

Intel Unnati Industrial Training Program-2025

Problem Statement-07

**Visual Search using VLM's
(Enhancing Forensic Investigations: A Multimodal Search
Engine with CLIP and FAISS)**

TEAM_7_METIS Visionaries

Team Members

**JUDE FRANKLIN M
CHANDAN N
H JOHN JOSHUA**

VideoLink:-https://drive.google.com/file/d/1-2LyCa_B0XFuASE60v784zsb7qqojdwK/view?usp=drive_link

APRIL-2025

Abstract

This project outlines the development of a **Forensic visual search engine** tailored for face to face retrieval, utilizing **Vision-Language Models (VLMs)** to effectively connect textual queries and image inputs with real images from the **IDOC Mugshot Dataset**. The system employs **Contrastive Language-Image Pre-training (CLIP)** to enable precise retrieval by embedding sketches, text descriptions, and real photographs into a shared semantic space, allowing for seamless multi-modal querying. Whether users input detailed textual descriptions or upload the criminal images, the system retrieves matching images efficiently using **Facebook AI Similarity Search (FAISS)** indexing and **Approximate Nearest Neighbor (ANN)** search techniques.

The architecture is built upon a foundation of deep learning, computer vision, and contrastive learning techniques, ensuring a robust and scalable solution for forensic applications. It features a user-friendly frontend designed for intuitive interaction, a Python-based backend for reliable query processing, and a PyTorch-driven machine learning pipeline that powers embedding generation and system efficiency. Performance is rigorously assessed through quantitative metrics, including precision@k, recall@k, and mean average precision, which evaluate retrieval accuracy and ranking quality, while qualitative validation from forensic experts ensures the system meets practical investigative needs. Developed over 50 days, including lab work, this project highlights the potential of AI-driven tools to transform forensic investigations by providing a functional and efficient platform for law enforcement. The resulting search engine delivers a scalable and user-friendly tool optimized for forensic tasks, such as criminal identification and suspect matching, addressing real-world challenges in public safety.

INTRODUCTION:-

Forensic investigations often encounter significant challenges when traditional photographic evidence is either incomplete or entirely absent. To address these challenges, this project presents an advanced visual search system that harnesses cutting-edge artificial intelligence techniques. The system is specifically tailored for forensic applications, where law enforcement agencies require rapid, accurate, and scalable methods to retrieve relevant images from large databases using either textual descriptions or suspect images. At the heart of the system is the integration of multi-modal embeddings—a methodology that aligns visual and textual data within a shared semantic space. By converting both images and descriptive text into high-dimensional vectors, the system can effectively bridge the gap between different data modalities. For example, an image or a textual description of a suspect is mapped to a representation that is directly comparable to the embeddings of images stored in the database. This shared representation enables the system to perform efficient and precise similarity searches, even in cases where traditional photographic evidence is missing. The implementation leverages state-of-the-art models and libraries. The project uses the pre-trained CLIP model (ViT-B/32) from OpenAI for feature extraction, ensuring robust encoding of image data. This model, which has been rigorously validated in various vision-language tasks, extracts semantic features from images and transforms them into a form suitable for multi-modal matching. Beyond the technical architecture, this project significantly enhances investigative workflows. It provides law enforcement agencies with a scalable and high-performance tool that not only streamlines suspect identification but also improves tracking.

Objectives

1. Unified Embedding Representation

- Implement a cutting-edge vision-language model (VLM) such as CLIP, BLIP, or ALIGN to create a unified embedding space for both text and images.
- Ensure that semantically similar visual and textual inputs are positioned closely in the embedding space to improve search relevance.

2. Efficient Indexing & Scalable Retrieval

- Develop an optimized indexing pipeline using high-performance retrieval libraries such as FAISS, Annoy, or Milvus to store and efficiently search embeddings at scale.
- Integrate fast similarity search algorithms (k-nearest neighbors, approximate nearest neighbors) to facilitate quick and accurate image retrieval, even from large forensic databases.

3. Multi-Modal Query Support

- Enable search functionality through multiple query modalities:
 - **Text-based search:** Retrieve images based on natural language descriptions (e.g., “A person wearing a black hoodie at night”).
 - **Image-based search:** Find visually similar images from a given input image..

4. Performance Evaluation & Accuracy Enhancement

- Assess retrieval effectiveness using standard image retrieval metrics, including Precision@K, Recall@K, and Mean Average Precision (mAP).

- Conduct qualitative evaluations to ensure the system returns images that align with forensic investigation requirements.

CHALLENGES FACED:-

Dataset Selection and Availability:-One of the most significant challenges was identifying an appropriate dataset that contained both images and corresponding metadata. Initially, we started with one dataset but had to switch to another due to constraints such as lack of diversity, incomplete metadata, and limited scalability. Finding a dataset that met forensic requirements while also aligning with our vision-language model (VLM) constraints proved to be time-consuming.

Handling Missing Metadata in Images:-Crucial aspect of our system was linking images with their corresponding metadata to enhance search accuracy. However, a substantial portion of the dataset had missing metadata, making it difficult to align visual and textual representations. This required us to preprocess the data extensively by filtering out images without metadata, which reduced our dataset size and introduced additional challenges in ensuring sufficient training data.

Evaluation Metrics for Multi-Modal Queries:-Evaluating retrieval accuracy was straightforward for image-based queries but complex when combining textual queries with images. While image-based searches yielded consistent results, text-based and mixed queries produced varying performance. The challenge was in fine-tuning the model to balance retrieval accuracy across different modalities while ensuring the system returned relevant forensic matches.

Managing Different Angles in the Dataset:-Since the dataset contained both front-facing and side-profile images of individuals, ensuring that the system recognized and retrieved relevant matches from multiple perspectives was a

challenge. The embedding space needed to be robust enough to generalize across different angles, requiring additional augmentation and fine-tuning of the model.

Optimizing Search Speed and Scalability:-Given that forensic databases can contain millions of images, optimizing the retrieval speed while maintaining accuracy was a critical challenge. The initial brute-force similarity search methods were inefficient for large-scale retrieval. To overcome this, we integrated FAISS for efficient indexing, but tuning it for optimal balance between speed and accuracy required multiple iterations.

Ensuring Robustness Against Variability in Image Quality:-The dataset contained images with varying resolutions, lighting conditions, and noise levels, which affected retrieval consistency. Standardizing image quality without distorting important forensic details was a challenge that required careful preprocessing and augmentation techniques.

Computational Constraints During Model Training and Inference:-Training vision-language models and processing large datasets required significant computational resources. Running experiments on high-dimensional embeddings was time-intensive, requiring us to optimize GPU usage, batch processing, and memory allocation to maintain efficiency.

Dataset Description(IDOC Mugshot Dataset)

- **Source & Composition:**

The IDOC dataset is a large-scale forensic mugshot collection from the USA. It contains 70,008 pairs of images for unique prisoners. Each prisoner has a pair of images: one front-facing and one side-facing.

Key Characteristics:

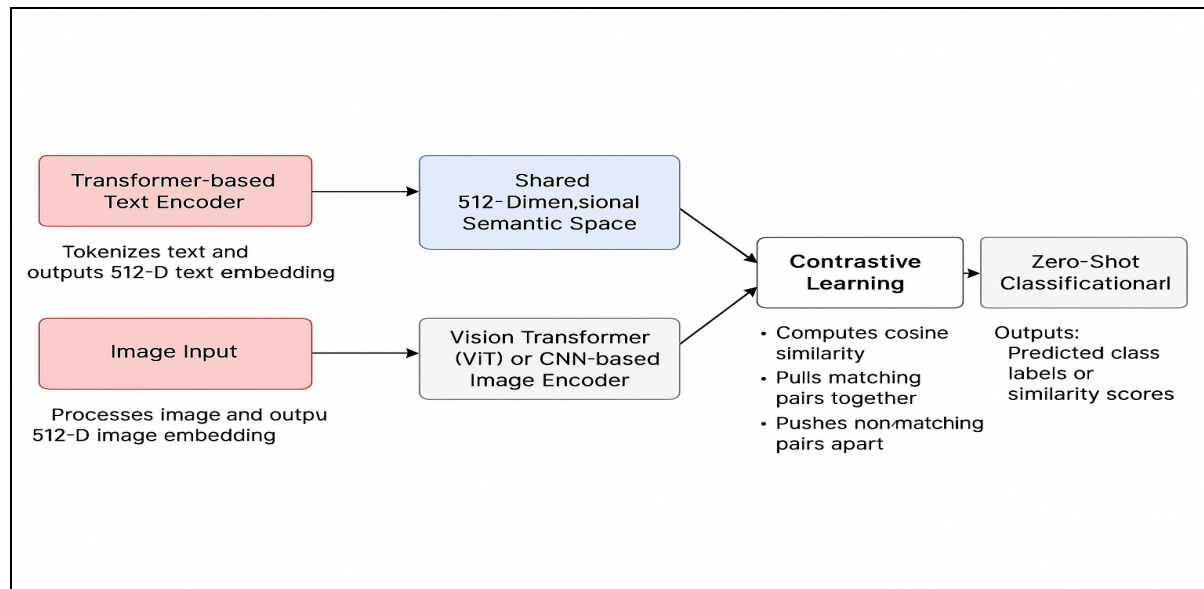
- **Image Pairs:-**Each prisoner is represented by a pair of mugshots (front and side views). For our visual search system, we primarily used the front-facing images but for the testing purpose we used side views to enhance the better performance of the model.
- **Metadata and Attributes:-**Out of 70,008 pairs, 69,827 include detailed labels describing each prisoner. These labels cover attributes such as sex, height, weight, hair color, eye color, race, and details about the type of offense. This rich metadata can be used to generate textual descriptions for multi-modal queries.
- **Scale and Diversity:-**With tens of thousands of unique individuals, this dataset is both large and diverse. Such scale is beneficial for building a robust search engine that can handle real-world variability and provide accurate retrieval results.
- **Use in Project:-**The mugshot images are embedded into a shared representation space (using the fine-tuned CLIP model) and indexed (using FAISS). When a query is made—whether it is a sketch, photo, or text description—the system searches for the closest matching mugshots from this database.
- **Multi-Modal Querying:-**The IDOC dataset’s rich metadata allows for text-based queries (e.g., “male, black hair, brown eyes”) while its large-scale photographic data supports image-based retrieval. By embedding both types of inputs into the same space, the system can perform multi-modal search seamlessly.

MODEL DESCRIPTION AND ARCHITECTURE:-

To effectively align multiple input modalities—specifically real images and textual descriptions—**Contrastive Language-Image Pretraining (CLIP)** was chosen as the core embedding model. CLIP is particularly powerful for this application because it learns a shared semantic space where visually and textually similar concepts are positioned close together. By leveraging its pre-trained knowledge and fine-tuning it on the IDOC Mugshots dataset, the model is optimized to handle forensic queries where text-based inputs need to be accurately mapped to real face images.

CLIP is used to generate embeddings for two types of inputs:

- **Real Images:** Photographs from the IDOC Mugshots dataset are processed through CLIP's image encoder to obtain corresponding embeddings.
- **Text Descriptions:** Textual inputs (for example, "a young man with short black hair and a scar on the left cheek") are processed via CLIP's text encoder to create semantic representations that are aligned with the visual data.

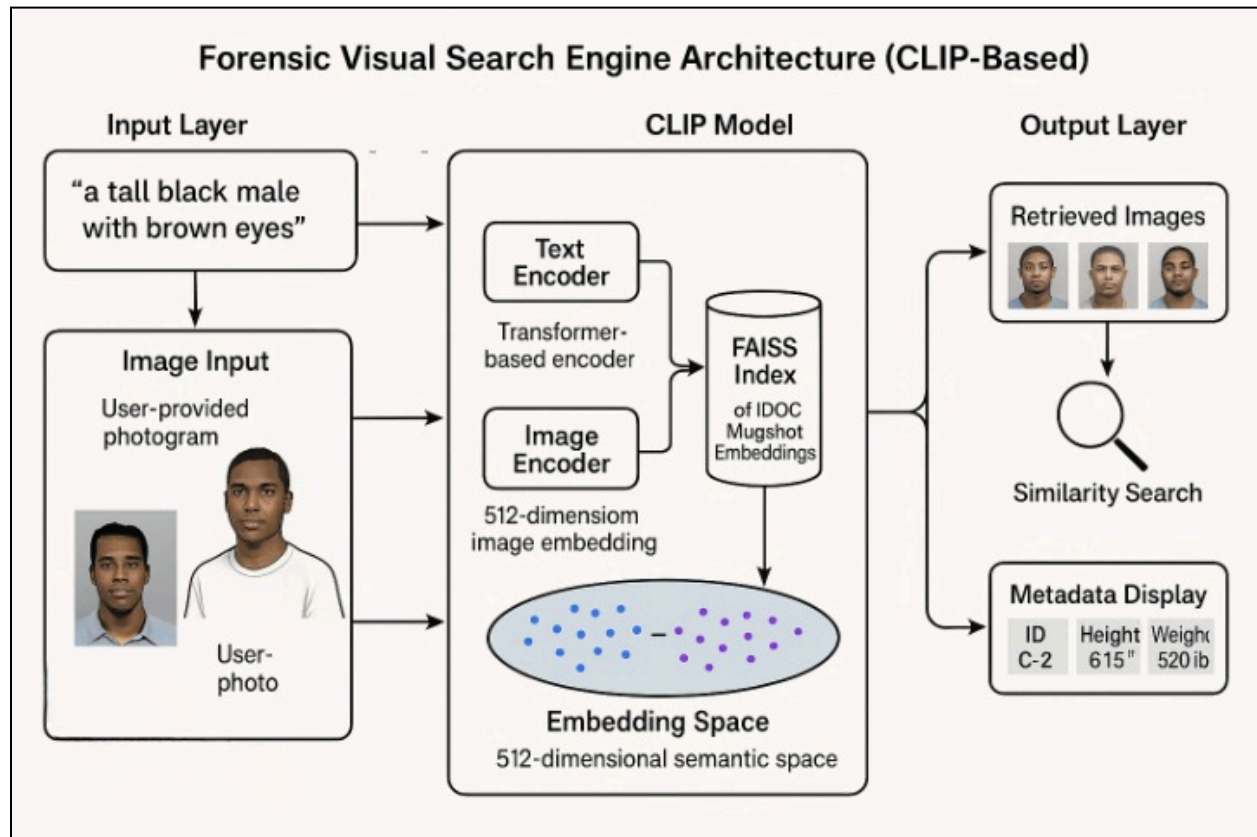


CLIP (Contrastive Language–Image Pre-training) is a neural network that learns visual concepts directly from natural language supervision. Rather than relying on traditional, labor-intensive labeled datasets, CLIP is trained on a vast collection of image–text pairs gathered from the internet. This training paradigm enables the model to understand a wide range of visual concepts simply by associating them with their natural language descriptions. By learning from noisy, unfiltered data, CLIP is able to perform zero-shot classification—meaning it can correctly classify images into categories it hasn’t been explicitly trained on by using natural language prompts. At its core, the architecture of CLIP consists of two main components: a text encoder and an image encoder, both of which project their respective inputs into a shared semantic space. The text encoder is based on a Transformer architecture, similar to those used in models like GPT. It processes user-provided textual descriptions by tokenizing the input and then generating a fixed-length embedding (typically a 512-dimensional vector) that captures the semantic meaning of the text. In parallel, the image encoder—often implemented as a Vision Transformer (ViT) or sometimes as a modified convolutional neural network (e.g., ResNet)—processes input images to produce its own 512-dimensional

embedding. The key innovation in CLIP's training is its contrastive learning objective. During training, the model is presented with batches of image-text pairs. For each image in a batch, the model attempts to predict the correct textual description among a set of randomly sampled texts. This is achieved by maximizing the cosine similarity between the embeddings of the true image-text pair while minimizing the similarity for all other, non-matching pairs in the batch. In effect, the model "pulls" the embeddings of corresponding images and texts closer together in the shared embedding space and "pushes" apart those that do not match. This strategy not only helps CLIP to learn rich, generalizable features for both modalities but also underpins its impressive zero-shot capabilities.

Once trained, CLIP can be applied to a variety of visual classification tasks without any additional fine-tuning. To classify an image, one simply provides a set of candidate textual descriptions (such as "a photo of a dog" or "a photo of a cat"), and CLIP computes the similarity between the image's embedding and each text embedding. The description with the highest similarity score is chosen as the prediction. This flexibility allows CLIP to adapt quickly to new domains or tasks, making it a powerful tool for applications like content moderation, search, and even artistic creation.

PROJECT FLOW



1. Input Layer – Query Input

Text Input:

Users can start with a textual query that describes a suspect or scene (e.g., “A tall black male with brown eyes”). The text is fed into a text encoder that transforms natural language into a dense vector representation. This modality is particularly useful when detailed descriptions are available but no visual example exists.

Image Input:

Users can also provide an image, such as a forensic sketch or a photo, as a query. The image is processed through an image encoder, generating a corresponding

embedding that captures visual features. This approach is beneficial when a visual reference is needed to find similar images in the database.

Multimodal Input (Text + Image):

In some cases, a combination of text and image is used. For instance, a user might supply an image with an accompanying note (e.g., “Find similar images to this photo, but with a beard”).

The system processes both the text and image, generating separate embeddings, which are then fused using a weighted sum or concatenation to create a single hybrid representation. This multimodal approach refines the search by combining the strengths of both modalities—visual details from the image and contextual or attribute information from the text.

2. CLIP Model – Embedding Generation

Text Encoder:

The text encoder component of CLIP (or similar state-of-the-art vision-language models) processes the text input to produce a 512-dimensional embedding. It translates the nuances of natural language into a format that can be compared with visual data.

Image Encoder:

The image encoder processes the visual input (photo, sketch, etc.) and converts it into a 512-dimensional vector embedding. This embedding captures essential visual features like shape, color, and texture.

Multimodal Fusion:

When both text and image are provided, their individual embeddings are combined into a unified representation. Fusion techniques (like weighted sums or concatenation) are used to integrate the information, ensuring that both the descriptive context and visual cues contribute to the final embedding. This combined embedding is more robust for retrieval as it leverages complementary information from two different sources.

3. FAISS Index – Efficient Storage & Retrieval

Indexing of Embeddings:

All image embeddings in the forensic database are precomputed and stored in a FAISS index. FAISS (Facebook AI Similarity Search) is a library that enables fast, scalable similarity searches in high-dimensional spaces. The use of FAISS ensures that even with large databases, the search process remains efficient and responsive.

Scalability:

FAISS is optimized for high-dimensional vector searches and can handle millions of entries, which is critical in forensic applications where datasets are large. This efficient indexing makes it possible to quickly compare a query embedding (from text, image, or multimodal input) with all stored embeddings.

4. Similarity Search – Image Retrieval

Distance Computation:

Once a query embedding is generated (from any of the input types), the system calculates the distance between this query and the embeddings stored in the FAISS

index. Techniques like Approximate Nearest Neighbors (ANN) are used to find the closest matches.

Retrieval Process:

The system retrieves the top matching images based on the computed similarity scores. This step is crucial for forensic investigations, as it ensures that the most semantically similar images (whether queried via text, image, or a hybrid approach) are presented to the user.

Handling Multimodal Queries:

In the case of multimodal inputs, the fused embedding is compared against the stored image embeddings, ensuring that the search accounts for both visual and contextual textual details.

5. Output Layer – Retrieved Images & Metadata Display

Result Presentation:

The final step involves presenting the retrieved images to the user. Each image is displayed along with its associated metadata (e.g., identification numbers, physical attributes, etc.). The metadata provides additional context, which is especially important in forensic applications for quickly narrowing down suspects or identifying key details.

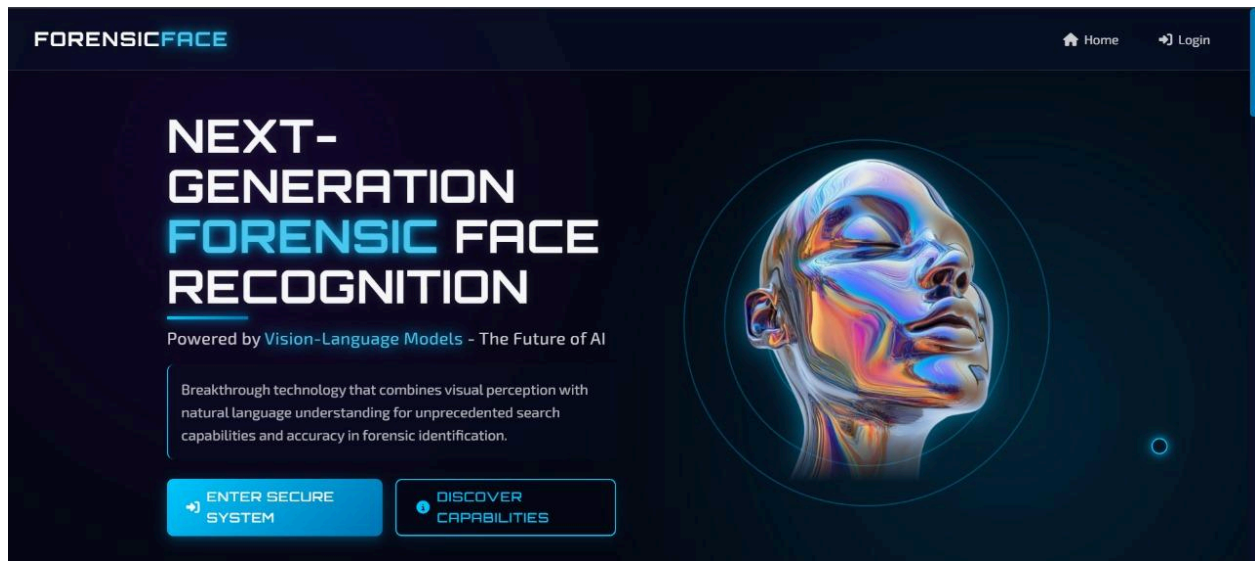
Enhanced Forensic Analysis:

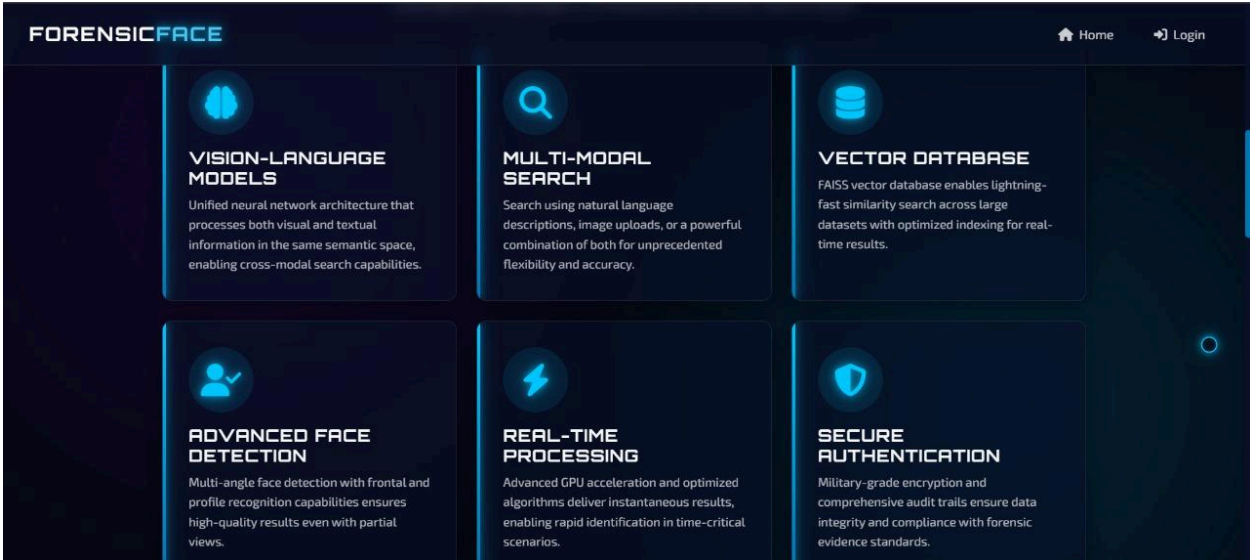
By combining both visual data and rich metadata, the output layer ensures that investigators receive a comprehensive view of potential matches. The interface is

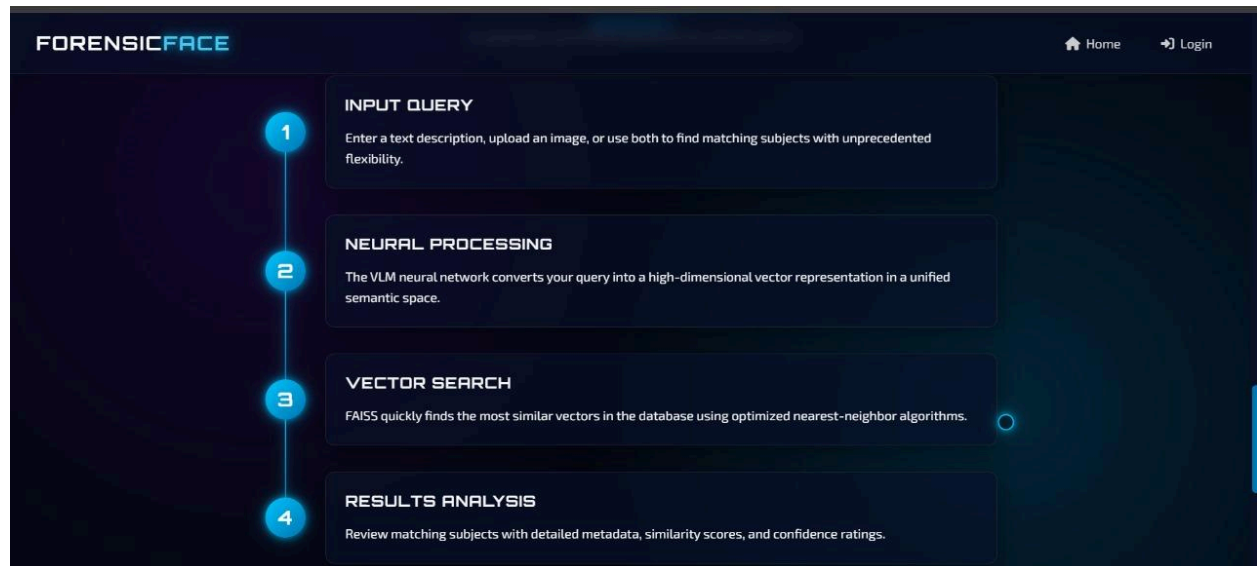
designed to allow users to review and verify matches efficiently, contributing to a more streamlined and accurate forensic investigation process.

OUTPUT AND INFERENCE

INTERFACE:-







1. Login Page via Dashboard

FORENSICFACE Home Login

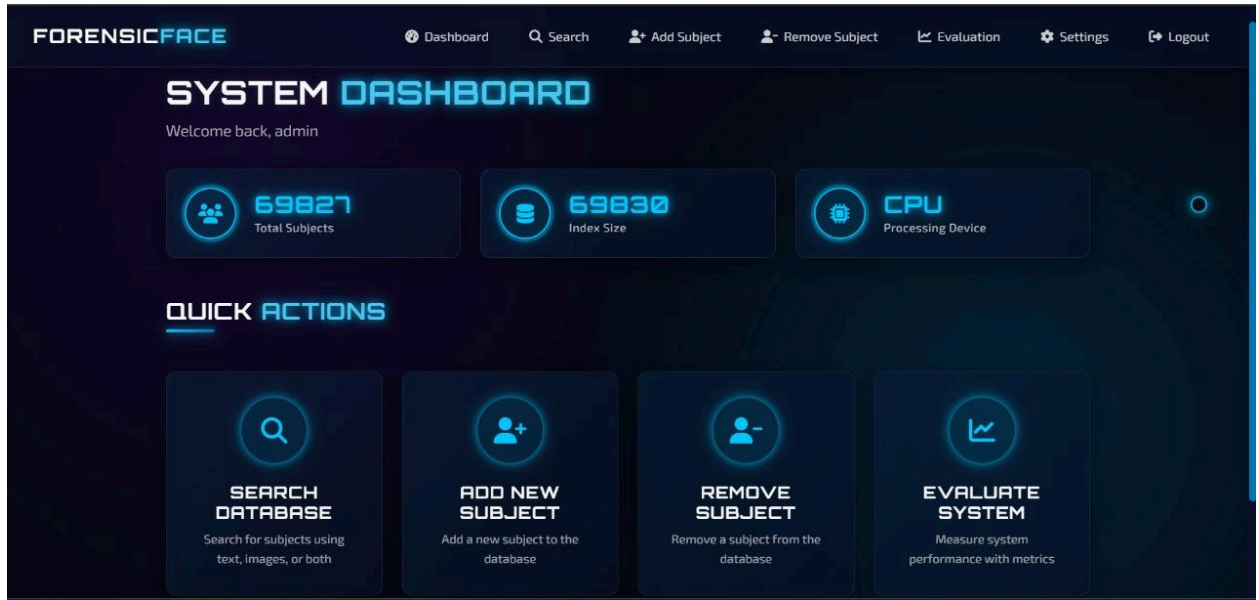
SYSTEM LOGIN
Enter your credentials to access the system

Username

Password

LOGIN

© 2025 Forensic Face Recognition System. All rights reserved.

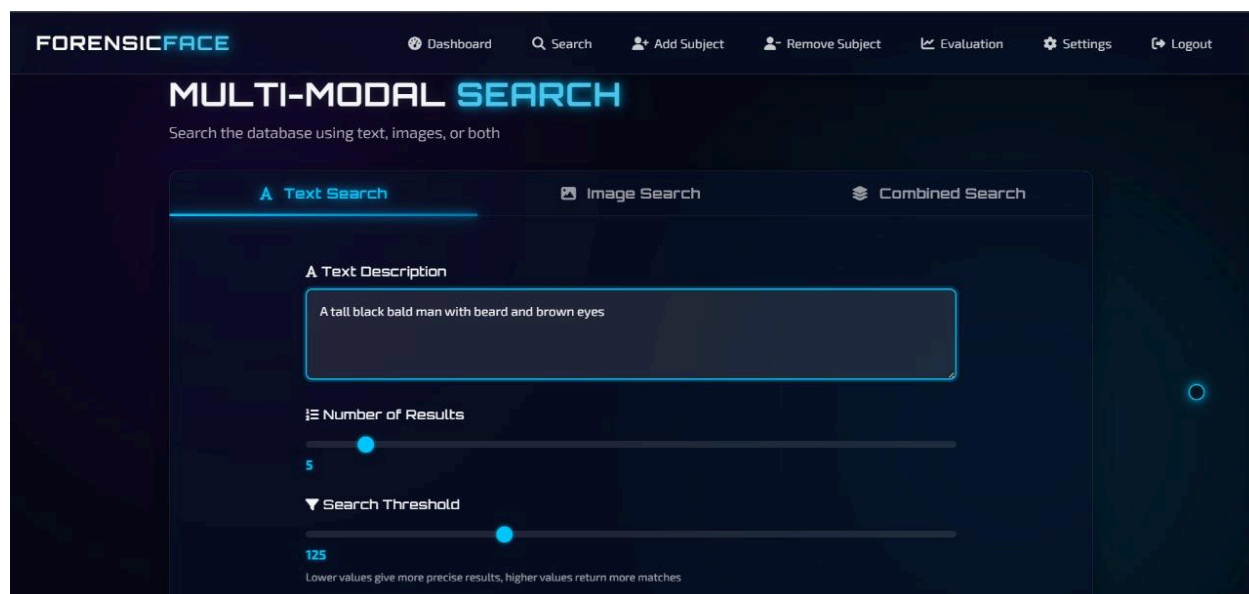


2. Text-Based Queries

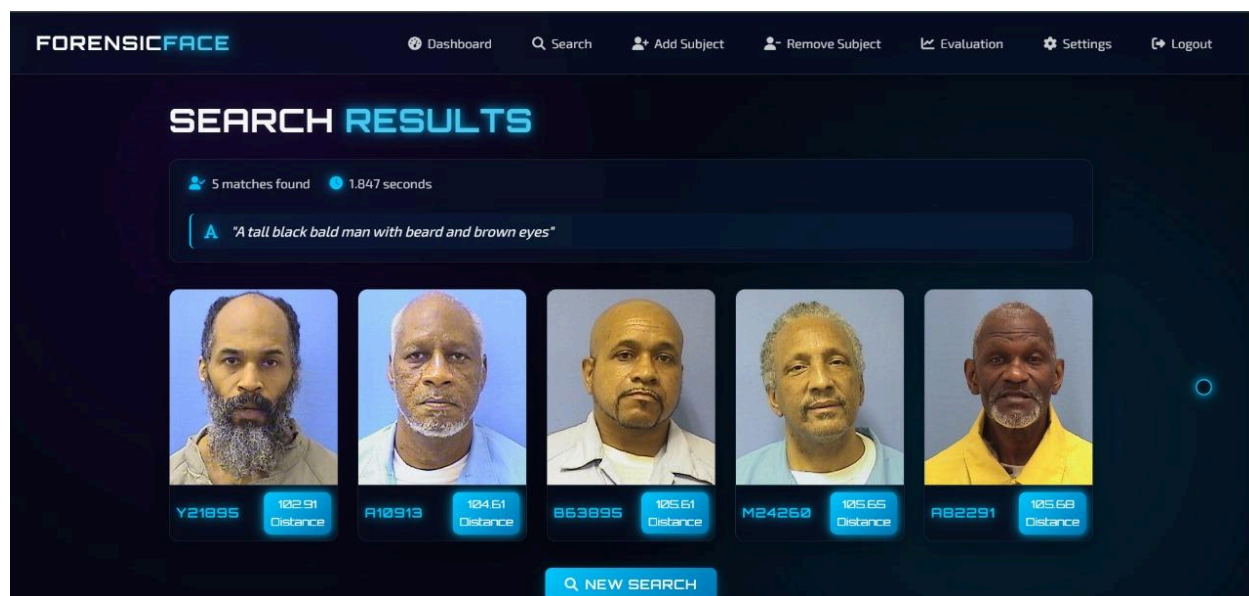
User input in textual description: *“a tall black bald man with beard and brown eyes”*,

the system:

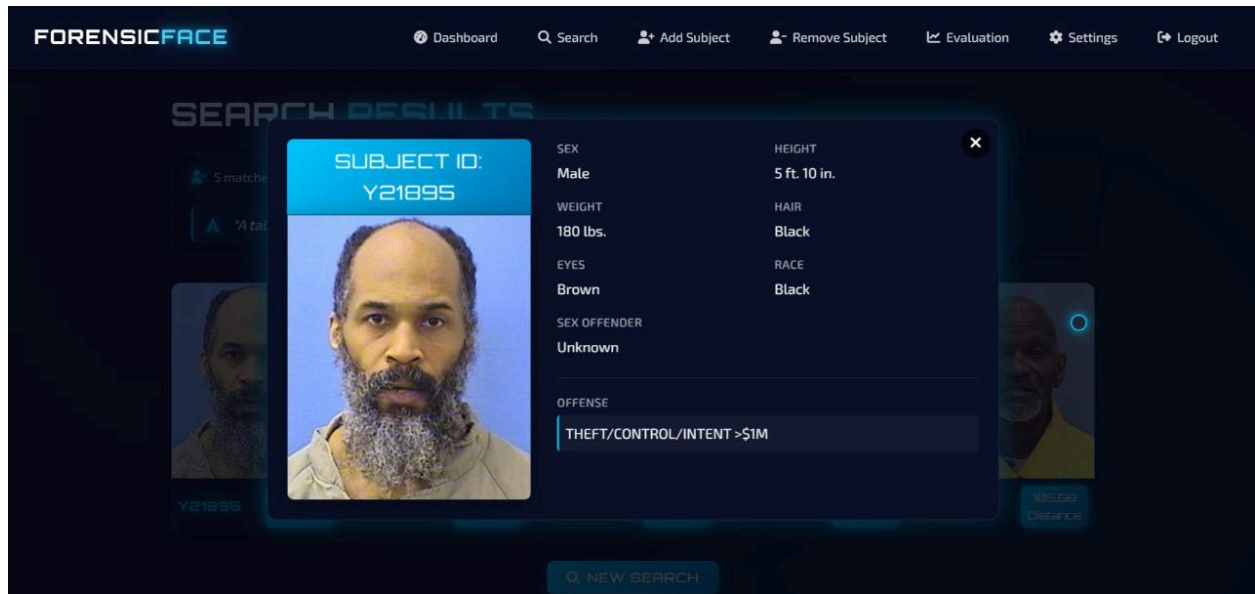
1. **Processes the text** through CLIP’s text encoder to generate an embedding.
2. **Searches the FAISS index** for visually similar faces based on the computed embedding.
3. **Displays ranked results**, prioritizing those that best match the description



Multimodal Interface



Retrieval Output



Subject Metadata

Inference:-

Overview of Distance and Similarity Score:

- In the embedding-based retrieval system depicted in the images, the distance measures the separation between an image's embedding and the query embedding within CLIP's vector space.
- A smaller distance indicates a closer match, signifying that the image is more similar to the query.
- The similarity score is typically the negative of the distance (or a related function), where a smaller distance corresponds to a higher (less negative) similarity score, and a larger distance results in a lower similarity score.

Analysis Based on Retrieval Output Image:

- The query *"a tall black bald man with beard and brown eyes"* produces the following results:
 - First result (Subject ID: Y21895) has a distance of 102.91, with an implied similarity score of approximately -101.91.

- Second result (Subject ID: A10913) has a distance of 124.61, with an implied similarity score of approximately -123.61.

- Since 102.91 is less than 124.61, the first result (Y21895) is ranked higher, indicating it is more similar to the query.

Validation and System Performance:

- The ranking demonstrates the system's capability to prioritize matches based on proximity in the embedding space.

- This is corroborated by the metadata and visual alignment in the **Subject Metadata** image for Y21895, validating the accuracy of the top-ranked result.

3. Image-Based Queries

Forensic criminal face images are uploaded by investigators, and the system follows these steps:

1. **Preprocesses and standardizes the photo** (resizing, normalization, etc.).
2. **Passes the image through CLIP's image encoder** to generate an embedding.
3. **Conducts a similarity search in the FAISS index** to find real faces that resemble the image.
4. **Returns the most similar faces**, ranked by confidence scores.

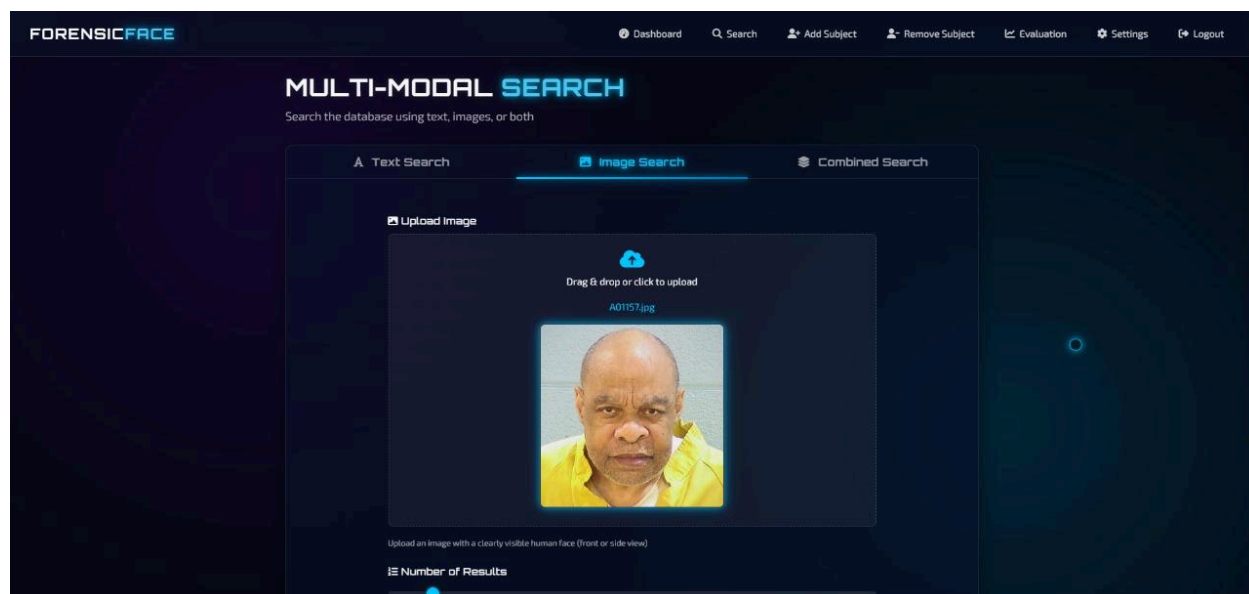
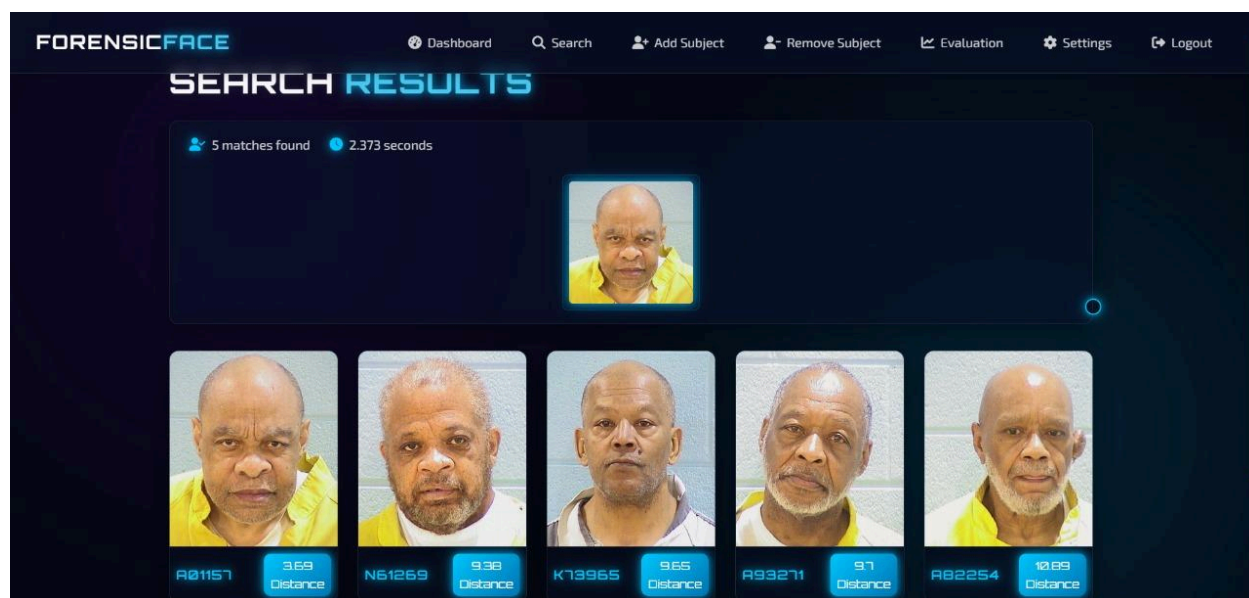
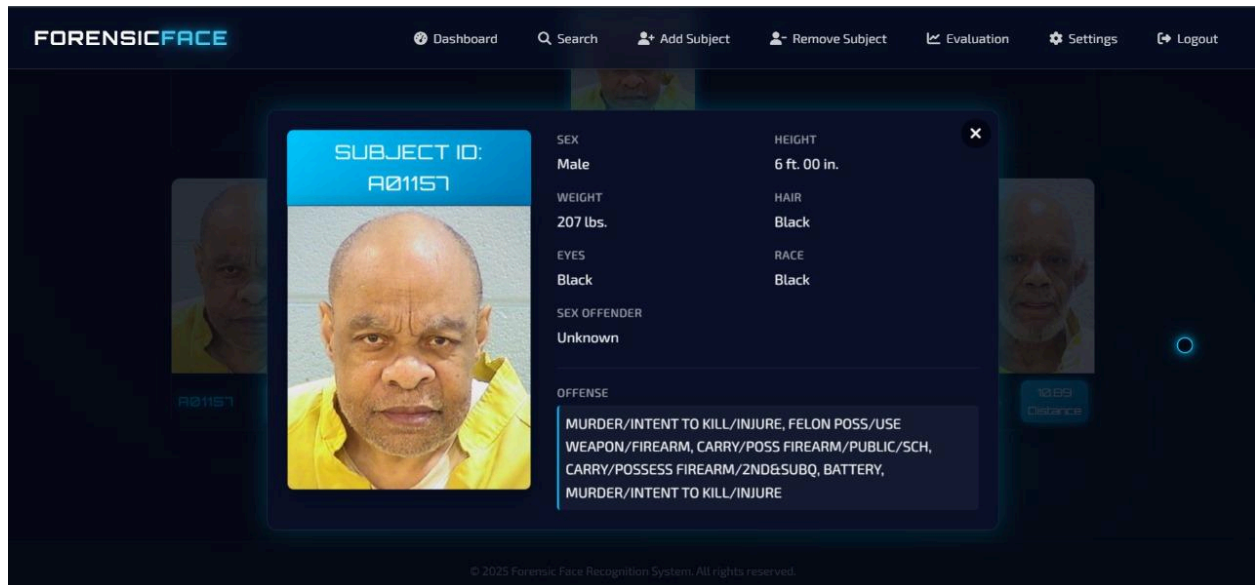


Image Search Interface



Subject Details (A0157)



Search Results (A0157)

Inference:-

Exact Match (Distance = 0.000, Similarity Score = 1.000):

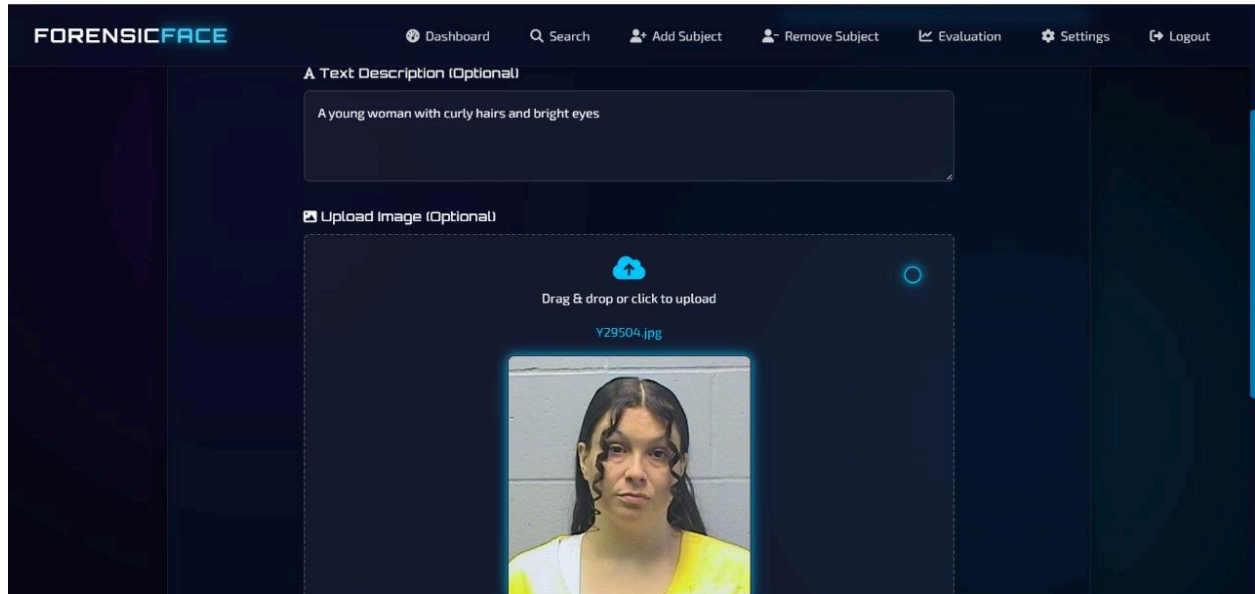
- Distance = 0.000: Indicates that the query image (e.g., from **Image Search Interface** with file A0157.jpg) is found exactly in the database, implying the embedding vectors are identical.
- Similarity Score = 1.000: Represents a perfect similarity, confirming that the individual in the query image matches the subject in the database (e.g., Subject ID: A0157 in **Subject Details (A0157)**).

Second Result (Distance \approx 3.69, Similarity Score \approx -3.69):

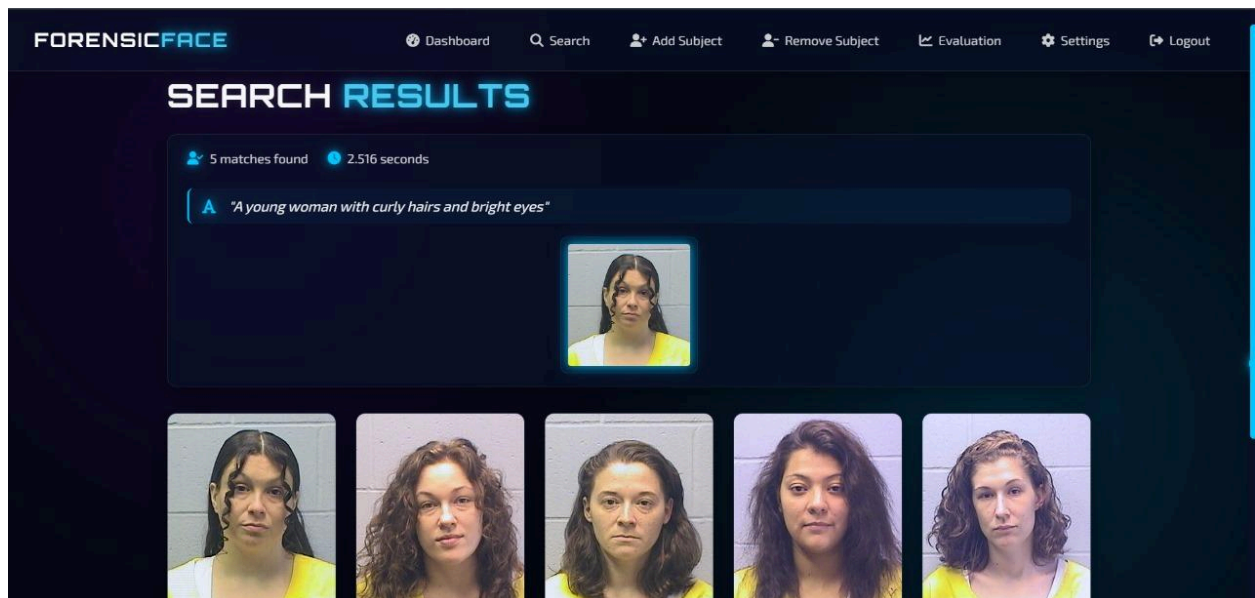
- A distance of approximately 3.69 (as shown in **Search Results (A0157)** for Subject ID: A0157) indicates the second result is farther in the embedding space compared to an exact match, suggesting it is less similar to the query image.
- The negative similarity score of approximately -3.69 reflects that a higher distance corresponds to lower similarity. Despite this, it ranks as a top result due to shared visual features (e.g., general facial structure, bald head) with the query, as validated by the metadata in **Subject Details (A0157)**.

3. Multi Model Approach (Text +Photo Combination)

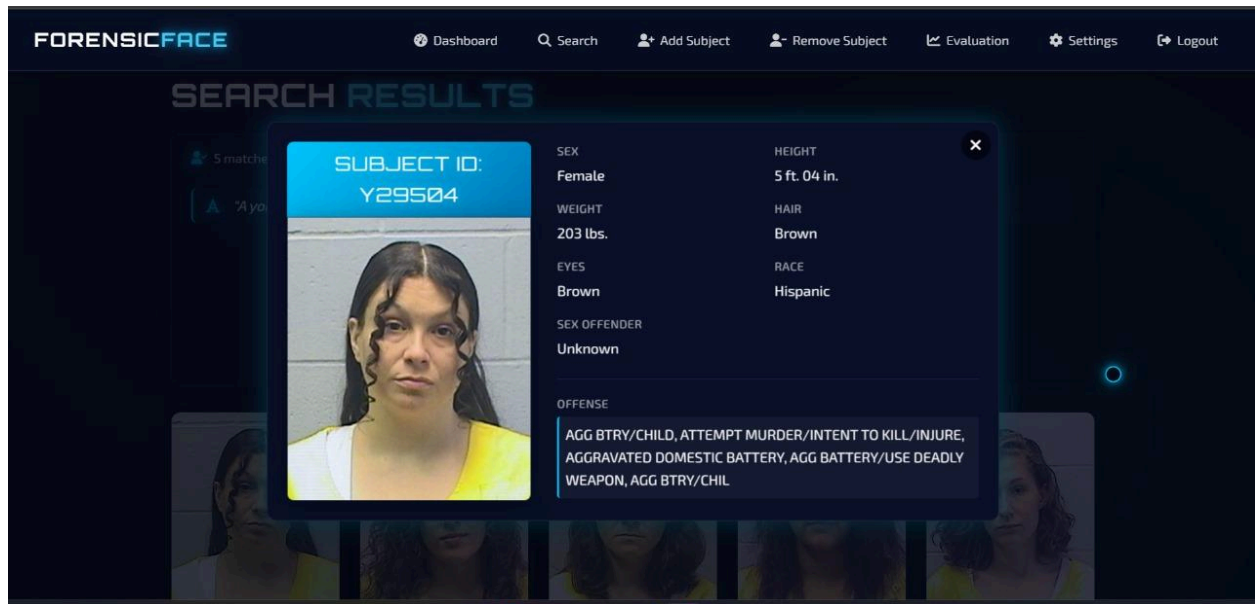
For enhanced search flexibility, the system supports **multi-modal queries**, where a photo and a text description are combined. The embeddings of both modalities are **fused to refine search results**, allowing investigators to narrow down matches using additional descriptive details.



Combined Search Interface



Search Results (Y29504)



Subject Details (Y29504)

Inference

Exact Match (Distance = 0.000, Similarity Score = 1.000):

- Distance = 0.000: Indicates that the query image (e.g., from **Combined Search Interface** with file YZ9504.jpg) is found exactly in the database, implying the embedding vectors are identical.
- Similarity Score = 1.000: Represents a perfect similarity, confirming that the individual in the query image matches the subject in the database (e.g., Subject ID: (Y29504) in **Subject Details(Y29504)**).

Second Result (Distance \approx 9.38, Similarity Score \approx -9.38):

- A distance of approximately 9.38 (as shown in SearchResults(Y29504) for Subject ID: (N61269)) indicates the second result is farther in the embedding space compared to an exact match, suggesting it is less similar to the query image.
- The negative similarity score of approximately -9.38 reflects that a higher distance corresponds to lower similarity. Despite this, it ranks as a top result due to shared visual features (e.g., curly hair, general facial structure) with the query, as validated by the metadata in **Subject Details(Y29504)**.

Deployment and Testing

After demonstrating robust performance during evaluation, the system proceeds to a structured deployment and testing phase designed to ensure smooth real-world integration and efficient handling of forensic-scale data. In the initial prototype testing phase, the system is rigorously evaluated on a smaller subset of data to validate core functionalities. For instance, during query processing, we verify that the CLIP model accurately generates embeddings for both text and image inputs, ensuring that each query is correctly transformed into its semantic representation. The FAISS-based indexing and retrieval pipeline is then tested to confirm that it consistently retrieves the expected matches from the database, while the ranking performance is scrutinized to ensure that the most relevant images align closely with the ground truth data. Subsequently, the system is gradually scaled to process the full iDOC Mugshot database, comprising 70,000 images, to stress-test its performance under realistic forensic conditions. In addition to these tests, we also evaluate the system using random face images that are not part of the dataset. In this scenario, we have adjusted the similarity threshold such that if the distance between the query embedding and the database embedding falls below the set value, the system returns the most relevant matches, thereby ensuring accuracy even with out-of-dataset queries. Finally, the system is tested with non-face images to verify its robustness; if an uploaded image does not contain a human face, the system is designed to return a clear message indicating that a human face was not detected. This comprehensive deployment and testing approach ensures that our forensic visual search system is both reliable and scalable for real-world applications.

1. Adding New Images (Case 1):

FORENSICFACE

[Dashboard](#)[Search](#)[Add Subject](#)[Remove Subject](#)[Evaluation](#)[Settings](#)[Logout](#)

ADD NEW SUBJECT


Add a new subject to the database with front and side images

SUBJECT IMAGES

Front Image

Drag & drop or click to upload


em-front.jpeg



Side Image

Drag & drop or click to upload

em-side.jpeg



FORENSICFACE

[Dashboard](#)[Search](#)[Add Subject](#)[Remove Subject](#)[Evaluation](#)[Settings](#)[Logout](#)

SUBJECT METADATA

Sex

Female

Height

5

ft.

05

in.

Weight

115

lbs.

Hair Color

Brown

Eye Color

Brown

Race

White

Sex Offender

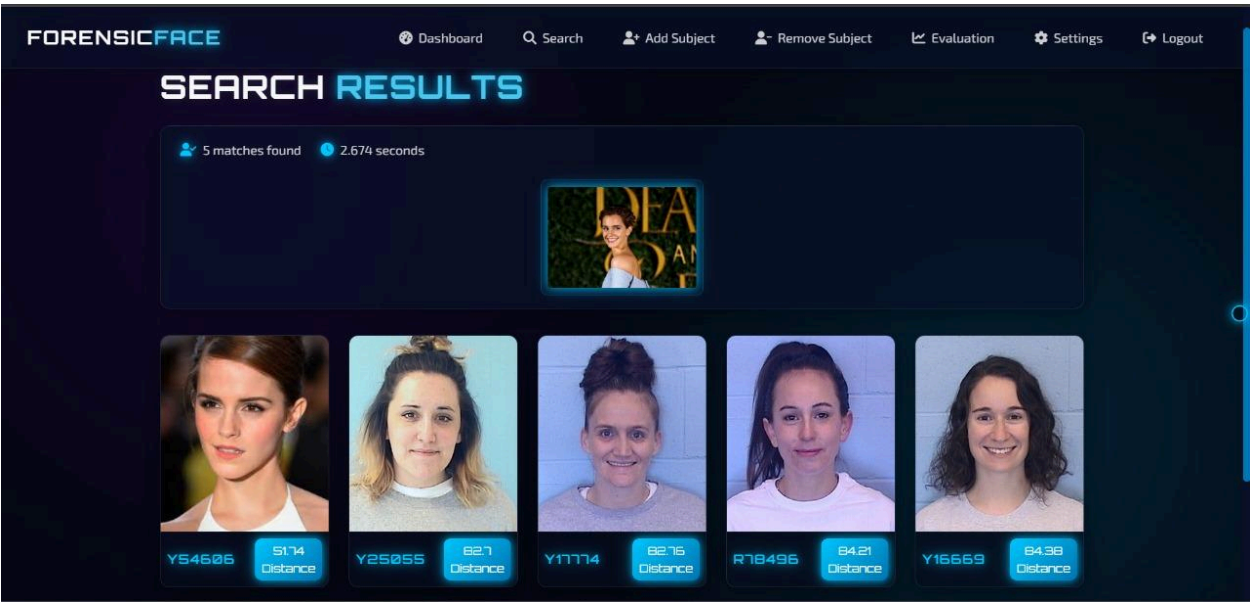
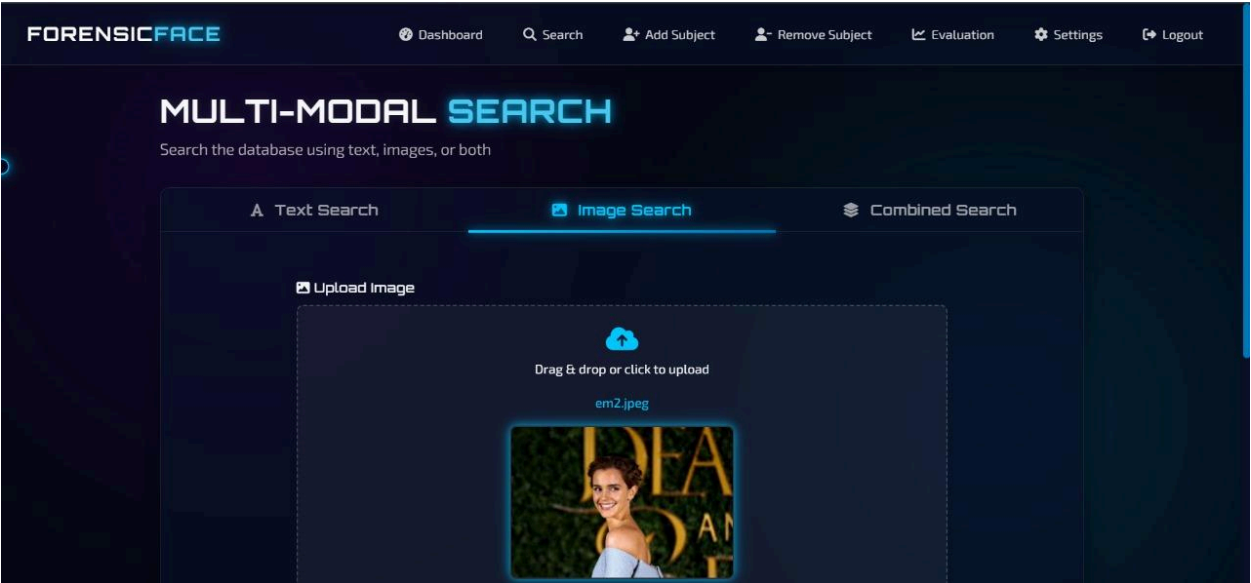
No

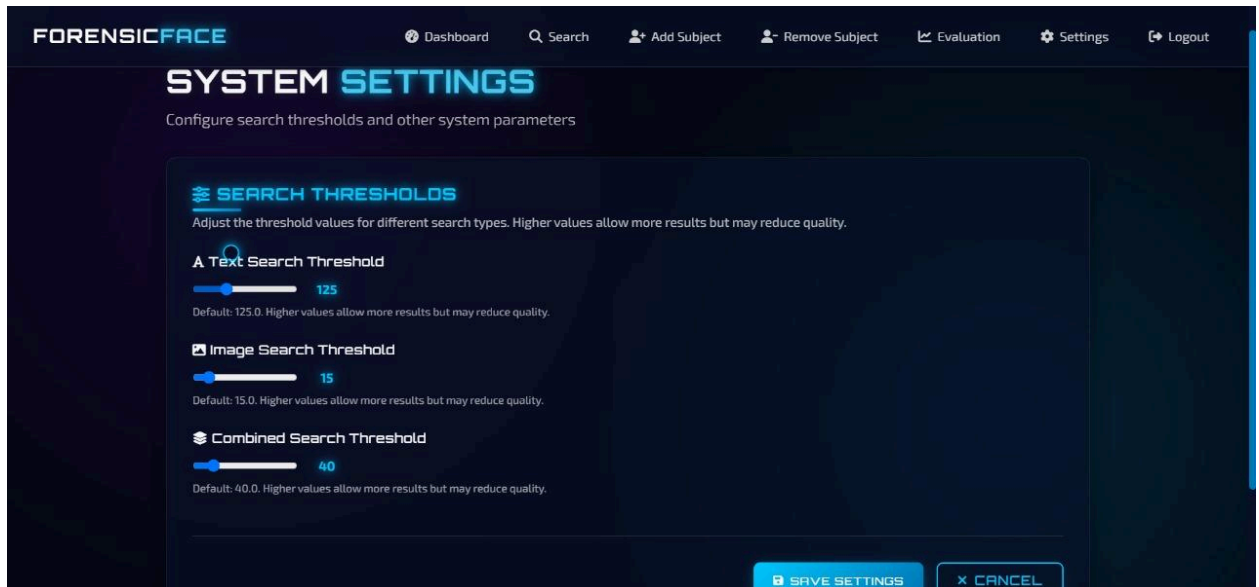
Offense(s)

AGGRAVATED THEFT

ADD SUBJECT

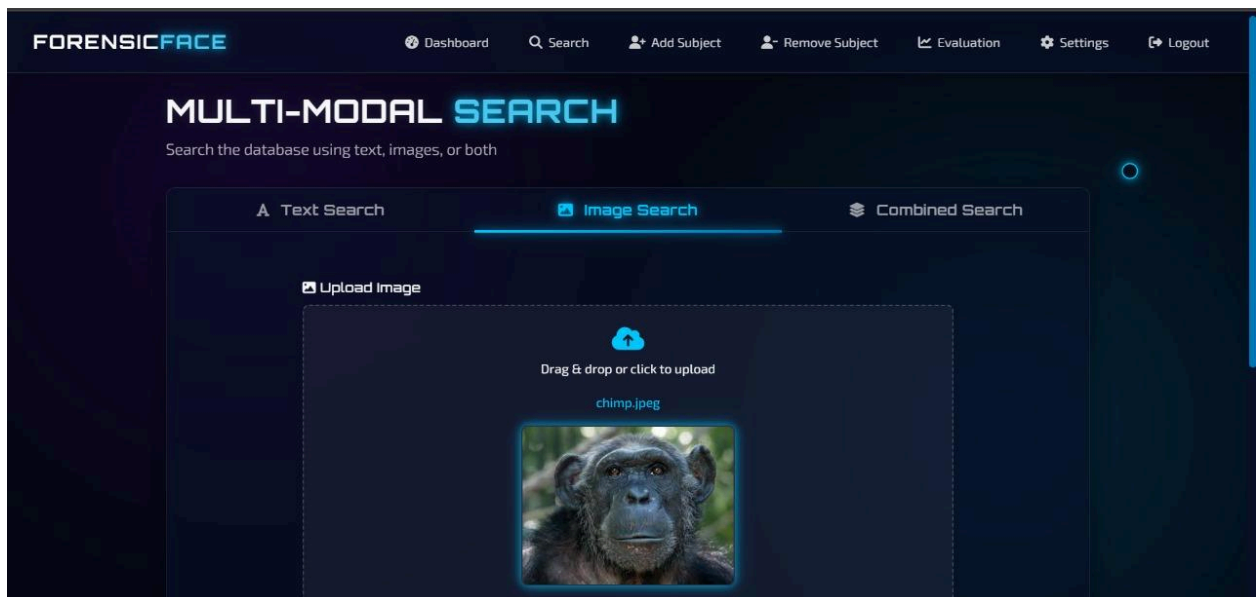
CANCEL

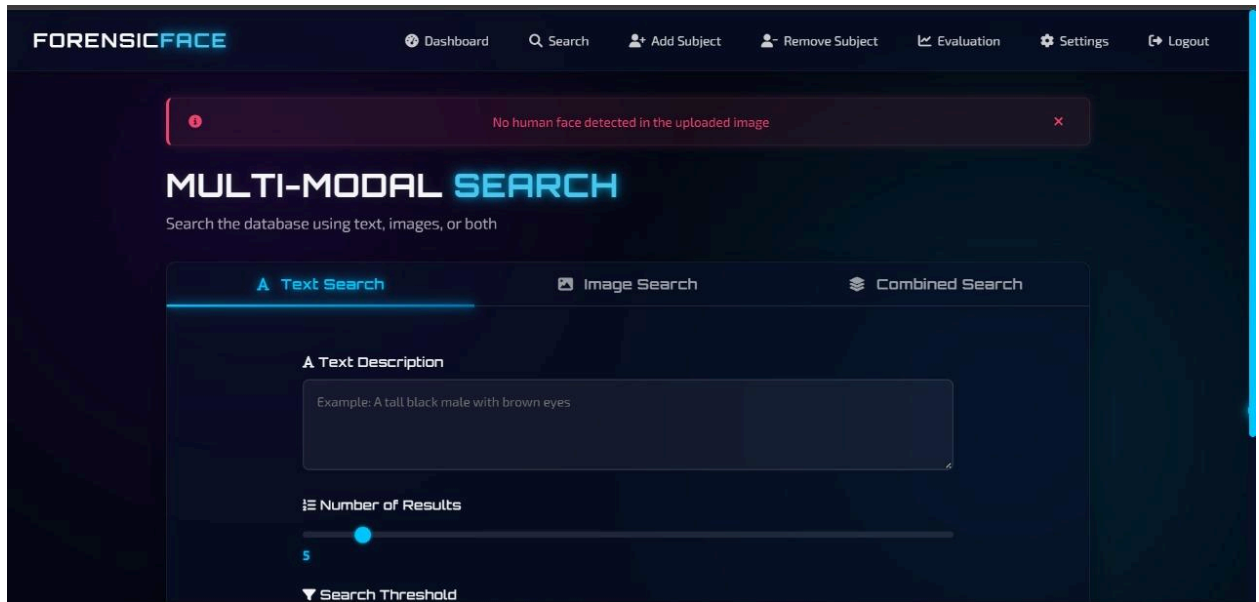




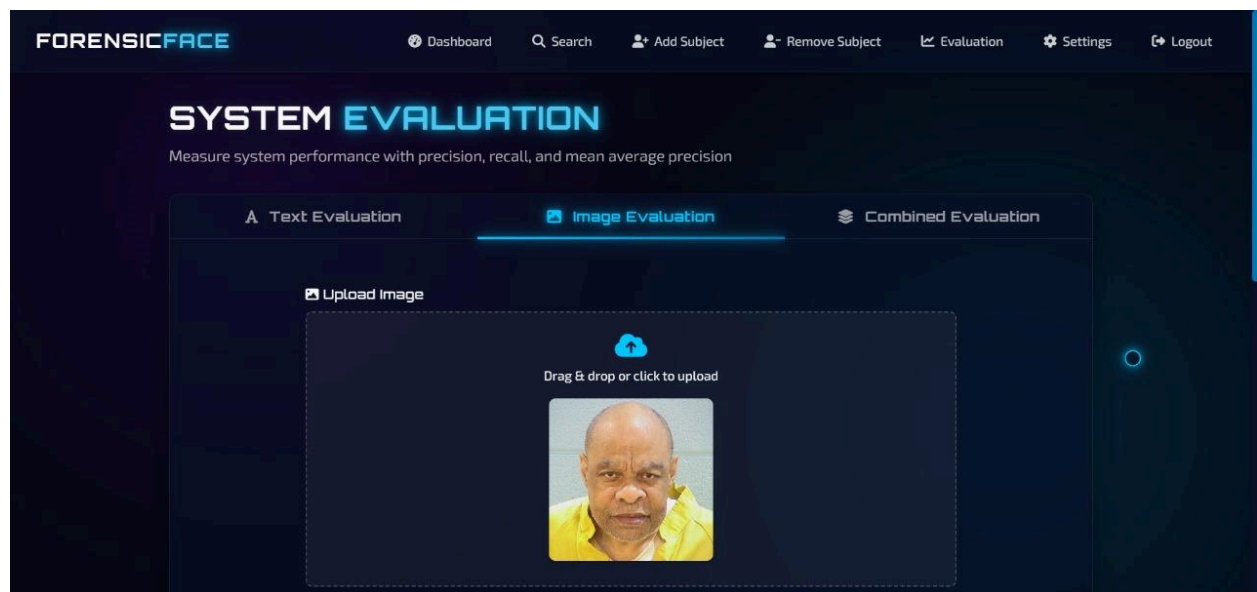
Threshold Adjustments

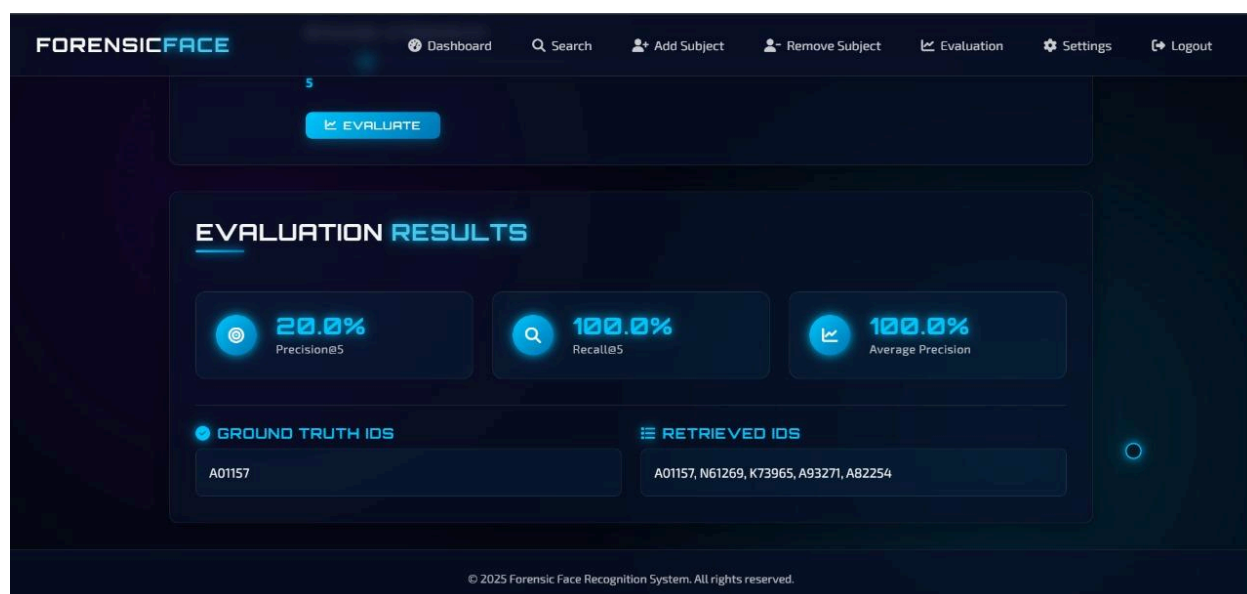
2. Trying with Non human face (Case 2):





Evaluation Metrics:-





The evaluation results indicate a 100% recall@5, demonstrating that the system successfully retrieves all relevant IDs, such as the ground truth ID A0157, within the top five results, alongside additional retrieved IDs like N61269, K73965, A9327N, and AB2254. However, the precision@5 of 20% and average precision of 10% reveal significant limitations, with 80% of the retrieved IDs being irrelevant, pointing to a high rate of false positives and a clear need for enhancing the system's accuracy to improve its reliability in forensic contexts. The interface also supports image-based evaluation, as evidenced by the uploaded photo of a man in a yellow jacket, enabling users to assess performance by uploading images for processing against a database.

The evaluation metrics precision, recall, and mean average precision provide a comprehensive framework for assessing the ForensicFace system's performance. Precision@5 measures the proportion of relevant items among the top five retrieved results, in this case reflecting the system's struggle to filter out irrelevant matches. Recall@5 evaluates the system's ability to retrieve all relevant items within the same top five, showcasing its strength in inclusivity despite accuracy issues. Mean average precision averages the precision scores across multiple ranks or queries, offering an overall indicator of performance consistency, which at 10% suggests room for improvement. Together, these metrics highlight the system's capability to identify all relevant matches but underscore the importance of

refining its algorithms to enhance precision and reduce erroneous results, critical for effective forensic use.

CONCLUSION:-

In conclusion, the development of our forensic visual search engine demonstrates the powerful potential of integrating state-of-the-art vision-language models with efficient similarity search mechanisms for forensic investigations. By leveraging the CLIP model, we successfully established a unified embedding space that enables accurate and scalable retrieval of images using text, image, or combined multimodal queries. The implementation of a FAISS-based indexing pipeline further ensured that even large forensic databases could be searched rapidly and effectively.

Throughout the project, we addressed several challenges ranging from dataset selection and handling missing metadata to optimizing search performance and managing multiple image perspectives. Overcoming these obstacles has not only enhanced the robustness of our system but also provided valuable insights into the practical integration of deep learning and efficient search algorithms in real-world applications.

Overall, our work contributes to the modernization of forensic investigation methods by offering a tool that improves the speed and accuracy of suspect identification. Future enhancements could focus on refining multimodal fusion techniques and expanding evaluation metrics to further adapt to the evolving needs of forensic analysis.

REFERENCES

- [1] <https://huggingface.co/>
- [2] <https://www.intel.com/content/www/us/en/developer/tools/opencv-toolkit/overview.html>
- [3] <https://github.com/opencv-toolkit/opencv>
- [4] <https://iab-rubric.org/index.php/iiit-d-sketch-database>
- [5] <https://www.kaggle.com/datasets/elliott/idoc-mugshots>
- [6] <https://medium.com/@az.tayyebi/bridging-vision-and-language-exploring-clip-blip-and-owl-vit-f8f6a99a263e>