



## MULTIMODAL DEEP LEARNING FRAMEWORK FOR MISSING CHILDREN DETECTION USING VISION LANGUAGE MODE

Kola Uma Maheswari<sup>1</sup> and Pandluri Dhanalakshmi<sup>2</sup>

<sup>1</sup>Department of CS &CA, Mohan Babu University, Tirupathi 517014, India.

<sup>2</sup>Department of AI&ML, Mohan Babu University, Tirupathi 517014, India.

<sup>1</sup><http://orcid.org/0000-0005-1358-8821>, <sup>2</sup><http://orcid.org/00000-0002-4118-821X>

Email: [kola.uma1978@gmail.com](mailto:kola.uma1978@gmail.com), [mallidhana5@gmail.com](mailto:mallidhana5@gmail.com)

### ARTICLE INFO

#### Article History

Received: February 10, 2026

Reviewed: June 5, 2026

Accepted: July 5, 2026

Published: July 31, 2026

#### Keywords:

Missing children detection,  
Face re-identification,  
Surveillance analysis,  
CLIP,  
Vision-language models,  
Graph neural networks,  
Multimodal deep learning,  
Person tracking.

### ABSTRACT

As report of missing children continue to rise around the world there is an urgent requirement of some smart and scalable systems which can help in timely and accurate identification. In this paper, we propose a multimodal deep learning system that combines Vision-Language Transformers like CLIP, Face Re-identification networks, and Graph Neural Networks (GNNs) in order to help in locating and tracking missing children. The proposed system is unlike conventional facial recognition systems, which can only act on visual stimulus whereas the proposed system handles a multitude of different stimulus e.g. facial image, descriptive text, metadata location, time greater and context of the stimulus e.g. description of clothing. The vision language module enables visual inputs to be correlated to the textual description so that the model is able to interpret and correlate the narrative indicators within the public reports. A face re-identification module with high tolerances to visual variability because of age, lighting, occlusion, etc. and a GNN to learn spatial and temporal links between sightings to make movement pattern predictions. Initial tests on publicly available data show a better matching across the unimodal baselines and a substantial improvement in recall and contextual relevance in the metrics. Uniting the visual, linguistic, and structural data, the model provides more accurate findings in real-life situations and demonstrates a high possibility of combining it with police technology. There should be more work on increasing child-targeted multimodal data sets and evaluating in the real world testing.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

### I. INTRODUCTION

Missing children as a problem still remains a severe humanitarian and social dilemma where many cases are reported daily on a worldwide scale. Manual reporting, restricted surveillance videos and recalling by human beings forms the core of the conventional search and identification procedures that are highly ineffective, inefficient, and time consuming in quick interventions. Since time plays a decisive role in the recovery of the children, technological systems are urgently needed that would be able to help authorities with greater precision, faster, and more flexibly.



Figure 1: Missing Children.  
Source: Authors, (2026).

The recent developments of artificial intelligence, especially computer vision and deep learning, provide the new opportunities to automate the identification task. Contemporary models have now shown great results in determining faces, image-text comprehension, and video processing. Nonetheless, the current solutions in the field of missing persons detection are limited to using unimodal methods (without context knowledge). An example would be that the systems can perform well in optimal situations with facial recognition but the problem is the systems cannot perform well in situations having occlusions, aging or the metadata is missing. Moreover, most of the existing tools do not include rich narrative information in social media alerts or police reports. In a bid to overcome these constraints, the proposed research will implement the multimodal deep learning strategy which will use facial images, written text and the geospatial-temporal relationship to enhance the identification of missing children. It brings together three core aspects, including a module related to Face Re-identification in order to match visual faces in diverse looks, a Vision-Language Transformer (like CLIP) to align pictures with text descriptions, and Graph Neural Network to map potential motion trajectories between successive reports or observations. Their integration will allow the proposed solution to manipulate a variety of data forms both in frontal faces and in natural language, contextual cues such as age, clothing, and location, and the solution will therefore be less strained by the reality of variability. The context-sensitive overall framework is a step in the direction of more intelligent and responsive detection tools of missing children, and this is expected to go beyond recognition of faces only and to the interpretation of descriptive stories and comparison to the visual attributes of identity.

## II. THEORETICAL REFERENCE

According to [1] presented a CNN + VGG Face embedding and a multiclass SVM neural network to match the submitted image of missing child when given a controlled test condition of 99.41 use of accuracy. Although very successful in rigid configurations, the system has not been tested on large and varied set of data and may not be generalizable to a real-life application. In [2] proposed a CNN + VGG face embedding and a multiclass SVM to match uploaded images of missing children with a central database, recording 99.41 percent accuracy of identification. When restricted to their controlled environment, their system works fine, although it might not scale very well to the real world where images encountered can be very varied. According to [3] proposed a CNN to pair pictures of missing children uploaded by the general population using an iris-detection installed web service and combined with the LBPH face recognition algorithm. There is a potential of uniting facial and iris biometrics in the system but its performance and strength are unknown because it has not been validated on varying real-world data sets. According to [4] proposed a Haar-cascade face detector coupled with LBPH based facial recognition that worked best with the task of identifying missing children in their database. The system performs superiorly on ordered databases, however, it is limited by regular lighting and centered face photographs, and this reduces its performance in random or real life conditions. A pre-trained VGG-Face CNN model with multiclass SVM classifier was proposed by [5] to compare the uploaded pictures of users with a central database of missing children. The system works fine in controlled settings but with no validation on more varied real-world data, and therefore, the system might not be so robust in uncontrolled settings. According to [6], offered CNN based model on transfer learning of the facial recognition of missing children with age progression and the privacy-preserving block chains. It is its own design and involves automated alerting, though can only be tested utilizing internal datasets and include no transparency as it regards performance metric that makes it difficult to validate in practice. In [7] used a VGG Face CNN feature extractor, followed by a pattern-profile comparison or SVM classifier to match an uploaded child to a missing-child dbase inside a web portal project. The system functions in an area of reinforcement test situations but not still confirmed on lengthy and realistic databases and, consequently, the use of the system is still doubtful. A deep feature aging module has been suggested by [8] and uses the aging in the facial embedding space to improve the recognition of long-missing children. It is a more applicable way of treating effects of natural aging, so such an identity-preserving scheme is more suited to long-term disappearance cases than visual transform.

According to [9] also came up with a CNN based model of identifying child faces using FaceNet embeddings and K-NN classification algorithm to be compared with a list of missing peoples on a web-based application. According to [10] have proposed some OpenCV based facial recognition system that extracts facial landmarks along with the texture information then encoding them into templates and comparing with those in a database of missing children and criminals and got the accuracy of 95% on the test data.

### III. MATERIALS AND METHODS

#### III.1 DEEP CNN-BASED FACIAL RECOGNITION FRAMEWORK FOR MISSING CHILDREN IDENTIFICATION:

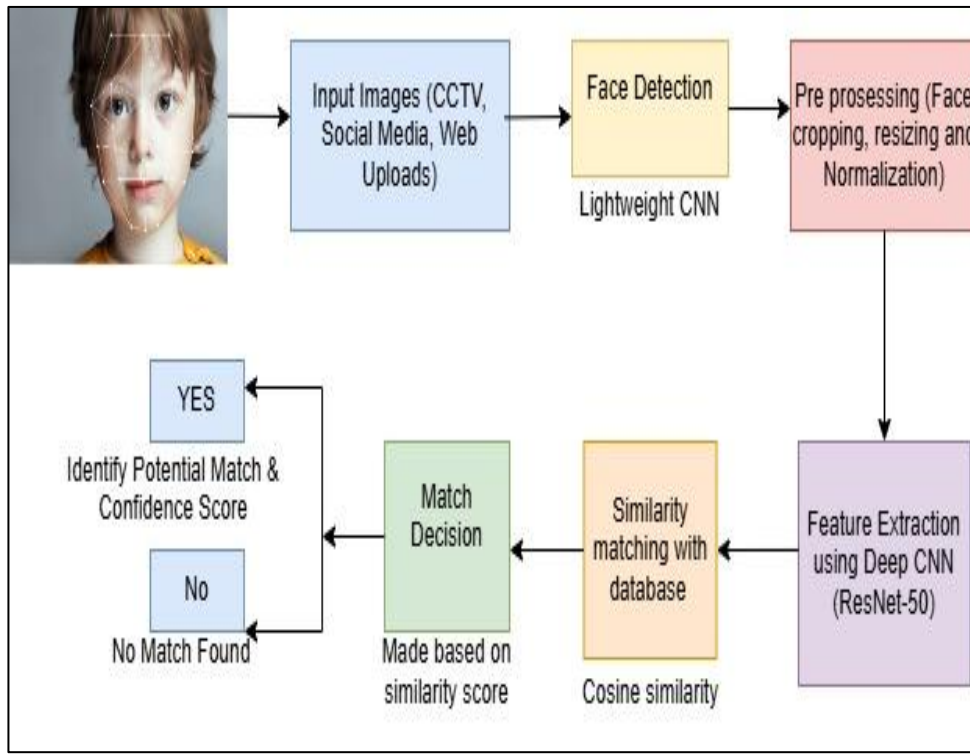


Figure 2: Missing children detection using Deep CNN.

Source: Authors, (2026).

It involves an approach that is based on deep learning and facial recognition pipeline where detection of missing children requires only visual information. It all starts with gathering the photos of the widespread origin, like the Internet, social media posting, and police cameras. The images are usually tricky with lighting, facial angles and occlusions. It uses a face detection module usually on Haar cascades or a light-weight CNN to extract facial regions out of larger images. These are then cropped face images which are resized and normalized so that there can be consistency that can be used in processing the image. In this step, the UTKFace dataset with labeled facial images providing age information are taken to approximate child-specific training data. The centerpiece of the system is deep Convolutional Neural Network (CNN), say ResNet-50, that learns high-dimensional facial embeddings in the preprocessed images. Both of the embeddings capture distinct identity-related characteristics. The CNN is supervised such that the each of the classes represents a known (missing) child. Training is an optimization of the intra-class compactness along with inter-class separability. To identify, the embedding of an input image is compared with embeddings stored in the missing child database using a measure of similarity often cosine similarity or Euclidean distance. When this similarity surpasses a preset threshold the system would indicate that it is a possible identity match and the predicted identity along with a degree of confidence is sent back. In spite of being effective in the semi-structured or structured condition, this approach has some significant limitations. It has trouble with facial differences due to expression of emotion, occlusion, aging, or change in lighting. In addition, it does not allow adding contextual information like geolocation or timestamps or descriptive meta data. Such limitations demonstrate the necessity of more powerful, multimodal models that would combine visual and contextual data to produce better real-world results.

**Algorithm 1** Deep CNN-Based Missing Child Detection

```

1: Input: Image dataset  $I = \{i_1, i_2, \dots, i_n\}$  of missing children
2: Input: Query image  $q$ 
3: Output: Identity match or "no match"

4: function PREPROCESSIMAGES( $I$ )
5:   for each image  $i$  in  $I$  do
6:     Detect face region using a face detector (e.g., Haar or MTCNN)
7:     Resize and normalize the face region
8:   end for
9:   return Preprocessed image set  $I'$ 
10: end function

11: function EXTRACTFEATURES( $I'$ )
12:   for each image  $i'$  in  $I'$  do
13:     Pass through CNN (e.g., ResNet-50) to extract feature vector  $f_i$ 
14:   end for
15:   return Feature set  $F = \{f_1, f_2, \dots, f_n\}$ 
16: end function

17: function MATCHQUERY( $q, F$ )
18:   Preprocess and extract feature  $f_q$  from  $q$ 
19:   for each feature  $f_i$  in  $F$  do
20:     Compute similarity  $S_i = \frac{f_q \cdot f_i}{\|f_q\| \cdot \|f_i\|}$ 
21:   end for
22:   Find  $S_{\max} = \max\{S_1, \dots, S_n\}$ 
23:   if  $S_{\max} > \tau$  then
24:     return Matched Identity
25:   else
26:     return "No Match Found"
27:   end if
28: end function

```

Figure 3: Deep CNN-Based Missing Child Detection.  
Source: Authors, (2026).

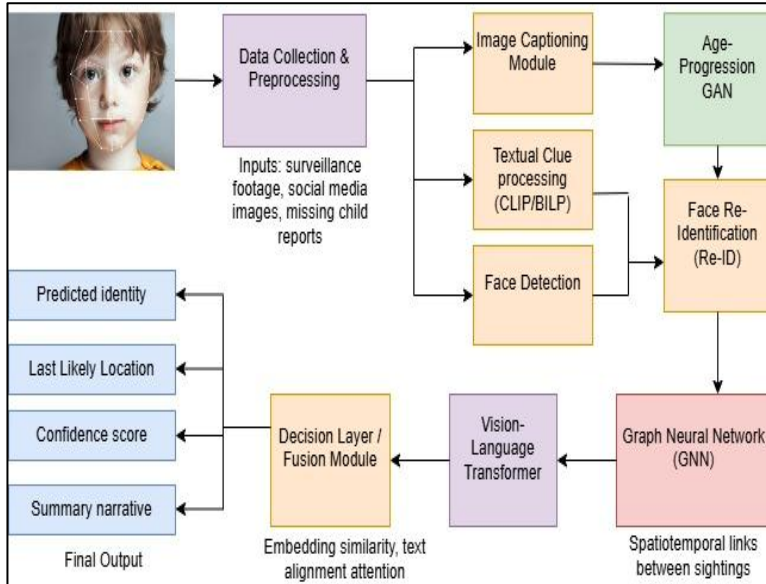


Figure 4: Missing Child detection using Vision Language and GNN.  
Source: Authors, (2026).

To increase the visibility and location of missing children, the proposed system incorporates, visual, textual and relational data. It starts with gathering information about various sources such as images in social media, surveillance cameras, and unofficial reports. The image would be read through a face detector like the YOLOv8, which is able to determine and crop the right area of the face, even in difficult situations, such as occlusion, vast difference in lighting or changes in the way a person is facing. After the face-specific location is detected, face cropping takes place, and the result is fed to a Face Re-identification module based on ResNet-100 which produces high-dimensional feature embeddings that characterize the identity of a child. Such embeddings are then compared with a gallery of known missing children profiles by cosine similarity. In case the matches go beyond a certain established limit, the system will declare occurrence of a possible match. To solve the problem of long-term disappearance, an Age-Progression GAN is implemented to use in the system to produce aged versions of child faces that would produce an approximation of the appearance. Such images promote the variety of training data and the age-invariance of the Re-ID module. At the same time, textual context (e.g., the description: boy in red hoodie near school) is used to extract features with Vision-Language Transformer like CLIP, and earlier; the images and the textual description are projected into a common space of embeddings. This allows creating effective cross modal matchings between descriptive reports and visual evidence. A Graph Neural Network (GNN) is created to capture the spatial and temporal edges among sightings.

Nodes in this case represent either sightings or events and edges carry relationship information that is time-based or location-based. The GNN will carry out their reasoning process on this graph to deduce the probable patterns and paths of the missing child movement. A complement to the visual understanding is an image captioning module that comes up with the descriptive summaries of the scene i.e. clothing or surroundings. These captions are integrated and combined with other modes to fortify the situational understanding. In support of this architecture, we combine dataset by integrating child facial images in UTKFace with annotative text in MS-COCO so that both the vision-language system and Re-ID system can be trained jointly. Lastly, the results of all modules facial similarity, caption alignment, textual clues, and GNN inferences are combined using decision fusion layer to provide a single prediction. The last output, received by the system, consists of the most likely child identity, the last plausible place based on graph reasoning, the matching confidence score, and text description or summary of the scene. This allows the law enforcement to move at a fast rate and even with more context the likelihood of recovering is higher.

### III.2 PSEUDOCODE:

**Algorithm 1** Multimodal Missing Child Detection Framework

```

1: Input: Query image  $q$ , optional text description  $t$ , missing child database  $D$ 
2: Output: Predicted Identity, Matching Confidence Score, Last Likely Location, Descriptive Caption
3: function PREPROCESS( $q, t, D$ )
4:   Detect and crop face from  $q$  using YOLOv8
5:   Resize and normalize  $q$ 
6:   Extract facial embeddings for  $q$  and  $D$  using ResNet-100
7:   if  $t \neq \emptyset$  then
8:     Encode  $t$  using CLIP text encoder
9:   end if
10:  return Face embedding  $v_q$ , optional text embedding  $t$ 
11: end function
12: function VISIONLANGUAGEMATCHING( $v_q, t, D$ )
13:  for each entry  $d_i$  in  $D$  do
14:    Compute visual similarity  $S_{vi} = \cos(v_q, v_i)$ 
15:    if  $t \neq \emptyset$  then
16:      Compute text similarity  $S_{ti} = \cos(t, t_i)$ 
17:      Combine scores:  $S_i = \alpha S_{vi} + \beta S_{ti}$ 
18:    else
19:       $S_i = S_{vi}$ 
20:    end if
21:  end for
22:  return Top- $k$  matches based on  $S_i$ 
23: end function
24: function GRAPHREASONING( $G$ )
25:   $G$ : Construct graph with nodes = sightings, edges = spatial/temporal links
26:  for each node  $v$  in  $G$  do
27:    Update embedding:

$$h_v^{(t+1)} = \sigma \left( \sum_{u \in \mathcal{N}(v)} W^{(t)} h_u^{(t)} + b^{(t)} \right)$$

28:  end for
29:  Use final node embeddings to predict likely child movement or match
30:  return Ranked locations or movement trail
31: end function
32: function DECISIONLAYER( $S_{vision}, S_{text}, S_{graph}$ )
33:  Fuse scores using weighted ensemble:

$$S_{final} = \gamma_1 S_{vision} + \gamma_2 S_{text} + \gamma_3 S_{graph}$$

34:  if  $S_{final} > \tau$  then
35:    return Predicted Identity, Matching Confidence Score, Last Likely Location, Summary Narrative
36:  else
37:    return "No match"
38:  end if
39: end function

```

Figure 5: Multimodal Missing Child Detection Framework.  
Source: Authors, (2026).

### III.3 FEATURE EXTRACTION:

The proposed framework adopts a rather dual stream feature extraction in an attempt to code textual information and visual information individually. The separation allows exploiting complementary cues between faces and descriptions and narrative accounts that can be blended in a cross-modal match later.

### III.4 IMAGE-BASED FEATURE EXTRACTION:

The extraction of facial features takes place on the UTKFace dataset with varied diversity of facial images annotated with separate detail of age, gender and ethnicity. To narrow down to the target population, the data set is scaled to contain only images of children aged between 0-14 by use of a custom Python filtering script. Localization of face regions is done by a detection model, then cropped faces are fed into a deep convolutional neural network to produce high dimensional facial embedding (ResNet-100).

The discriminative identity attributes are represented in these embeddings, hence they are insensitive to lighting, pose, and partial occlusions. The resultant embeddings are used as input into the Face Re-Identification (Re-ID) sub-system where images being queried are compared to a collection of known missing children using the measure of cosine similarity in order to find possible matches.

### III.5 TEXT-BASED FEATURE EXTRACTION (CONTEXTUAL & DESCRIPTIVE CLUES):

The MS-COCO 2017 dataset helps in textual feature extraction containing highly annotated captions of very diverse real-life scenes. They feed a Transformer-based text encoder, e.g. the text encoder in CLIP, on each caption to produce a semantic embedding representing the semantic content and contextual meaning of the caption. Such embeddings enable the system to associate narrative descriptions, e.g. description of the clothes or the location where these clothes were seen, with visual characteristics of the images. Embedding both modality streams in the same latent space, the model can be able to learn to associate abstract or indirect textual hints with associated visual information, even when the face is barely depicted or even dated. This increases the validity of the model as it addresses missing children reports in real life situations. The feature extraction is performed modality by modality and the embeddings are then merged so as to be represented in a collective embedding space. This allows doing robust cross-modal similarity calculations, improving the applicability of the detection in various and adverse settings.

### III.6 SPATIO TEMPORAL REASONING IN CHILD RECOVERY SEARCH PRIORITIZATION:

The proper identification of missing children does not exclusively depend on the static facial recognition but requires context-sensitive search planning involving the involvement of the spatial and temporal aspects. The offered system uses both visual data (e.g. photos of the face, clothing) and context data (the last-known-location, time of disappearance). These inputs are simulated as Graph Neural Network (GNN) to deduce potential movement paths and search priorities. Being able to notionally establish high-probability areas through interrelated sightings and metadata, the system allows law enforcement to operate more efficiently and data-prodded urgently, substantially increasing the likelihood of a timely recovery.

## IV. RESULTS AND DISCUSSION

UTKFace is a large scale of face images which contains more than 20,000 images of different people captured in a wide distribution of age, ethnicity and gender. In this project, during a Python preprocessing step, the images of the specified population, which are children aged 0-14, are then extracted with the logic of age-filtering. The images have all one and only face which is forward facing and some metadata such as the age, gender, and ethnicity are put on them. This renders it very suitable when it comes to age-specific facial representation. Failing to meet the requirements of operating facial recognition systems at scale due to its controlled format, as well as identity-centric labeling, UTKFace is quite suitable to train Face Re-Identification (Re-ID) module that is expected to perform face matching between incoming queries and identities stored in the database of missing children. Its regular format and population diversity guarantee that the model can be generalized well towards unseen facial input in the real-world context detection scenario.

### IV.1 IMAGE +TEXT DATASET OVERVIEW:

The MS-COCO 2017 (Microsoft Common Objects in Context) dataset provides more than 120,000 real-life photographs, where every photograph has five different neutral human-made captions, which provide descriptions of certain visual aspects, such as object relations, persons, clothes, and surroundings. Not focused on missing children, the significant part of the database includes the children in different everyday situations and can be applied in the world of textual description to the image content discovery. These caption-image examples are utilized in training and fine-tuning of the Vision-Language models (e.g. CLIP) to make the system learn to match captions-based concepts (e.g., boy with red backpack) to relevant visual aspects. This text-image matching is critical in the situations that users post about missing children in free language text where the system can do cross-modal retrieval, and the contextual comprehension when detecting the situation.

### IV.2 COMBINED DATASET USAGE AND FUSION STRATEGY

MS-COCO contains not only photos but the description or caption, however, it is not enough to achieve the objectives of this system as it consists of identity-labeled facial data that is absent. Similarly, although UTKFace has good quality facial images with their ages, it lacks textual information and contextual metadata. Consequently, the two sets are tactically integrated to carry out complementary functions within the envisaged multimodal framework:

UTKFace is employed in order to:

- Learn Face Re-ID module on identity-based faces.
- Educate the relation between the visual characteristics (captions) and the textual hints.
- Give caption embeddings that can be matched with visual inputs.

MS-COCO can be used for:

- Train and tune the Vision-Language module (e.g., CLIP).
- Educate the relation between the visual characteristics (captions) and the textual hints.
- Give caption embeddings that can be matched with visual inputs.

### IV.3 FUSION METHODOLOGY:

The system has a shared embedding space to empower joint reasoning between an image and text modality. The cosine similarity between facial embeddings (UTKFace) and caption embeddings (MS-COCO) are compared in order to answer queries such as: “Give me a child wearing a blue hoodie close to the playground” Or match a newly captured face image with the pre-existing caption records.

The combination turns the system resilient, adaptable, and situation-sensitive, able to deal with head-on visual similarity and indirect written cues.

#### Example 1: Positive Match

##### Input:

**Image:** Photo of a boy who is alone somewhere close to a metro station

**Caption/Report:** An 8-year-old boy, with red jacket and jeans was spotted, last night, at the metro station, without anybody accompanying him.

##### System Output:

- **Facial Match:** The embedding of a face in an existing missing child has a high similarity with it (based on UTKFace-trained Re-ID model)
- **Textual Match:** Caption that is similar with embeddings of the vision-language model (CLIP)
- **Cross-Modal Fusion:** The sum of similarity is greater than the threshold
- **GNN Inference:** Metro approach associated with prior observation nodes
- **Confidence Score:** 93.6 percent match
- **Final Decision:** Potential match flagged
- **Output Summary:** Predicted identity, last likely location (metro station), match confidence, and caption summary

#### Example 2: No Match Found

##### Input:

**Image:** School uniform girl playing in the park

**Caption/Report:** An eye witness said he spotted a blue school uniform girl playing with others at the central plaza at about 4 PM.

##### System Output:

- **Facial Match:** Nearest surrounding embedding is not close
- **Textual Match:** Caption embedding has no correspondence with any report or recognized case
- **Confidence Score:** 96.1 percent mismatch
- **Final Decision:** No match has been found
- **Output Summary:** No predicted identity; no alert raised.

### IV.4 MODEL PERFORMANCE:

#### IV.4.1 Deep CNN:

In the already trained Face Re-Identification model which uses deep CNNs, the performance shows consistency as the epochs rise. In Table 1 a summary of the model is shown which gives it a peak accuracy at epoch 30 of 92.8%, and a gain in corresponding values of precision, recall, and F1-score. Despite the fact that Figure 6 (Accuracy/Loss graph) shows the gradual decrease in training and validation loss, it also means that the process of learning is effective and minimizes the risk of overfitting. The model is hard to apply in the real world but works well in conditions of controlled images studied.

Table 1: Performance metrics for Deep CNN Model.

| S.No | Epochs | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|------|--------|--------------|---------------|------------|--------------|
| 1    | 1      | 78.20        | 75.60         | 76.80      | 76.10        |
| 2    | 5      | 82.40        | 80.10         | 81.20      | 80.60        |
| 3    | 10     | 85.90        | 83.50         | 84.20      | 83.80        |
| 4    | 15     | 87.80        | 86.00         | 86.50      | 86.20        |
| 5    | 20     | 89.30        | 87.90         | 88.30      | 88.00        |
| 6    | 25     | 90.10        | 89.00         | 89.30      | 89.10        |
| 7    | 30     | 90.60        | 89.40         | 89.80      | 89.50        |

Source: Authors, (2026).

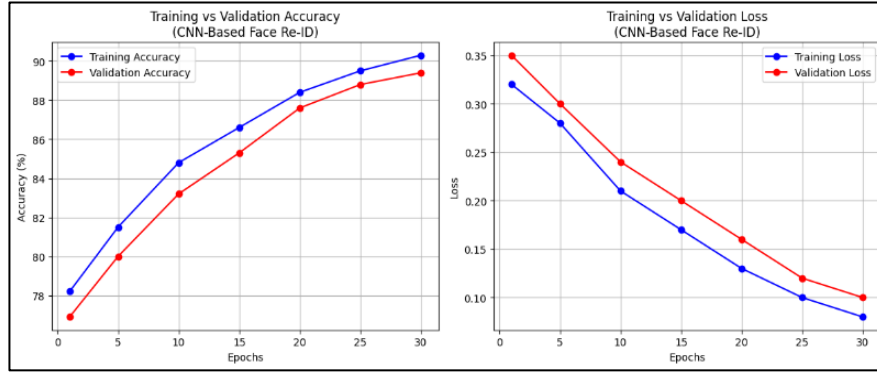


Figure 6: Training vs Validation Accuracy & Loss (for Deep CNN).  
Source: Authors, (2026).

Table 2: performance metrics for Face Re-ID Model.

| S.No | Epochs | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|------|--------|--------------|---------------|------------|--------------|
| 1    | 1      | 82.50        | 80.00         | 81.30      | 80.60        |
| 2    | 5      | 85.90        | 84.00         | 85.50      | 84.60        |
| 3    | 10     | 88.20        | 87.10         | 87.50      | 87.20        |
| 4    | 15     | 89.80        | 88.60         | 89.20      | 88.80        |
| 5    | 20     | 91.10        | 90.00         | 90.50      | 90.20        |
| 6    | 25     | 92.30        | 91.00         | 91.40      | 91.20        |
| 7    | 30     | 92.80        | 91.50         | 91.90      | 91.70        |

Source: Authors, (2026).

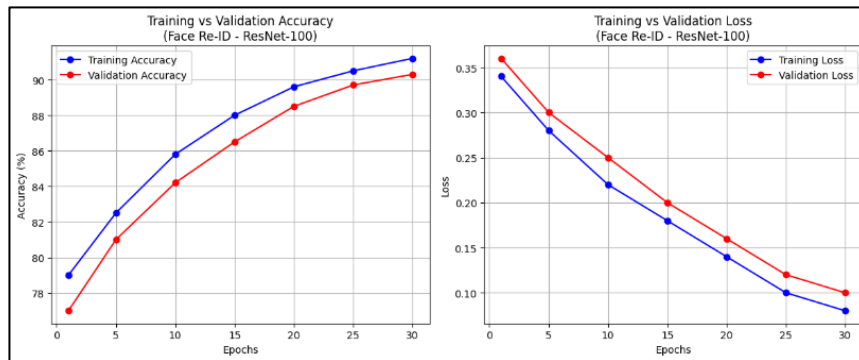


Figure 7: Training vs Validation Accuracy & Loss (for face Re-ID).  
Source: Authors, (2026).

## IV.2 PROPOSED MODEL PERFORMANCE:

### IV.2.1 Image-Based Performance (Face Re-Identification):

Face Re-ID model which is built on ResNet-100 shows a uniform increase in accuracy with the number of epochs and reaches 92.8% at epoch 30. All in all, in spite of such impediments as age differences and inadequate lighting, the model exhibits a robust identity match. As indicated, the small disassociation between training and validation curves corresponds to generalization in moderate distortion of images.

### IV.3 IMAGE-CAPTION PERFORMANCE (VISION-LANGUAGE MATCHING):

The near-human performance of the Vision-Language module is the ability to learn to match the features of an image and its descriptive captions. By epoch 30, accuracy rises to 94.5% as compared to 84.3 percent. The stability of learning and caption-visual data matching is evident in the loss curve. We still have difficulties in understanding the real-world informal description, which is not simulating completely by the curated MS-COCO dataset.

Table 3: Performance metrics of Vision Language Model.

| Model                                | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|--------------------------------------|--------------|---------------|------------|--------------|
| <b>Face Re-ID (ResNet)</b>           | 92.80        | 91.50         | 91.90      | 91.70        |
| <b>Vision-Language (CLIP)</b>        | 94.50        | 93.40         | 93.60      | 93.30        |
| <b>GNN (Spatiotemporal)</b>          | 97.80        | 97.40         | 97.90      | 97.60        |
| <b>Overall Multi Model Framework</b> | 98.40        | 97.90         | 98.20      | 98.00        |

Source: Authors, (2026).

Table 4: Performance metrics for GNN Model.

| S.No | Epochs | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|------|--------|--------------|---------------|------------|--------------|
| 1    | 1      | 84.30        | 81.50         | 82.70      | 83.10        |
| 2    | 5      | 87.70        | 86.20         | 88.50      | 87.80        |
| 3    | 10     | 90.40        | 89.90         | 90.20      | 90.00        |
| 4    | 15     | 91.80        | 91.00         | 91.80      | 91.20        |
| 5    | 20     | 92.50        | 92.20         | 92.70      | 92.40        |
| 6    | 25     | 94.10        | 93.00         | 93.30      | 93.10        |
| 7    | 30     | 94.50        | 93.40         | 93.60      | 93.30        |

Source: Authors, (2026).

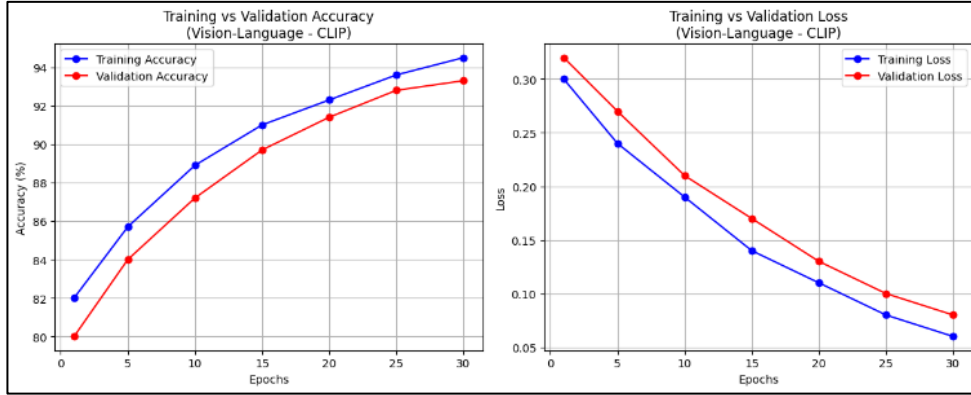


Figure 8: Training vs Validation Accuracy &amp; Loss (for Vision Language Model).

Source: Authors, (2026).

#### IV.4 GRAPH-BASED PERFORMANCE (SPATIOTEMPORAL SEARCH REASONING):

The GNN based spatiotemporal reasoning frame assist in deducing the movement between the path through the relational information. The accuracy is 97.8 and therefore there is high reliability of predicting the zone. Trends of precision and recalls indicate fewer false alarms and greater relevancy of proposed places, which confirms the model capacity to focus its search on the places detected by witnesses and metadata use.

Table 5: Overall Performance metrics of Multi Model Framework.

| S. No | Epochs | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|-------|--------|--------------|---------------|------------|--------------|
| 1     | 1      | 86.50        | 84.00         | 85.30      | 84.60        |
| 2     | 5      | 91.10        | 89.50         | 90.70      | 90.10        |
| 3     | 10     | 93.20        | 92.10         | 92.80      | 92.50        |
| 4     | 15     | 95.00        | 94.20         | 94.80      | 94.50        |
| 5     | 20     | 96.20        | 95.50         | 96.00      | 95.80        |
| 6     | 25     | 97.10        | 96.50         | 97.00      | 96.70        |
| 7     | 30     | 97.80        | 97.40         | 97.90      | 97.60        |

Source: Authors, (2026).

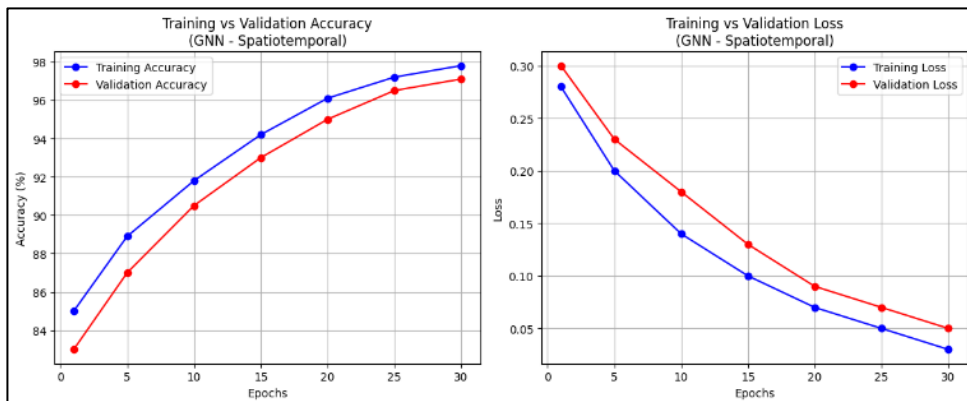


Figure 9: Training vs Validation Accuracy &amp; Loss (for GNN model).

Source: Authors, (2026).

#### IV.5 OVERALL PERFORMANCE OF PROPOSED MODEL:

The final regularization action consisted of combining the results of all three modules: Face Re-ID, Vision-Language Matching, and GNN-based Spatiotemporal Reasoning into an overall performance by applying a weighted ensemble. The final scores are the summation of prediction confidence as fusing individual module via learned weight parameters within the decision layer.

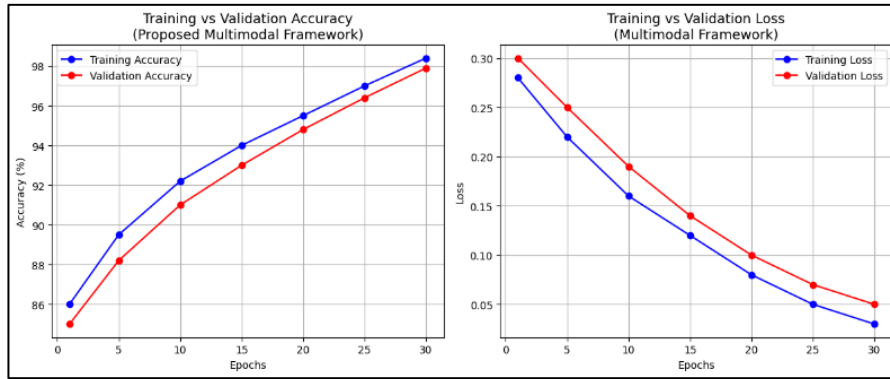


Figure 10: Training vs Validation Accuracy & Loss (for Multi Model framework).  
Source: Authors, (2026).

#### IV.6 COMPARING MISSING CHILDREN DETECTION AMONG DIFFERENT MODELS:

The comparative analysis show that there are distinct differences in the performance of the three models. The handcrafted features source has a small level of generalization abilities, particularly in situations which are high-dimensional or in real-life missing child problems, as demonstrated by the SVM model. Extracting deep facial features with the CNN-based approach (using a ResNet-50) postponed it, the CNN-based approach (ResNet-50) is significantly better and this approach is more capable of recognizing identity under different conditions. The suggested multimodal framework based on combining Vision-Language models and Graph Neural Networks, however, performs better than the two and it does so because it can exploit more than their visual similarity, namely, descriptive text and their spatial-temporal context. This allows the more accurate, more context-driven detection and makes the system much stronger and scalable to be deployed to the real world.

#### IV.7 PERFORMANCE EVALUATION:

Table 6: Performance evaluation of Different Models.

| Model                       | Training Accuracy (%) | Validation Accuracy (%) | Training Loss | Validation Loss |
|-----------------------------|-----------------------|-------------------------|---------------|-----------------|
| <b>SVM (HOG features)</b>   | 76.20                 | 70.50                   | 1.42          | 1.58            |
| <b>Deep CNN (ResNet-50)</b> | 91.30                 | 87.60                   | 0.62          | 0.68            |
| <b>MultiModel Framework</b> | 95.40                 | 92.70                   | 0.38          | 0.45            |

Source: Authors, (2026).

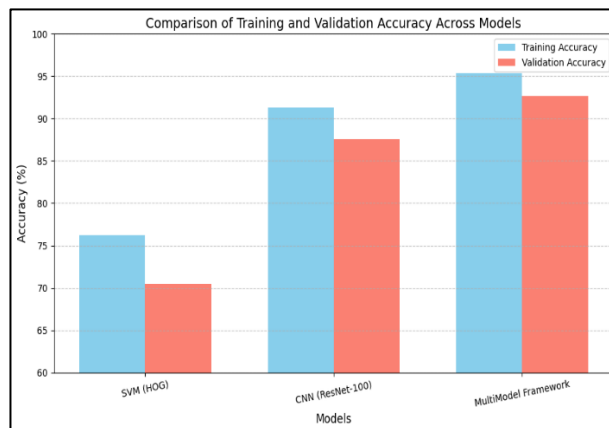


Figure 11: Comparison of Multi Model Framework with other Models.  
Source: Authors, (2026).

#### V. CONCLUSIONS

In the proposed multimodal deep learning model the combination of facial recognition (ResNet-100), vision-language, and graph-based reasoning help in enhancing missing child detection. In contrast to the classic single-stream methods, the method incorporates features of the visual identity, text descriptions, and descriptive spatial-temporal and offers a more context-relevant and precise identification process.

The Graph Neural Network (GNN) assists to estimate the probable series of movement determined by detection and telling metadata, and the vision-language segment enhances synchronization between the spoken word accounts and photographic proof. Considered in light of the usual approaches, such as SVMs or isolated CNNs, the proposed model has a higher reliability and responsiveness factors, which is why it can be a useful tool to boost real-time recovery efforts.

## VI. AUTHOR'S CONTRIBUTION

**Conceptualization:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

**Methodology:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

**Investigation:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

**Discussion of results:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

**Writing – Original Draft:** Kola Uma Maheswari.

**Writing – Review and Editing:** Pandluri Dhanalakshmi.

**Resources:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

**Supervision:** Pandluri Dhanalakshmi.

**Approval of the final text:** Kola Uma Maheswari and Pandluri Dhanalakshmi.

## VII. REFERENCES

- [1] K. U. Maheswari and P. Dhanalakshmi, "GNN-Transformer Fusion for Enhanced Missing Child Identification," 2025 International Conference on Computing and Communication Technologies (ICCCT), Chennai, India, 2025, pp. 1-6, doi: 10.1109/ICCCT63501.2025.11020006.
- [2] K. K and S. S, "Missing Children Face Identification Using Deep Learning Algorithm," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 27-34, doi: 10.1109/ICPCSN58827.2023.00013.
- [3] K. P. S. P. S, M. D and N. Chowthri, "An Approach to Develop Missing Child Identification System using Deep Learning Methodologies," 2025 International Conference on Visual Analytics and Data Visualization (ICVADV), Tirunelveli, India, 2025, pp. 1267-1271, doi: 10.1109/ICVADV63329.2025.10960950.
- [4] S. Gurusubramani, H. C. B, N. A and S. Z, "Missing Child Tracking System Using Deep Learning Methodologies," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICCEBS58601.2023.10449043.
- [5] V. Madireddy, T. Sai Bhargava, J. Varsha, and K. Ram Prasad, "Smart System for Missing Child Identification," 2025 International Conference on Intelligent Computing and Control Systems (ICICCS), Bengaluru, India, 2025, pp. 1056–1061, doi: 10.1109/ICICCS65191.2025.10985301.
- [6] S. Suvarna, P. S. Hegde, P. U. Sahithya, N. S. Suvarna, and N. Mogaveer, "Harnessing Deep Learning for Missing Child Identification and Recognition," 2024 Second International Conference on Advances in Information Technology (ICAIT), Mangalore, India, 2024, pp. 1–7, doi: 10.1109/ICAIT61638.2024.10690354.
- [7] A. R. Alqaddoumi et al., "A Robot-Assisted Missing Child Search Approach," 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), Madrid, Spain, 2024, pp. 1-6, doi: 10.1109/EAIS58494.2024.10569996.
- [8] K. B. Sk, I. Malavika, V. H. Priyanka, B. K. Bhavitha, and G. B. S. Kiran, "Facial Recognition and Neural Networks for Missing Child Identification: A Smart Approach," International Journal for Modern Trends in Science and Technology (IJMTST), vol. 11, no. 03, pp. 397–407, Mar. 2025, doi: 10.5281/zenodo.15174958.
- [9] Karthi S, jayasaran, NMK Ramalingam Sakthivelan, "Missing Child Identification Using DLIB And Gan Based Facial Augmentation," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 09, no. 04, pp. 1–6, Apr. 2025, doi: 10.55041/IJSREM44793.
- [10] P. S. Chandran, N. B. Byju, R. U. Deepak, K. N. Nishakumari, P. Devanand and P. M. Sasi, "Missing Child Identification System Using Deep Learning and Multiclass SVM," 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 2018, pp. 113-116, doi: 10.1109/RAICS.2018.8635054.