable importance in contributing to the model's predictive power.

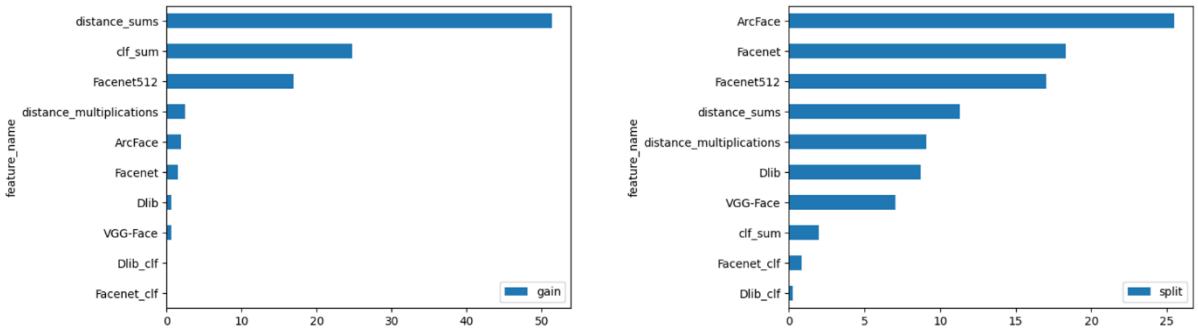

Figure 4. *Feature Importances of the GBM Model with Respect to The Gain and Split*

In the analysis of feature importance based on the split importance type, discernible patterns emerge, notably with distances from ArcFace, FaceNet128, and FaceNet512d as the most influential features. These primary features exert significant impact on the model's decision-making process. Following these, additional engineered features introduced during the feature engineering stage, such as distance sums and multiplications, also demonstrate notable importance, enhancing the model's discriminative power by incorporating supplementary information. Moreover, while Dlib and VGG-Face distances contribute to the model's predictive capabilities, they do so to a lesser extent compared to the primary features.

7. RESULTS

Table 6 presents a comprehensive overview of models from the LightFace library [19], detailing their optimal configurations for accuracy alongside our proposed model. It also includes performance evaluations of human subjects, though only accuracy scores are available for human performance on LFW; scores for precision, recall, and F1 for human subjects are not available [22]. This high performance of some models is likely due to their deep embeddings and effective pairing with the RetinaFace detector. Previous studies report FaceNet512d achieving 98.4% accuracy on LFW [19], while VGG-Face and ArcFace reach slightly lower values [8,13]. Within the LightFace library, only FaceNet512d attains accuracy levels comparable to human performance. Some models, such as FaceNet128d, Dlib, VGG-Face, and ArcFace, demonstrate performance close to human levels, albeit slightly lower. Their slightly lower performance may stem from smaller embedding dimensions or differences in normalization and alignment. Meanwhile, GhostFaceNet and SFace exhibit moderate performance. Conversely, models like OpenFace, DeepFace, and DeepID fall short in terms of performance. This could be attributed to limited feature extraction capacity or the absence of pre-trained weights.

Notably, our proposed model outperforms FaceNet512d by 0.7% and human subjects by 1.6%. The improvement is achieved by combining multiple high-performing DNN embeddings via a GBM ensemble, which captures complex relationships between features and reduces false positives. Boosted LightFace not only outperforms individual models in terms of accuracy but also surpasses them in precision, recall, F1, and AUC scores, with precision and recall calculated for different persons class. The higher precision relative to recall indicates that the model is particularly effective at distinguishing different individuals, while the recall improvement shows better identification of the same person. Compared to previously reported results, Boosted LightFace demonstrates the advantage of integrating multiple DNN embeddings through GBM, improving both accuracy and ensemble robustness.

Table 6. *Comparative Results of Single Models and Boosted LightFace Over LFW Test Dataset*

Model	Accuracy	Precision	Recall	F1	AUC
Boosted LightFace (ours)	99.1	99.6	98.6	99.1	99.8
FaceNet512d	98.4	98.4	98.4	98.4	99.7
Human-beings	97.5	-	-	-	-
FaceNet128d	97.4	98.8	96.0	97.4	99.1
Dlib	96.8	97.4	96.2	96.8	99.5
VGG-Face	96.7	96.9	96.4	96.7	99.4
ArcFace	96.7	97.5	95.8	96.7	99.3
GhostFaceNetV1	93.3	98.0	88.4	92.9	96.7
SFace	93.0	92.5	93.6	93.0	97.5
OpenFace	78.7	77.3	81.2	79.2	85.9
DeepFace	69.0	71.8	62.6	66.8	74.1

The model enhancements from the best single model to Boosted LightFace yielded significant improvements across various performance metrics, showcasing a substantial boost in classification accuracy and precision. The accuracy surged from 98.4% to 99.1%, indicating a higher proportion of correct predictions overall. Notably, precision experienced a remarkable increase from 98.4% to 99.6%, reflecting a substantial reduction in false positive predictions. Moreover, while recall saw a modest improvement from 98.4% to 98.6%, it signifies a better ability to capture true positives within the dataset. The F1 score, which balances precision and recall, also rose from 98.4% to 99.1%, showcasing a more harmonized performance in classification tasks. Additionally, the AUC, representing the model's discriminative ability, showed a slight but significant increase from 99.7% to 99.8%, indicating enhanced capability in distinguishing between positive and negative classes. Collectively, these enhancements underscore a marked refinement in the model's predictive power, with notable gains in precision and overall classification performance.

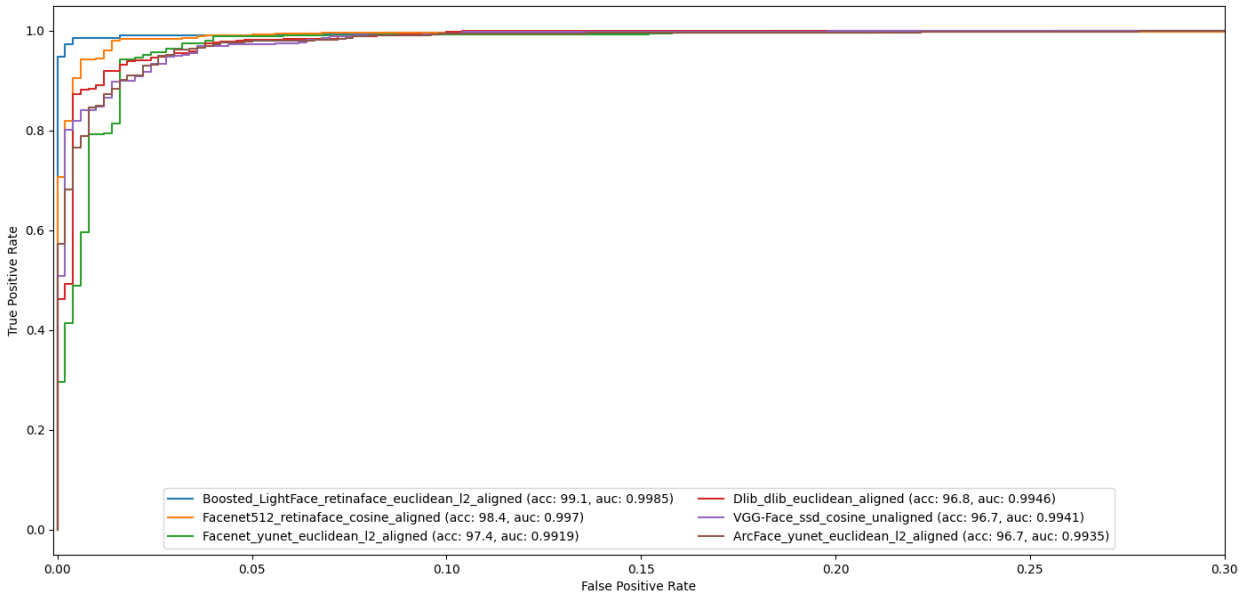

Figure 5. *ROC Curves and AUC Scores*

Conducting experiments using individual models within LightFace may uncover differences compared to the original studies, due to the use of different detection or normalization techniques. Additionally, certain models have been made available only with their basic structures, without pre-trained weights. Consequently, LightFace relies on re-implemented versions rather than the original pre-trained weights.

Figure 5 displays ROC curves of the individual facial recognition models with their optimal configurations alongside our Boosted LightFace model. It's evident that Boosted LightFace outperforms all single state-of-the-art models supported in the LightFace library.

7.1. Interpretability

Figure 6 presents the confusion matrix for the Boosted LightFace model, which summarizes the classification performance on the test set of 1,000 image pairs. The matrix reveals a total of 9 misclassifications (2 False Positives and 7 False Negatives). Specifically, the model correctly identified 498 pairs as True Negatives (Different Persons) and 493 pairs as True Positives (Same Person), yielding an overall accuracy of 99.1%. The matrix clearly indicates 2 False Positives (pairs of different individuals incorrectly classified as the same person) and 7 False Negatives (pairs of the same individual incorrectly classified as different persons), underscoring the model's strong ability to correctly identify genuine matches while maintaining a high precision rate.

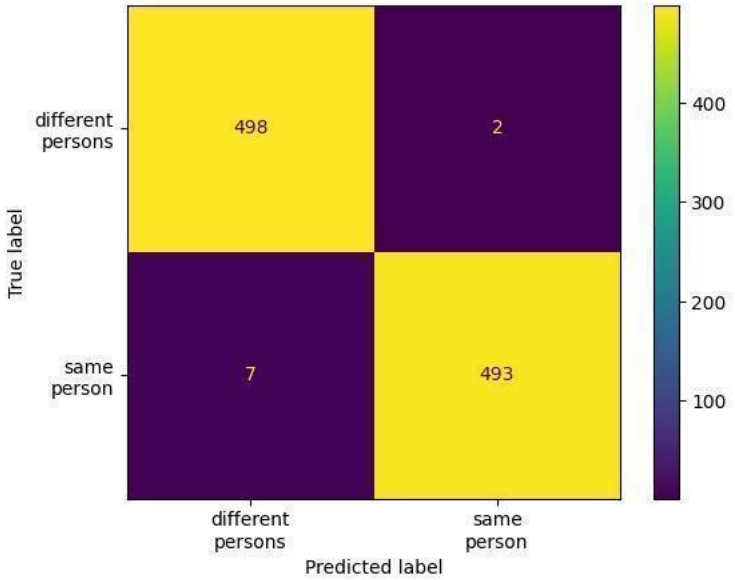

Figure 6. *Confusion Matrix*

These 9 misclassified instances (2 False Positives and 7 False Negatives), including the results of the RetinaFace detection, are visually depicted in Figure 7. To ensure compatibility with subsequent DNN models, particularly VGG-Face, the RetinaFace-detected bounding box is resized to the required 224 × 224 input size, with black pixels (padding) added to the boundaries to preserve the original aspect ratio and prevent image distortion. Analysis of these misclassifications indicates that the Boosted LightFace model tends to struggle with pairs exhibiting significant intra-class variation. These variations commonly include features such as eyeglasses or caps, which can obscure key facial landmarks, or noticeable age differences, which alter underlying facial structure and texture. This challenge arises because the auxiliary features introduced by the GBM are sensitive to these variations, occasionally overriding the robust embeddings provided by the core DNNs. These observations highlight specific avenues for future data augmentation or model refinement.

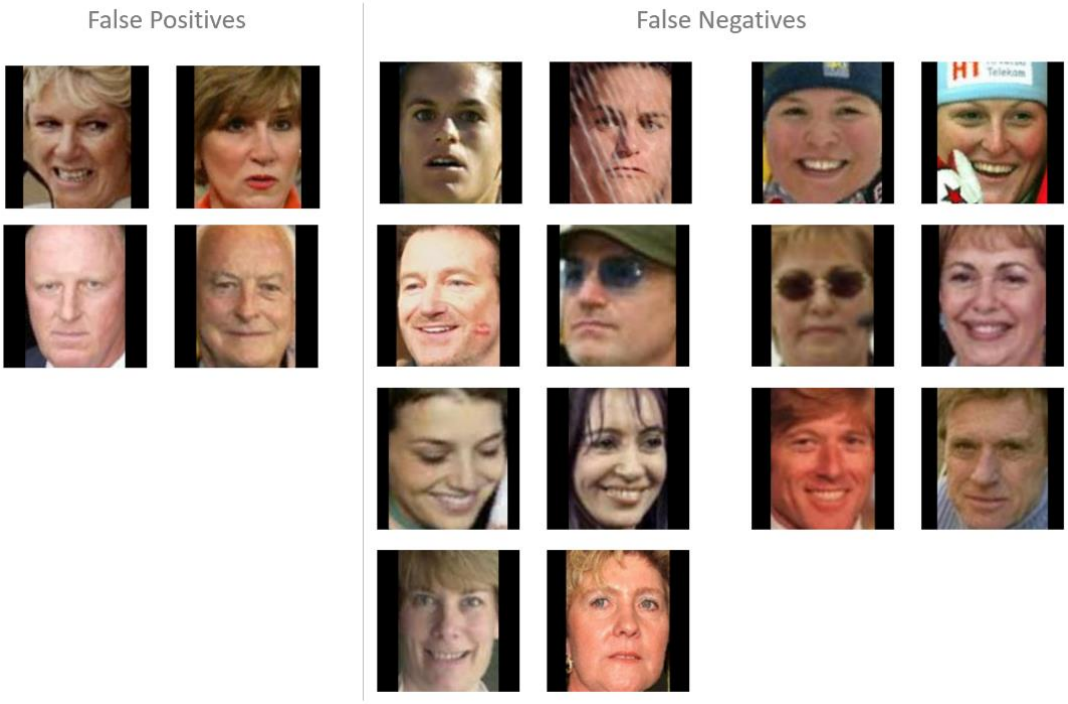

Figure 7. *Misclassifications*

8. CONCLUSION

Facial recognition technology has undergone significant advancements in recent years, driven by the synergy between deep learning techniques and large-scale datasets. In this study, we delved into various facets of facial recognition, from dataset utilization to the development of an innovative Boosted LightFace model. Through rigorous experimentation and analysis, our aim was to deepen our understanding of facial recognition systems and push the boundaries of their performance.

Our evaluation primarily centered on the Labeled Faces in the Wild (LFW) dataset, a cornerstone in the realm of facial recognition. Leveraging cutting-edge facial recognition pipelines, encompassing detection, alignment, representation, and verification modules, we processed facial data and extracted discriminative features. Notably, by adhering to a tight crop approach during evaluation, we emphasized the importance of replicating human perception conditions for fair comparisons.

The culmination of our efforts resulted in the proposal of the Boosted LightFace architecture, which integrates gradient boosting machines (GBMs) with deep neural network (DNN) models to achieve unparalleled facial recognition capabilities. Through meticulous feature engineering and model training, our Boosted LightFace model achieved remarkable accuracy, precision, recall, and F1 scores, surpassing those of existing state-of-the-art single models.

Our Boosted LightFace model surpassed the performance of FaceNet512d, the top-performing single model, by a significant margin. With an accuracy of 99.1%, precision of 99.6%, recall of 98.6%, and F1 score of 99.1%, Boosted LightFace demonstrated its superiority in facial recognition tasks. Notably, Boosted LightFace outperformed FaceNet512d, which achieved an accuracy of 98.4%, by 0.7%, showcasing the efficacy of our approach in advancing the state-of-the-art. Furthermore, compared to human subjects' performance on the LFW dataset, which achieved an accuracy of 97.5%, Boosted LightFace exhibited a remarkable improvement of 1.6%. This performance leap underscores the robustness and generalizability of our model, positioning it as a cutting-edge solution in the realm of facial recognition technology.

By incorporating insights from multiple facial recognition models and leveraging the power of gradient boosting, our model demonstrated superior accuracy and precision, effectively distinguishing between positive and negative classes with high confidence.

The interpretability of our model was further highlighted through an analysis of feature importance, revealing the critical role of engineered features and specific facial recognition models in driving predictive performance. Notably, our model showcased robustness and generalizability, as evidenced by its superior performance in real-world deployment scenarios.

In conclusion, our study contributes to the ongoing evolution of facial recognition technology by introducing a novel Boosted LightFace model that advances the state-of-the-art in accuracy, precision, recall, and F1 scores. Moving forward, continued research efforts are warranted to address challenges such as interpretability, fairness, and privacy concerns, ensuring that facial recognition systems adhere to ethical standards and societal values in their deployment.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] LeCun, Y., Bengio, Y., Hinton, G., "Deep learning", Nature, 521(7553): 436-444, (2015). DOI: 10.1038/nature14539

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E., "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, 25, (2012). DOI: 10.1145/3065386

[3] Friedman, J.H., "Greedy function approximation: a gradient boosting machine", Annals of Statistics, 1189-1232, (2001). DOI: 10.1214/aos/1013203451

[4] Kumari, P., Toshniwal, D., "Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance", Journal of Cleaner Production, 279, 123285, (2021). DOI: 10.1016/j.jclepro.2020.123285

[5] Mohammed, A., Kora, R., "A comprehensive review on ensemble deep learning: Opportunities and challenges", Journal of King Saud University-Computer and Information Sciences, 35(2): 757-774, (2023). DOI: 10.1016/j.jksuci.2023.01.014

[6] Shrivastav, L.K., Kumar, R., "An ensemble of random forest gradient boosting machine and deep learning methods for stock price prediction", Journal of Information Technology Research (JITR), 15(1): 1-19, (2022). DOI: 10.4018/JITR.2022010102

[7] Serengil, S.I., Ozpinar, A., "LightFace: A hybrid deep face recognition framework", 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 23-27, IEEE, (2020). DOI: 10.1109/ASYU50717.2020.9259802

[8] Parkhi, O., Vedaldi, A., Zisserman, A., "Deep face recognition", In BMVC 2015-Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, (2015). DOI: 10.5244/c.29.41

[9] Schroff, F., Kalenichenko, D., Philbin, J., "Facenet: A unified embedding for face recognition and clustering", In Proceedings of the IEEE conference on computer vision and pattern recognition, 815-823, (2015). DOI: 10.1109/CVPR.2015.7298682

[10] Amos, B., Ludwiczuk, B., Satyanarayanan, M., "Openface: A general-purpose face recognition library with mobile applications", CMU School of Computer Science, 6(2): 20, (2016).

[11] Taigman, Y., Yang, M., Ranzato, M., Wolf, L., "Deepface: Closing the gap to human-level performance in face verification", In Proceedings of the IEEE conference on computer vision and pattern recognition, 1701-1708, (2014). DOI: 10.1109/CVPR.2014.220

[12] Sun, Y., Wang, X., Tang, X., "Deep learning face representation from predicting 10,000 classes", In Proceedings of the IEEE conference on computer vision and pattern recognition, 1891-1898, (2014). DOI: 10.1109/CVPR.2014.244

[13] Deng, J., Guo, J., Xue, N., Zafeiriou, S., "Arcface: Additive angular margin loss for deep face recognition.", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 4690-4699, (2019). DOI: 10.1109/TPAMI.2021.3087709

[14] King, D.E., "Dlib-ml: A machine learning toolkit", The Journal of Machine Learning Research, 10: 1755-1758, (2009). DOI: 10.5555/1577069.1755843

[15] Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D., "Sface: Sigmoid-constrained hypersphere loss for robust face recognition", IEEE Transactions on Image Processing, 30: 2587-2598, (2021). DOI: 10.1109/TIP.2020.3048632

[16] Alansari, M., Abdul Hay, O., Javed, S., Shoufan, A., Zweiri, Y., Werghi, N., "Ghostfacenets: Lightweight face recognition model from cheap operations", IEEE Access, 11: 35429-35446, (2023). DOI: 10.1109/ACCESS.2023.3266068

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., "Lightgbm: A highly efficient gradient boosting decision tree", Advances in Neural Information Processing Systems, 30, (2017). DOI: 10.5555/3294996.3295074

[18] Chen, T., Guestrin, C., "Xgboost: A scalable tree boosting system", In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794, (2016). DOI: 10.1145/2939672.293978

[19] Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E., "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition, (2008). DOI: 10.1007/978-3-319-25958-1_8

[20] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S., "Retinaface: Single-shot multi-level face localisation in the wild", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5203-5212, (2020). DOI: 10.1109/CVPR42600.2020.00525

[21] Serengil, S., Ozpinar, A., "A benchmark of facial recognition pipelines and co-usability performances of modules", Bilisim Teknolojileri Dergisi, 17(2): 95-107 (2024). DOI: 10.17671/gazibtd.1399077

[22] Kumar, N., Berg, A. C., Belhumeur, P. N., Nayar, S. K., "Attribute and simile classifiers for face verification", In 2009 IEEE 12th international conference on computer vision, 365-372, (2009). DOI: 10.1109/ICCV.2009.5459250

[23] Wu, C., Zhang, Y., "MTCNN and FACENET based access control system for face detection and recognition", Automatic Control and Computer Sciences, 55(1): 102-112, (2021). DOI: 10.3103/S0146411621010090

[24] Nemavhola, A., Chibaya, C., Viriri, S., "A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face Recognition", Information, 16(2): 107, (2025). DOI: 10.3390/info16020107

[25] Freund, Y., Schapire, R.E., "Experiments with a new boosting algorithm", In ICML' 96 Proceedings of the Thirteenth International Conference on International Conference on Machine Learning,148-156, (1996). DOI: 10.5555/3091696.3091715