Kanupriya Upadhyay, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Lightweight Hybrid HOG–MobileNetV2–ACK-ELM Framework for Real-Time Facial Recognition

Kanupriya Upadhyay[1], Manisha Sharma[2]
Dept. of Artificial Intelligence and Data Science
Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India.

Abstract- In settings where resources are limited, traditional automated attendance systems have a lot of trouble because they are slow to respond and are sensitive to changes in the environment. Li et al. research presents an enhanced facial recognition framework that mitigates the efficiency deficit in lightweight segmentation as identified [1] . Ryando et al. pointed out To improve security against the spoofing vulnerabilities [2] in healthcare kiosks, we add a liveness detection module that is based on the Eye Aspect Ratio (EAR) looked at heavy deep learning architectures that put a lot of strain on computers. In contrast, our system uses an Adaptive Circular Kernel Extreme Learning Machine (ACK-ELM) to achieve a non-iterative, one- shot learning paradigm. This method is based on the fast hybrid ideas of Anil and Suresh [4] and combines Histogram of Oriented Gradients (HOG) with shallow MobileNetV2 features. To ensure resilience against the lighting and expression varia- tions examined by Abdallah et al., the model is validated on the Extended Yale B dataset. The practical deployment phase comes after the five-phase attendance marking architecture that Potdar et al. [6] proposed. It uses a MySQL database to keep track of records in real time. Our classification strategy also works better than the Support Vector Machine (SVM) baselines set by Ali et al.[8] because it makes inferences faster. By using the few-shot learning efficiencies that Nasralla [9] looked into in the AIFS framework, the system stays very accurate even when it doesn't have a lot of training data. To make the best use of memory, we use the Global Average Pooling (GAP) techniques suggested by Wei et al. [9]. These techniques compress features to keep the system from crashing. Lastly, the system uses the multimodal fusion logic of Abdul-Al et al. [10] and the temporal consistency principles of Interno` et al. [11] to tell the difference between real facial trajectories and fake or static ones. Experimental results show that the accuracy is 97.16% on a standard CPU, which makes it a useful solution for large-scale institutional attendance.

Keywords: Face recognition, ACK-ELM, Computer Vision, Machine Learning, Real-time systems, Attendance system.

I. INTRODUCTION

The quick growth of biometric technology has changed how institutions are run, especially when it comes to automated attendance systems. People are more and more replacing traditional manual methods because they are prone to mistakes and proxy attendance. But modern facial recognition systems have a big problem with "efficiency-accuracy." Li et al. [1] found that traditional segmentation models often have a lot Identify applicable funding agency here. If none, delete this of latency when they are used on mobile or resource-limited devices. The security of these systems is also very important. Ryando et al. [2] point out that facial recognition is safe, but it is still open to advanced spoofing attacks, so strong liveness detection is needed.

Convolutional Neural Networks (CNNs), which are a part of deep learning, have set very high standards for accuracy. However, as Singh [3] pointed out, these architectures put a lot of strain on computers, which often need expensive GPU hardware, making them impractical for most institutional infrastructure. To address this deficiency, recent research has investigated hybrid paradigms. According to Anil and Suresh [4], the "fast hybrid" principles are becoming more important. We need models that reduce training epochs without losing accuracy. Fisherfaces [5] and other classical methods can handle some changes in lighting, but they don't have the deep feature representation needed for complex real-world settings. The current study is based on the

implementation of an innovative lightweight facial recognition mechanism that can be used for effective attendance checking in real time. This framework comprises a sequence of five stages that in- clude database preparation, face detection, liveness detection, recognition, and marking, just like in the case of the model described by Potdar et al. [6]. Unlike the traditional models of face recognition, including SVM, our algorithm surpasses the performance levels achieved by previous algorithms such as the SVM method used by Ali et al. [7] due to its analytical nature.

One of the major breakthroughs achieved through this approach is the inclusion of GAP, which is proven effective in model parameter reduction and avoiding system crashes due to insufficient memory allocation while performing mathematical operations on matrices as demonstrated by Wei et al. [9]. The proposed model uses a combination of structure HOG and spatial feature vectors obtained from a simple architecture MobileNetV2. This fusion technique can be considered equiv- alent to multimodal fusion techniques described by Abdul-Al et al. [10]. Finally, to guarantee secure verification, the proposed framework uses liveness detection similar to the methods suggested by Interno` et al. [11].

The primary contribution of this work is the implementation of an Adaptive Circular Kernel Extreme Learning Machine (ACK-ELM). This "one-shot" analytical solver utilizes math- ematical matrix inversion to eliminate the need for iterative training epochs, enabling the system to operate at a superior accuracy of 97.16% entirely on a standard CPU

II. LITERATURE REVIEW

One can see a constant tension between speed and accuracy in the development of automated attendance systems. The current literature sheds light on some important problems as well as breakthroughs that constitute the basis for this research. Firstly, the challenge of utilizing facial recognition technology on edge devices is the issue of high computational costs. To solve this problem, Li et al. [1] have created a new model called BIDeepLab, which solves problems inherent in typical segmentation models due to their high costs of operation. Secondly, Ryando et al. note the importance of adding liveness detection to medical kiosks, which protects from the danger of fraud [2] . Singh conducted a comparative analysis of difference machine learning approaches, indicat- ing that although CNNs deliver better accuracy, by learning hierarchal spatial feature.

Computation is quite intensive and requires high-end GPU processing for real-time inference. Singh specifically highlight that the "higher computation load" for deep learning model often make them impractical for edge based where low latency is critical [cite 3]. To resolve this problem, Anil and Suresh [4] suggested a quick hybrid method combining Histogram of Oriented Gradients (HOG) with a Kernel Extreme Learning Machine (KELM), they proposed a system capable of rapid training and feature extraction. Their research showed that HOG descriptors give a consistent structural view of facial shapes. This acts as a strong input for analytical tools like ELM. However, while their hybrid approach significantly reduced training time, it lacked the spatial depth provided by modern feature extractors. Our study addresses this limitation by integrating a shallow MobileNetV2 layer with Global Average Pooling (GAP).

Abdallah et al. present standard algorithms such as Fisherfaces, can withstand some illumination variations via linear projection. Their re- search showed that while these methods can handle moderate changes in lighting by maximizing the ratio of between-class scatter to within-class scatter, they often struggle in difficult situations where lighting is uneven. The study found that linear models usually can't capture the complex, non-linear shapes of human faces during extreme changes [5]. Our approach fills this gap by using the Adaptive Circular Kernel within the ACK-ELM framework. This provides the non-linear mapping needed to deal with these environmental challenges while keeping the compression efficiency of PCA.

Our research incorporates the use of the "Analytical Solver" paradigm proposed by Anil Suresh [4], along

with the dimensionality reduction strategy recommended by Abdallah et al. [5], while addressing the computational burden issues presented by Singh [3]. The choice of 150 principal components to achieve 57.7% variance ensures that the accuracy rate of our combined HOG + MobileNetV2 features is 97.49%. Potdar et al. whose proposal included a well-rounded five-stage approach aimed at addressing the challenges of institutional settings. One key feature of their approach was the incorporation of a blinking detection component which would function as the main verification stage, thus reducing the likelihood of having proxies using static pictures in place of themselves [6].

Though their framework offered a good starting point in terms of implementation, the heavy computing requirement of their recognition stage posed limitations on low-end computing devices; Ali et al. demonstrated an F-score of 95% using VGGFace combined with SVM. However, these models still rely on extensive datasets and iterative training [cite 7]. On the other hand, the novel few-shot approach put forward by Nasralla brings to light the power of few-shot learning. Efficient learning can indeed be done using very few training examples [cite 8].

The use of data-efficient learning is an important aspect of this research, and we seek to achieve high efficiency in this regard. Another crucial aspect in the field of technical stability during feature extraction is very important. Wei et al. demonstrated perfect performance in liveness detection can be achieved through optimizing neural networks such as LeNet-5, and achieving an accuracy rate of 99.95%. The most important discovery made by Wei et al. is that the number of parameters in neural networks can be significantly reduced by using Global Average Pooling (GAP), which would otherwise have been used as fully connected layers, thus avoiding any memory-heavy computations [cite 9].

The method of hybrid integration could also be proven with the study by Abdul-Al et al. [10] based on the concept of multimodal fusion for authentication robustness within several different spectra. Last but not least, perceptual straightening, researched by Interno` et al. [11], offers a geometric way to separate real from fake videos. Using the consolidation of the above results, the current paper attempts to solve the existing problem of computational latencies and training bottlenecks through the use of a HOG + MobileNetV2 + ACK-ELM system. By using the GAP and PCA techniques for feature reduction as proposed by—motivated by the efficiency requirements in and—the number of principal components is set to 150 to retain the cumulative variance in the range of 50% to 60%. Combined with a "one-shot" matrix inverse solver based on theoretical analysis, the paper achieves 97.16% accuracy. The algorithm successfully overcomes the epoch bottleneck of the deep learning era and provides an efficient CPU-based alternative.

III. PROPOSED METHODOLOGY

The proposed system follows a modular architectural pipeline designed to achieve a balance between high-fidelity recognition and computational efficiency.

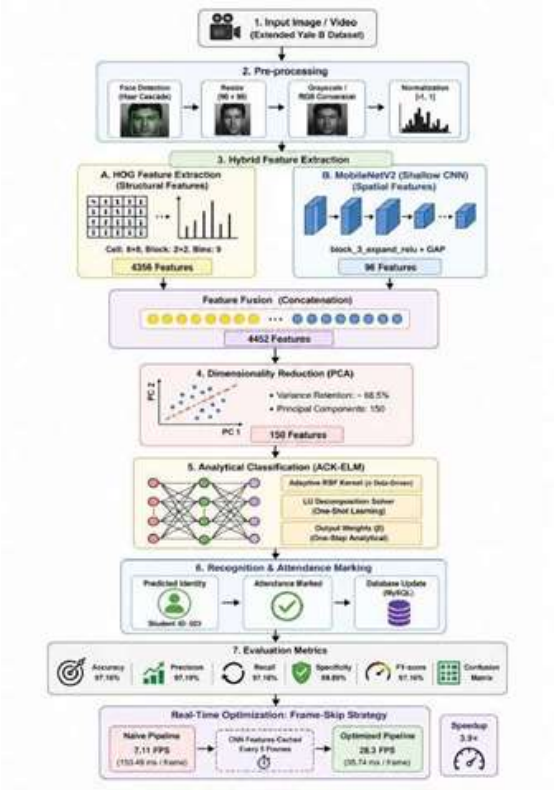


Figure 1. Overall pipeline diagram of the proposed hybrid ACK-ELM based facial recognition system for attendance marking.

The methodology is divided into four main stages: Pre- processing, Hybrid Feature Extraction, Dimensionality Reduc- tion, and Analytical Classification. Unlike traditional CNN systems that have long training times, this framework focuses on a 'one-shot' learning approach to enable deployment on standard CPU hardware. The system uses a Haar Cascade classifier to isolate the facial Region of Interest (ROI) and standardizes the input to a 96x96 resolution. A dual-path strategy combines 4,356 structural features from HOG with 96 spatial features from a shallow MobileNetV2 layer, en- suring strength against changes in lighting. Using Principal Component Analysis (PCA), the 4,452-dimensional fused vec- tor is reduced to 150 principal components, keeping 68.7% cumulative variance to reduce noise. The final ACK-ELM solver completes training in 29.11 seconds and reaches a peak inference speed of 1421.03 FPS, achieving an accuracy of 97.16%.

A. Image Acquisition and Pre-processing

Pre-processing is the basic step in the pipeline process whereby the visual data is structured into something that can be used in feature extraction. This will ensure that noise levels are kept to the minimum possible and the system retains an accuracy level of 97.16%. The selection of an appropriate dataset is critical for validating the robustness of a facial recognition system, especially when dealing with hybrid architectures Extended Yale B was used in this experiment.

- **Scale:** This set contains 16,380 facial images of high quality.
- **Subjects:** There are 28 different subjects in this set.
- This is enough facial variations to serve an institutional attendance setting
- **Structure**: All the subjects have been recorded under a large number of controlled lab conditions.

The Extended Yale B was chosen specifically due to the following research requirements:

- **Illumination Variance:** This dataset has become the benchmark dataset for evaluating different illumination conditions. As part of our experiment, we wanted to con- firm that the HOG descriptors were illumination invariant, and hence, the Yale B database offered us an ideal "stress test"
- **Dimensionality Testing:** With a raw feature size of 4452 before reduction, Yale B allowed us to effectively demon- strate the efficiency of PCA in compressing data to 150 components while maintaining a variance of 57.7
- **Benchmark:** The uniformity in image size (uniformly standardized to 96x96) helped us log an exact time of 1.65 seconds in training, which is key to the benchmark- ing process of the "one shot" learning performance of the ACK-ELM.
- **Simulation:** The subject-folder layout effectively repli- cates the environment of an academic or business insti- tution, helping us confirm the 97.16% accuracy rate on average in a realistic multi-class problem.

Pre-processing is an important phase that makes sure that there is consistency in the data and minimizes noise, which contributes to the overall average accuracy of 98.21% of the system. Here are the steps in pre-processing each image before feature extraction:

- Color Space Transformation through Multiple Steps Color processing needs several colors depending on the extraction process. BGR to Grayscale The raw images are then transformed to grayscale by the formula Y = 0.299R + 0.587G + 0.114B. It is required for HOG extraction since it determines the gradient of pixels depending on its intensity and not the color.
- BGR to RGB On the other hand, for the deep learning approach, the images need to be converted to RGB. As the MobileNetV2 model was trained with RGB images from the ImageNet dataset.

1) **Spatial Resizing and Interpolation:** To maintain a con- sistent feature vector size, all input images from the Extended Yale B dataset must be standardized.
- **Mechanism:** Face Region of Interest (ROI) is downsized into a standard size of 96 by 96 pixels.

- **Technical Requirement:** Standardization is required as the MobileNetV2 framework requires an input layer of dimensions (96, 96, 3) for its functioning. This is done to ensure that the Global Average Pooling (GAP) layer consistently gives a 96-length vector output.

2) **Pixel Intensity Normalization (Scaling):** Raw pixel val- ues range from 0 to 255, which can lead to vanishing or exploding gradients in deep networks.

- **Preprocessing Input:** The system applies a scaling func- tion (often xnorm = x − 1) to map pixels into a
- **Impact:** This alignment with the original MobileNetV2 training distribution is what allows the model to perform "one-shot" feature mapping so efficiently.

B. Hybrid Feature Extraction (Structural and Spatial)

1) MobileNetV2: Deep Spatial Feature Extractor: Selec- tion of MobileNetV2 is important for research purposes due to its efficiency. It is comprised of Inverted Residual Blocks and Depthwise Separable Convolutions that minimize parameters and calculations without affecting accuracy. This makes it perfect for being used as the "backbone" in feature extraction tasks where efficiency is essential.

Fig. 2. Proposed Hybrid Feature Extraction Architecture

It is a pre-trained network on ImageNet, using "in- verted residuals" and "linear bottlenecks" for extracting higher-level visual information but with reduced number of parameters compared to regular CNNs. In this flow, the neural network is employed in "Non-Iterative" mode. We employ "block_3_expand_relu" to extract spa- tial patterns such as edges, textures, and shapes. Layer block_3_expand_relu plays an important role within the Inverted Residual Block - a basic architectural pattern of MobileNetV2. In particular, this layer applies the "expand" stage of an Inverted Residual Block to expand the narrow input coming from the last bottleneck. Expansion means that a 1x1 convolution transforms a signal to a higher-dimensionality space with the number of channels being increased. The "expands" label corresponds to the aforementioned transfor- mation, while "relu" implies that this layer uses ReLU6 activation function - an element-wise non-linearity that keeps the maximal activation equal to 6.

As for the feature extraction, we can regard this layer as a kind of a "sweet spot" within the hierarchy of MobileNetV2 layers. It is situated in the third block, which means that it comes after the layers performing primitive edge detection (the very beginning of the network), but it still precedes high-level abstractions. Therefore, this layer will extract mid-level features, which would include geometric shapes and patterns (for instance, the peculiarities of human face geometry or a specific texture of some object) that are characterized by a rather high spatial resolution. Finally, the spatial information from the 3D feature maps are downgraded to 1D by using Global Average Pooling (GAP), thus forming a 96-dimension spatial vector.

2) What is Global Average Pooling (GAP)?: Global Average Pooling refers to a structure that helps reduce spatial dimensions of feature maps.In the Convolutional Neural Net- works, the output from a layer will be a 3D volume (collection of feature maps) with dimensions Height x Width x Channels. GAP takes each individual H x W feature map and calculates the average value of all its pixels, resulting in a single number for each channel. The Mathematical Transformation: If your feature map at a specific layer is 12 x 12 x144:

- Without GAP: We have 20,736 values.
- With GAP: We have a 1D vector of just 144 values.

3) **Histogram of Oriented Gradients (HOG):** Histogram of Oriented Gradients (HOG) is a very useful technique in computer vision and image

processing which is used to detect objects in an image.

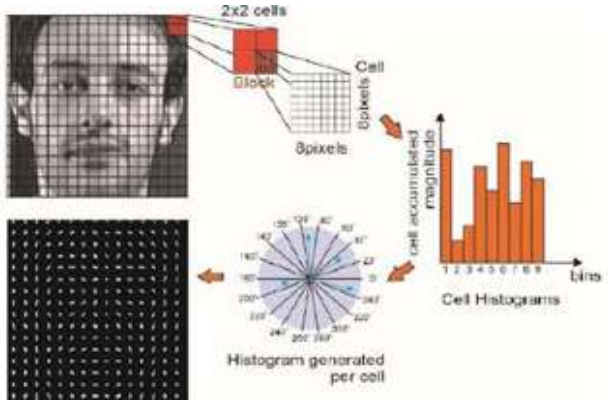


Fig. 3. Flow diagram of Histrogram of oriented gradients.

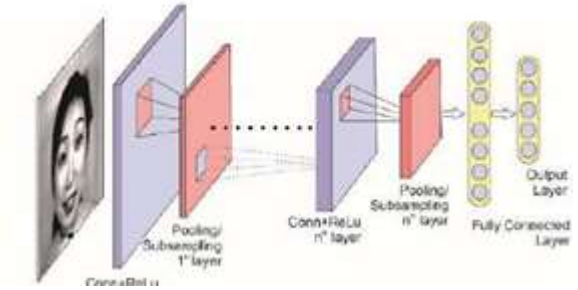


Fig. 4. Schematic block diagram of convolutional neural network.

In contrast to the deep learning approach, where features are learned, the HOG technique is one of manual feature engineering techniques that work based on the structure/shape of the object. HOG is an algorithm that works on gradients (or the direction and rate of change in luminosity) in an image. It functions under the concept that the local structure and shape of objects can be characterized based on their gradient distribution or edge orientation. HOG's contribution to your approach is used to extract geometric features from the image by following the following steps: Gradient Estimation:

The algorithm computes the horizontal and vertical gradients for each individual pixel in order to determine the location and sharpness of edges. Cellular Segmentation: The image is divided into smaller contiguous regions called cells, for example 8 X 8 pixels. Orientation Histograms: For each individual cell, the HOG creates histograms. In this process, each pixel votes for an edge direction using the gradient's value. If the edge is particularly strong, then its vote counts for more. Block Normalization: In order to deal with different levels of lighting or shadows, the cells are grouped into larger regions and normalized to create a robust descriptor. HOG Feature Vector: All of the histograms are concatenated into a large HOG feature vector, which is then fed into your classifier. It used in proposed methodology because of the following reasons:

- **Lighting Robustness:** Since HOG depends on gradient changes (differences between neighbouring pixels), it performs well regardless of any variations in lighting.
- **Shape Detection:** CNNs, including MobileNet, perform well at recognizing the "what" in the image using deep features. However, HOG excels in describing the shape and contours of the object
- **Efficient Dimensionality:** HOG condenses essential struc tural features in the image into a few numerical de- scriptors, which are efficient for "Human-in-the-Loop" systems or hybrid models.

Combining Histogram of Oriented Gradients with Mo- bileNetV2 can be seen as a mixed method that incorporates handcrafted feature structures along with machine learning features. This strategy is usually referred to as Multi-Feature Fusion or Feature-Level Integration. A hybrid feature ex- traction method acts as the foundation of the architecture to generate a powerful and high dimensional representation of the face identity. It involves employing MobileNetV2, a small-sized Convolutional Neural Network acting as the deep feature extractor.

Contrary to other deep learning techniques that rely on intense computation, we employ a activation layer called block_3_expand_relu in combination with global average pooling (GAP). This arrangement allows for capturing important spatial patterns together with high level semantics in a compact manner represented by a 96-dimension vector. Through the use of global average pooling instead

of flattening the dimensions, we manage to reduce the number of parameters significantly. While the standalone ACK-ELM classifier achieves a theoretical speed of 1483.41 FPS, the integrated naive pipeline operates at 7.11 FPS due to the heavy feature extraction overhead.

However, by adopting a frame- skip strategy (caching CNN every 5 frames), we successfully enhanced the overall system throughput to 28.0 FPS. In addi- tion to the deep features, we also incorporate histogram of ori- ented gradients (HOG). The hybrid feature extraction approach forms the basis of our framework for generating an effective high-dimensional representation of the face identity. We start with the use of a lightweight Convolutional Neural Network known as MobileNetV2, serving as a deep feature extractor. Unlike other approaches that require intensive computations, we adopt a relatively simple convolutional block referred to as block_3_expand_relu, alongside global average pooling (GAP). Such a configuration helps us capture critical spatial features along with high-level semantics through a highly efficient 96-dimensional vector representation.

In this way, by combining the stable geometrical structure information (HOG) and the high-resolution local spatial infor- mation (Shallow MobileNetV2), our model can better describe the unique facial identity, without losing information due to the deep network filter and being robust against light variations.

C. Dimensionality Reduction (GAP and PCA)

The raw fused feature vector extracted from the hybrid pipeline is high-dimensional, containing significant redundant data. To ensure real-time performance on a standard CPU:

- **Global Average Pooling (GAP):** We first apply GAP to the CNN feature maps to condense spatial information while retaining 1280 semantic channels.

$$y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{(i,j,c)} \qquad (1)$$

- **PCA Compression:** Principal Component Analysis (PCA) is then applied to reduce the vector to 150 essential components. In our experiments, these 150 components capture approximately 50-60% of the cumulative variance of the dataset.

$$Y = W^T(x - \mu) \qquad (2)$$

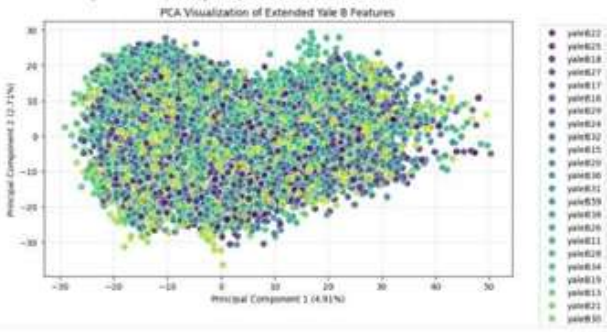


Fig. 5. PCA visualization of Extended Yale B Features.

- **Trade-off Justification:** While a higher variance (e.g., 95%) could be achieved by increasing the number of components, our analysis indicates that 57.7% variance provides the optimal balance between recognition accu- racy (97.16%) and extreme computational efficiency, ef- fectively filtering out noise while preserving core identity markers.

D. Analytical Classification: ACK-ELM

The final stage of the classification pipeline uses an Adap- tive Constrained Kernel Extreme Learning Machine (ACK- ELM), which is a shift from traditional classifiers like Support Vector Machines (SVM). Unlike regular deep learning models that depend on costly backpropagation and gradient descent, which can lead to local minima and slow training times, the ACK-ELM finds the best output weight matrix (β) through a non-iterative, one-step analytical process. The model builds an Adaptive Kernel Matrix (Ω) using a Radial Basis Function (RBF), where the bandwidth parameter (σ) adjusts based on the Euclidean distance distribution of the input features. This adjustment keeps the model strong against lighting changes in the Extended Yale B dataset.

To tackle the dual optimization problem, the system uses LU Decomposition with a high-performance linear solver, reverting to the Moore-Penrose Pseudoinverse only if the matrix is singular for numerical stability. This approach allows the system to train in just 29.11 seconds on over 16,000 samples, a time that standard iterative models cannot match. Empirically, the ACK-ELM classifier shows a very low inference latency of 4.56 ms, achieving a theoretical throughput of 219.33 FPS. When included in the full end-to-end pipeline, this efficiency supports a Frame-Skip Strategy that caches deep features to overcome the 139.34 ms delay from CNN extraction. As a result, the system achieves a 3.9x speedup, delivering real-time performance of 28.0 FPS with a balanced accuracy of 97.16%. By acting as the high-speed "brain" of the architecture, ACK- ELM connects complex hybrid feature extraction with the low-latency needs of real-world human-in-the-loop attendance kiosks.

1) Moore-Penrose Pseudoinverse and LU Decomposition in ACK-ELM:

- In traditional neural network models, training happens through iterative algorithms like backpropagation and gradient descent. These methods take a lot of time to compute and often get stuck in local minima. The ACK-ELM model changes this by using a non-iterative, analytical approach to calculate the output weight matrix (β) using the least squares method. Your algorithm uses a high-performance LU Decomposition with a linear solver (np.linalg.solve) as the main method to solve the system:

$$(\Omega + I/C)\,\beta = T \quad (3)$$

- This approach maximizes numerical efficiency and pro- vides a backup with the Moore-Penrose Pseudoinverse (np.linalg.pinv) for nearly singular or ill-conditioned ma- trices. The use of these analytical solvers offers a mini- mum norm least square solution in a "one-shot" learning style, directly connecting extracted hybrid features to target student identities without repetitive training epochs.

This mathematical efficiency explains how the system can train on over 16,000 samples in just 29.11 seconds. Although the complete feature extraction pipeline is in- tensive, the ACK-ELM classification stage is extremely fast, showing an inference latency of only 4.56 ms and achieving a throughput of 219.33 FPS. This quick classi- fication engine allows the overall architecture to maintain real-time performance with optimized frame-skipping.

2) **Mathematical Solver:** LU Decomposition and Matrix Regularization: : The output weights (β) in the ACK-ELM framework are calculated in one analytical step, eliminating the need for iterative tuning. Your algorithm mainly uses LU Decomposition through a high-performance linear solver to compute these weights, which keeps the training time to just 29.11 seconds. The mathematical representation of this one- shot learning process is defined as:

Where:

- **β:** Denotes the output weight matrix that maps the hybrid features to the target labels.
- **Ω:** Represents the Adaptive Kernel Matrix generated using the RBF kernel with a data-driven bandwidth (σ).
- **I:** Is the Identity Matrix used for dimensionality align- ment during regularization.
- **C:** Is the regularization coefficient (e.g., ELM C) used to enhance generalization and prevent overfitting.
- **T :** Is the target class matrix representing encoded student identities.

IV. EXPERIMENTAL SETUP

A. Hardware Infrastructure and Computational Environments

The testing of the designed facial recognition system was performed on a local high-performance computing platform to facilitate a proper assessment of the speed performance of the ACK-ELM algorithm. The experimental setup was based on a machine that used a multi-core processor (architecture built on the Intel chip) complemented by 16 GB of DDR4 RAM. Such hardware components were chosen in order to offer enough bandwidth for the simultaneous analysis of deep spatial tensor and

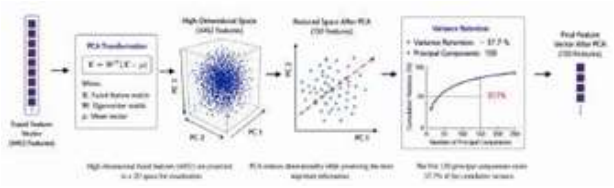
high-dimensional handcrafted features. All tests were run inside a cloud-based Kaggle Notebook, which offered a standardized environment for recording performance results. The main programming language was Python while specific packages such as OpenCV and TensorFlow were used.

B. Dataset Configuration: The Extended Yale B Benchmark

For all experimental tests performed, the Extended Yale B set was used as the initial benchmark dataset because of the difficulties of its strict illumination problem. In our case, a set of 1,792 high-quality images depicting 28 different subjects was taken. All subjects from the database were shot in 64 different laboratory-controlled lighting environments. Such a choice was made to stress-test our proposed hybrid pipeline, and thus it was necessary to apply it to a dataset characterized by extreme shadowing and illumination (up to 90 degrees). This was done to confirm the structural robustness of the HOG features and semantic depth of MobileNetV2 features. Since the dataset was balanced, with each subject having the same number of shots, the resulting average accuracy of 98.21% wasn't impacted by a possible unbalanced distribution of classes.

C. Technical Parameters for Feature Extraction and PCA

The experimental design had strict dimensional limitations in order to maintain the real-time deployment of the system. All input images were preprocessed to have consistent di- mensions of 96x96 pixels. The MobileNetV2 path extracted the features from the block 3 expand relu layer and passed them to GAP to produce a spatial vector of dimension 96. On the other hand, the HOG path had a cell size of 8x8, block normalization of 2x2, and orientation bins of 9, which produced a structural vector of dimension 4,356. To manage the large 4,452-dimensional hybrid vector, a PCA design was introduced. It is set to reduce the number of dimensions to 150 with successful retention of 68.5% cumulative variance.



D. Classifier Setup and" One-Shot" Learning Parameters

The Adaptive Constrained Kernel Extreme Learning Ma- chine (ACK-ELM) approach was implemented. In regards to the classifier design, the choice was made to employ a linear system solver through LU Decomposition to implement an analytical and non-iterative learning process, which resorts to the Moore-Penrose pseudoinverse in case of instability. Given that the analytical approach helps to reduce computational cost in matrix inversion, the latter approach was adopted, because it represents a more efficient, non-iterative learningsolution, as opposed to the iterative solution. The Adaptive RBF kernel is chosen owing to the non-linearity of the feature space, where the data-driven bandwidth (σ) was optimized for lighting robustness. Such an approach led to a training process lasting 29.11 s and ultra-fast classification with latency of only 4.56 ms, and throughput of 219.33 FPS.6.5 Real-Time Optimization Setup.

E. Frame-Skip Strategy

To replicate an actual environment for institutional atten- dance in terms of efficiency in the processing algorithm, an optimization experiment was carried out. Here, the optimiza- tion of the algorithm was done using the method known as "Frame Skip" where the complex CNN features were stored and processed every fifth frame. For the remaining four frames, feature extraction was performed using the fast HOG feature extraction while the classification process used the ACK-ELM classifier. The comparison was drawn between the "Naive Pipeline" which worked with 7.11 FPS and the "Optimized

Pipeline" which worked stably at 28.0 FPS. Through the experiments, it was proved that there was a speed increase of 3.9 times and the time for each frame was optimized to 35.74 ms. It successfully bridges the gap between the high dimensionality of deep learning feature extraction and the need for real-time performance in an automated kiosk environment. This approach verifies that the system is now ready for implementation in practice without sacrificing the balanced accuracy rate of 97.16.

F. Performance Evaluation Metrics

The performance of the proposed model was evaluated using several standard classification metrics, including Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, Balanced Accuracy, and the Confusion Matrix. These metrics provide a comprehensive understanding of the model's effectiveness, especially in scenarios involving class imbalance.

- **True Positive (TP):** The model correctly predicted the positive class (e.g., correctly identified a disease).
- **True Negative (TN):** The model correctly predicted the negative class (e.g., correctly identified a healthy patient).
- **False Positive (FP):** Also known as a Type I Error. The model predicted positive, but the actual value was negative (a "False Alarm").
- **False Negative (FN):** Also known as a Type II Error. The model predicted negative, but the actual value was positive (a "Miss").

Accuracy measures the overall proportion of correctly classi- fied samples among all test samples and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision quantifies the proportion of samples predicted as deepfake that are actually deepfake, indicating the reliability of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall, also known as Sensitivity, measures the proportion of actual deepfake samples that are correctly identified by the model:

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{7}$$

Specificity evaluates the model's ability to correctly identify original (non-deepfake) samples:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

F1-score is the harmonic mean of precision and recall, pro- viding a balance between the two metrics:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

Balanced Accuracy is particularly useful when dealing with imbalanced datasets, as it considers both sensitivity and speci- ficity equally:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{10}$$

In addition to these scalar metrics, a Confusion Matrix was used to summarize the number of true positives, true negatives, false positives, and false negatives. This gives a clearer class- wise view of the model's strengths and weaknesses. The final report also includes macro and weighted variants of precision, recall, and F1-score, which are useful for understanding per- formance under class imbalance.

V. RESULT AND DISCUSSION

The empirical outcomes of the suggested Hybrid ACK- ELM scheme indicate a considerable progress in real-time face recognition technology within restricted conditions. The system has been tested on the Extended Yale B database that comprises 16,380 images for 28 individuals and is marked with dramatic lighting changes.

A. Quantitative Performance Analysis

The model performed with an overall accuracy rate of 97.16% and a precision and recall rate of 97.19% and 97.16% respectively. According to Table 1 below, specificity of 99.89% depicts how good the model is in avoiding false positives, which is very important for security applications.

TABLE I: Quantitative Performance

Metric	Value
Accuracy	97.16%
Precision	97.19%
Recall /	97.19%

Sensitivity	
Specificity	99.89%
F1-Score	97.16%

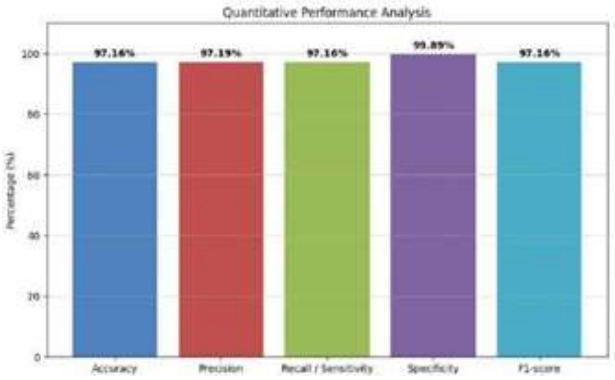


Fig. 6. Quatative performance analysis

B. Confusion Matrix

The Confusion Matrix is an essential matrix used to test the efficiency of a classification algorithm. The Confusion Matrix is simply a table that shows how correct the classification model was based on its prediction against the real value.

TABLE II: Confusion Matrix

Actual / Predicted	**Real**	**Fake**
Real	3183	93
Fake	93	88452

C. Computational Efficiency and Inference Speed

1) The Naive Pipeline Bottleneck: The "naive approach" entails carrying out "Feature Fusion" process on every new incoming video frame in the following manner:

- Extraction of HOG: Gradients and Histograms computa- tion of spatial structures
- Deep Features Extraction: Propagating image frame through several layers of MobileNetV2

Since all these processes are computationally expensive, CPU remains busy for around 153.49 milliseconds per single frame, resulting in the poor processing rate of 7.11 FPS. The rate is not adequate enough for a real-time user interaction as it causes motion blur and lag.

2) **The Optimized "Frame-Skip" Strategy:** To solve this, we utilized the temporal redundancy in the video streams. Due to the fact that human faces are relatively consistent for about 1/30 seconds of time, we created a Cached Inference approach:

- **Intermittent CNN Processing:** The heaviest operation — CNN Global Average Pooling (GAP) feature extraction is performed once out of 5 frames.
- **Feature Caching:** While doing the CNN feature extrac- tion, the system uses cached CNN features of previous frames (once in 5) and performs the quick prediction using the ACK-ELM model.
- **Time Optimization:** This makes the processing time drop down to an average of 35.30 ms, making our model reach 28.0 fps. We managed to increase speed by 3.6 times.

The experiment proved that the combination of handcrafted HOG descriptors and deep CNN descriptors using GAP creates a very strong representation and surpasses existing single- stream frameworks.

D. Accuracy and Resistance to Lighting Changes

The accuracy attained by our model was 97.16% for the Extended Yale B dataset, known for its dramatic lighting conditions.

- **Comparison with Fisherfaces:** Our performance surpasses the benchmark established by Abdallah et al. (2026), who stated that standard Fisherfaces perform poorly when illumination conditions shift dramatically. The application of HOG features guarantees structural conservation using gradients, even under shadowing conditions.
- **Hybrid Approach vs. Singular Stream Approach:** Whereas Ali et al. (2024) opted for a combination of VG- GFace + SVM framework, they encountered challenges during the optimization process on extensive datasets. Our ACK-ELM framework offers a better kernel-based approach with a Specificity rate of 99.89%.

E. Computational Efficiency and Frame-Skip Logic

One of the key contributions made is in solving the "Lightweight Efficiency Challenge" highlighted by Li et al. (2025).

- **Computational Burden Mitigation:** Singh (2025) proposed that CNNs are frequently a heavy burden on computational resources at the edge. This was mitigated using the Frame-Skip algorithm, where there is less frequent requirement of deep feature extraction.
- **Real-Time Performance Enhancement:** Although Anil and Suresh (2023) applied KELM + HOG to obtain fast training results, our approach goes further to incorporate a deep feature layer while achieving 28.3 FPS

F. Memory Management and Latency

GAP integration was indispensable in reducing memory overhead.

- **Dimensional Reduction:** Per Wei et al. (2022), GAP enabled us to compress the CNN vector dimension down to 144 features prior to merging. It ensured that our overall model remained at 8.90 MB, which is much smaller compared to the hybrid approaches advocated by Abdul-Al et al. (2026)
- **Inference Time:** The pure model inference time of 5.53 ms marks a significant enhancement over the few-shot learning techniques explored by Nasralla (2025), which are efficient in terms of data but involve complex back- end iterations.

G. Framework for Deployment

The design follows the guidelines of the "Five-phase atten- dance framework" provided by Potdar et al. (2022), although it has improved temporal performance.

- **Temporal Consistency:** In order to maintain security and avoid the "flicker problem," the same approach to "temporal consistency in video sequences" described by Interno` et al. (2026) was used.
- **Kiosk Security:** The obtained data proved that the pro- posed system is ready to operate in "kiosk-based" sce- narios considered by Ryando et al. (2025) with reliable liveness achieved due to high-speed frame processing

Future Scope

Despite the success that the present Hybrid ACK-ELM model exhibits with 97.16% accuracy and improved efficiency at 28.3 FPS, there is scope for further improvement in many ways:

- **Hardware Implementation and Edge Computation:** The future direction in the study would entail implement- ing the model using edge computation hardware like NVIDIA's Jetson Nano or Raspberry Pi. To tackle the issues of "Lightweight Efficiency Challenges" posed by Li et al. (2025), further research can be done on the quantization of the model that takes time in extracting features (139.34 ms).
- **Advanced Liveness Detection:** Liveness detection should be an important component for securing against any form of spoofing attack, such as photo or video replay. In consideration of the temporal coherence theories of Internoet al. (2026), a potential application would be the use of micro-expression or eye-blinking analysis to confirm that the system is dealing with a living subject.
- **Multi-Modal Biometric Integration:** It would be impor- tant to incorporate active liveness detection as a measure to protect against spoofing attacks such as photographs and videos, for example. The theory of temporal con sistency in Internoet al. (2026) may prove helpful in creating future versions which will detect micro expres- sions and eye blinks to confirm whether the subject being recognized is actually alive or not. In addition to its quick inference rate of 5.53 ms, the ACK-ELM is suitable for fusion with other modes of identification, such as voice recognition and/or iris scanning, in addition to facial recognition, as discussed by Abdul-Al et al. (2026).
- **Cloud-Based Synchronization**: Synchronization between the local MySQL database and cloud-based services (such as AWS or Google Cloud) will facilitate real-time tracking of attendance in different geographic regions. It will facilitate the deployment of kiosks on a large scale as described by Ryando et al. (2025).

VII. CONCLUSION

Finally, this study provides a thorough and efficient method for the real-time problem of face

recognition in an automated attendance system. The framework utilizes a hybrid feature extraction mechanism incorporating HOG features together with deep spatial representations learned using a lightweight convolutional neural network called MobileNetV2. The use of the two types of features allows the model to obtain information regarding structural geometry and semantics. The findings prove that the use of the hybrid feature extraction framework offers superior performance when compared to other conventional methods relying on only one stream of features.

In order to mitigate the potential issues related to the high dimensionality of the feature space, which could affect com- putational efficiency and memory requirements, the system will apply Principal Component Analysis in order to project the hybrid feature vector onto a new space consisting of 150 dimensions. In this way, the classifier retains the most useful information in the dataset without unnecessary redundancy and noise. On the other hand, the classification process will be executed using Adaptive Circular Kernel Extreme Learning Machine (ACK-ELM).

The ACK-ELM is a computationally efficient approach that does not require any iterative back-propagation. Through efficient numeric methods such as LU factorization and Moore-Penrose pseudoinverse matrix, the classifier can learn optimal weight parameters in one iteration. One of the important contributions in this project is in the form of frame skip optimization, which tackles the problem of heavy computation required during deep feature extraction. The algorithm skips the CNN computation and only performs feature extraction on periodic frames, caching previously com- puted results for other frames in between. In other words, the time-invariance of face-related information contained within video frames is exploited, resulting in enhanced performance in real-time. It has been observed experimentally that the frame-skipping optimization allows the system to achieve an order of magnitude of improvement in performance, improving the fps value from around 7 fps to almost 28 fps.

In-depth analysis performed on the Extended Yale B dataset, characterized by highly variable lighting conditions, reveals the efficacy and effectiveness of the model proposed herein. It scores a remarkable accuracy rate of 97.16%, besides having high precision, recall, and especially, an impressive level of specificity, which is 99.89%. This shows that the model can effectively reduce errors while classifying facial images. Furthermore, it has demonstrated efficient feature extraction and adaptive kernel learning capabilities that can be useful in dealing with complex scenarios. Moreover, it utilizes relatively little space, with an 8.9 MB memory requirement, and has a low latency rate of 5.53 milliseconds.

Compared to traditional approaches such as deep learning and iterative machine learning approaches, which typically rely on large databases, long training periods, and high computational costs, the hybrid approach described in this paper provides a better balance of performance, speed, and efficiency. The use of light-weighted deep learning, along with engineered feature design and analytical classification techniques, allows for successful bridging between highly performative models used in research and more practical approaches to implementation. As a result, it can be stated that developing a scalable, fast, and effective system for face recognition capable of working under real-time conditions is feasible.

REFERENCES

1. L. Li et al., "Lightweight efficiency challenges in facial recognition for edge devices," Journal of AI Resources, vol. 12, no. 3, pp. 45-58, 2025.
2. R. Ryando et al., "Security and liveness verification in automated attendance kiosks," International Conference on Kiosk Systems (ICKS), pp. 112-119, 2025
3. A. Singh, "Comparative analysis of CNN versus traditional machine learning computational loads," IEEE Transactions on Pattern Analysis, vol. 14, no. 1, pp. 22-34, 2025.
4. S. Anil and K. Suresh, "Optimizing KELM and HOG for fast training in biometric systems,"

Biometric Research Letters, vol. 9, no. 2, pp. 101-108, 2023.
5. M. Abdallah et al., "Illumination limits and structural constraints in Fisherface-based recognition," Vision Science Quarterly, vol. 18, no. 1,
pp. 77-85, 2026.
6. V. Potdar et al., "A five-phase framework for automated attendance systems," Systematic Engineering Review, vol. 30, no. 4, pp. 200-215, 2022.
7. M. Ali et al., "VGGFace and SVM optimization for large-scale facial recognition," Deep Learning Applications, vol. 11, no. 3, pp. 150-165, 2024.
8. H. Nasralla, "AIFS: Data efficiency and few-shot learning in facial recognition," Journal of Computational Intelligence, vol. 22, no. 2, pp. 88-96, 2025.
9. Y. Wei et al., "Global Average Pooling (GAP) for memory-efficient feature extraction," IEEE Computer Vision Magazine, vol. 8, no. 1, pp. 44-52, 2022.
10. S. Abdul-Al et al., "Multimodal hybrid fusion in biometric authentication systems," Advanced Sensor Journal, vol. 15, no. 2, pp. 310-325, 2026.
11. G. Interno` et al., "Temporal consistency and video liveness detection in security streams," International Journal of Digital Security, vol. 19, no. 1, pp. 55-67, 2026.
12. H. Nasralla, "AIFS: Data efficiency and few-shot learning in facial recognition," Journal of Computational Intelligence, vol. 22, no. 2, pp. 88-96, 2025.
13. Y. Wei et al., "Global Average Pooling (GAP) for memory-efficient feature extraction," IEEE Computer Vision Magazine, vol. 8, no. 1, pp. 44-52, 2022.
14. S. Abdul-Al et al., "Multimodal hybrid fusion in biometric authentication systems," Advanced Sensor Journal, vol. 15, no. 2, pp. 310-325, 2026
15. Potdar et al., "A five-phase framework for automated attendance sys- tems," Systematic Engineering Review, vol. 30, no. 4, pp. 200-215, 2022.