Advances in Artificial Intelligence and Machine Learning; Research 5 (2) 3736-3754 Received 28-02-2025; Accepted 14-05-2025; Published 21-05-2025

Enhancing Lost and Found Systems with Multi-Modal Deep Learning: Integrating SBERT and Siamese Networks for Improved Semantic Matching

B.M.P. Dhanawardhana 24piumal@gmail.com
Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

K.A.D. Chalana dishanchalana1999@gmai.com
Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

I.D.S.P. Abeywardena sachithrapiumal18@gmail.com
Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

Nalaka Lankasena nalaka@sjp.ac.lk
Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

M.H. Paul hansamalipaul@sjp.ac.lk
Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

Corresponding Author: Nalaka Lankasena

Abstract

Returning lost and found items in public spaces is challenging with traditional methods, and while technological advancements have led to systematic approaches, they often rely on query-based searches or image classification. This research provides a solution that combines textual and visual data to improve the semantic matching of lost and found items to address these problems. Three deep learning models for image similarity, text similarity, and fusion are implemented in a progressive web application (PWA) to support user data input and matching alerts. A fusion model was created by combining the SBERT model, which was refined using a dataset of 2,600 lost and found description pairs both for English and Sinhala languages, and the Siamese network, which was trained on 848 bag images using MobileNetV2. This fusion model also incorporates location and time features to give priority to recent activities and places to enhance matching accuracy. A neural network

Citation: B.M.P. Dhanawardhana, et al. Enhancing Lost and Found Systems with Multi-Modal Deep Learning: Integrating SBERT and Siamese Networks for Improved Semantic Matching. Advances in Artificial Intelligence and Machine Learning. 2025;5(2):212.

was trained using the dataset for the fusion model, which included image similarity, text similarity, location similarity and time similarity features as well as a target column that represents the similarity level of the two given bags. The accuracy of the Siamese model was 0.75, whereas the SBERT model demonstrated an accuracy of 0.9526 and an F1 score of 0.9405. The fusion model, which combined text and image data, achieved an accuracy of 0.87 and an F1 score of 0.98. The developed web application offers a community-driven platform to assist users in locating misplaced items, showcasing the system's practical usefulness.

Keywords: Deep learning, Siamese network, Transformers, Data fusion, Cosine similarity.

1. INTRODUCTION

Personal belongings are often misplaced by people, making recovery a challenge. Over time, the emergence of digital platforms and the integration of cutting-edge technology have significantly impacted various sectors, including the recovery of lost and found items. This transformation encompasses the use of AI-driven image recognition, GPS tracking, and centralized digital databases, enabling owners to search for their lost belongings across wide geographic regions. These technologies have made tracking down personal items much more efficient than traditional methods. Older methods, like putting up posters or writing notes, were less effective. Social media pages help by reaching a larger audience. However, they have also shown how unsecured traditional recovery systems can be. These older systems tend to be very slow. The tools they use are not precise enough to properly match lost items with their rightful owners. They usually rely on just a few types of data, such as text descriptions or attached pictures. In this paper, we propose a new method in multimodal deep learning to present semantic matching that can benefit tasks regarding recoverability for accuracy as well as efficiency. Deep learning and multimodal data processing have created new avenues for enhancing the precision and efficacy of lost-and-found systems. Incorporating semantic analysis and visual comparison architectures into advanced neural network architectures would be a very promising solution to the kinds of requirements that are being explored in this study. This study introduces a novel approach in the field of multimodal deep learning for optimizing semantic matching to improve accuracy and efficiency in tasks related to recovering lost and found items.

2. RELATED WORK

2.1 Lost and Found Systems

Various research have established systems in recent years to streamline the process of locating lost items [1–3]. These systems feature searchable databases where users are allowed to enter missing item descriptions and photos. The main distinction is that all these systems are limited by their reliance on query-based database searches for matching lost and found pairs. Suryani and Edy (2020) [4], developed an Android application called "Lost & Found" and used the Term Frequency-Inverse Document Frequency (TF-IDF) in order to weigh words that are significant for item descriptors and then employed cosine similarity to assess the similarity between these descriptors. Thus, the effectiveness of their strategy, which they implemented with an accuracy

of close to 88% and an error of 12%, was adequate to help them achieve the intended objectives. However, the weakness of relying on textual description is that different people may describe the same item in different words. Further, the proposed approach was not very reliable as there could be variations in the areas of comparison. Moreover, the software is compatible with only Android devices, and it lacks image comparison capability. A more advanced approach is described in the paper of Zhou et al. (2023) [5], "LostNet" which employs MobileNetv2 in conjunction with CBAM (Convolutional Block Attention Module) for enhancing the identification of lost items through a comparative analysis between user-submitted images and those stored within the system. Thus, this method was able to attain a testing accuracy of 96.8%. Nevertheless, due to the high accuracy of this approach, "LostNet" is designed to work only on laptops, which also restricts its application.

Prawira and Saputri (2023) [6], proposed a lost object identification version that integrates photo contrast using ResNet-50 and Natural Language Processing (NLP) for string matching. The model employs Euclidean distance for image similarity and cosine similarity for textual similarity, achieving 29.96% accuracy in image comparison and 97.92% in string matching. A notable feature is the use of background removal to handle varied image backgrounds. However, limitations include low image similarity accuracy, limited contextual feature integration (location and time), reliance on a small dataset, and exclusivity to iOS devices. The system uses threshold values for both Euclidean and cosine similarities to determine matches, potentially leading to suboptimal accuracy by not fully leveraging multimodal data's complementary nature.

2.2 Sentence Matching Algorithms

The development of sentence matching algorithms has advanced significantly, beginning with keyword-based methods such as Term Frequency and Inverse Document Frequency (TF-IDF) for information retrieval that focused on the presence and frequency of words to determine similarity [7]. These early methods struggled with understanding context and semantics. Advances came with models such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), which analyzed word distributions to better capture semantics [8, 9]. The introduction of DL produced models such as Word2Vec and GloVe, which created word embedding to capture semantic relationships [10, 11]. InferSent and the Universal Sentence Encoder are more advanced than word-level embedding in matching as they represent the context and relationships within sentences [12, 13]. Recent developments have been made along the lines of transformer-based models like BERT, RoBERTa, and SBERT which are responsible for learning the dependencies of sentences using attentional mechanisms that significantly impact the efficiency and reliability of sentence matching [14–16]. These advances make modern sentence-matching algorithms such as SBERT more suitable for understanding the context of user input in lost and found systems and for accurately matching the details of lost and found items.

2.3 Image Similarity Matching Algorithms

Image similarity matching algorithms have been improved from simple techniques like pixel-wise matching to CNNs, which have remarkably changed the field by learning hierarchical features from the raw image data. Application techniques such as AlexNet, VGG and ResNet have enhanced the efficiency of image similarity tasks with the help of deep architectures to learn and identify

complicated features on images [17–19]. Other developments have involved the employment of Siamese Networks to train two images to learn measures of similarity, and more recently GANs and transformers to improve feature extraction and image synthesis [20, 21]. Such advancements make contemporary image similarity algorithms very productive for applications in many fields, such as object detection, search engines, and automatic inspection.

2.4 Fusion Algorithms

Multimodal fusion helps combine records from multiple modes of sources to make consolidated information that enhances the accuracy and reliability of the decision-making systems. The traditional single-modal approaches are incapable of capturing the whole spectrum of records, leading to either incomplete or biased results. Some of the early fusion approaches, including the direct combining of the capabilities of modalities, offered an honest strategy to integrate records; however, they encountered topology-specific noise and differing reality features [22]. Advances in DL brought greater state-of-the-art strategies, inclusive of late fusion, which combines decisions from separate fashions, and hybrid fusion, which integrates capabilities and decisions [23]. The advent of DL has further improved multimodal fusion with models like Multimodal Deep Boltzmann Machines (DBMs) and CNNs that may study complex relationships among specific information types [24, 25].

3. OBJECTIVES AND SIGNIFICANCE OF THE STUDY

Several significant challenges are identified in the examination of the evolution of lost and found systems: Earlier systems were not suitable to accept various and flexible inputs even though they received general user inputs and data keywords. The textual models like the TF-IDF scores and cosine similarity failed to capture the context relevance for images and the image matching faced problems regarding lighting conditions, angles, and poses of the object. Also, there were no effective means for integrating the signal sources for the fused signal, which sometimes meant only partial or even prejudicial information. Thus, this study aims to contribute to solving this problem by presenting a two-branch Siamese network with a fusion model that can use the SBERT for semantic analysis and contrast in the visuals. This is an innovation strategy that is expected to assist individuals in efficiently locating lost items and within the shortest time possible.

The primary goal of this research is to create an application that can accurately match lost and found bags using a multimodal analysis approach. Specific objectives include: developing an image similarity matching model through data gathering and training a Siamese network to capture nuanced bag features; creating a textual similarity matching model by collecting a rich dataset of textual descriptions and training the SBERT model to handle varied and ambiguous descriptions; designing a data fusion model that combines outputs from the image and text similarity models to derive a unified similarity score and developing an intuitive and user-friendly UI to simplify the submission and searching of lost items, featuring detailed forms and a recommendation system for similar items.

The significance of this study lies in its potential to revolutionize lost and found systems through the integration of advanced DL techniques. By combining textual semantic matching and image simi-

larity comparison, this research addresses a critical need for more accurate and efficient methods of reconnecting missing items with their owners. The development of a multimodal approach enhances the robustness and precision of the matching process, ultimately improving user satisfaction and operational efficiency. This study not only contributes to the academic field of deep learning and textual context processing but also offers practical solutions to real-world problems, demonstrating the transformative impact of technology on everyday challenges.

4. DATA AND METHODS

4.1 System Design Overview

The system consists of several interconnected key components, which are shown in FIGURE 1, to address the research problem by integrating both textual and visual data to get a universal similarity score. It comprises three primary modules: a textual module that employs the SBERT model, an image module that employs the Siamese network, and a fusion module that employs a CNN to combine data from both the text and image modules into a single similarity score.

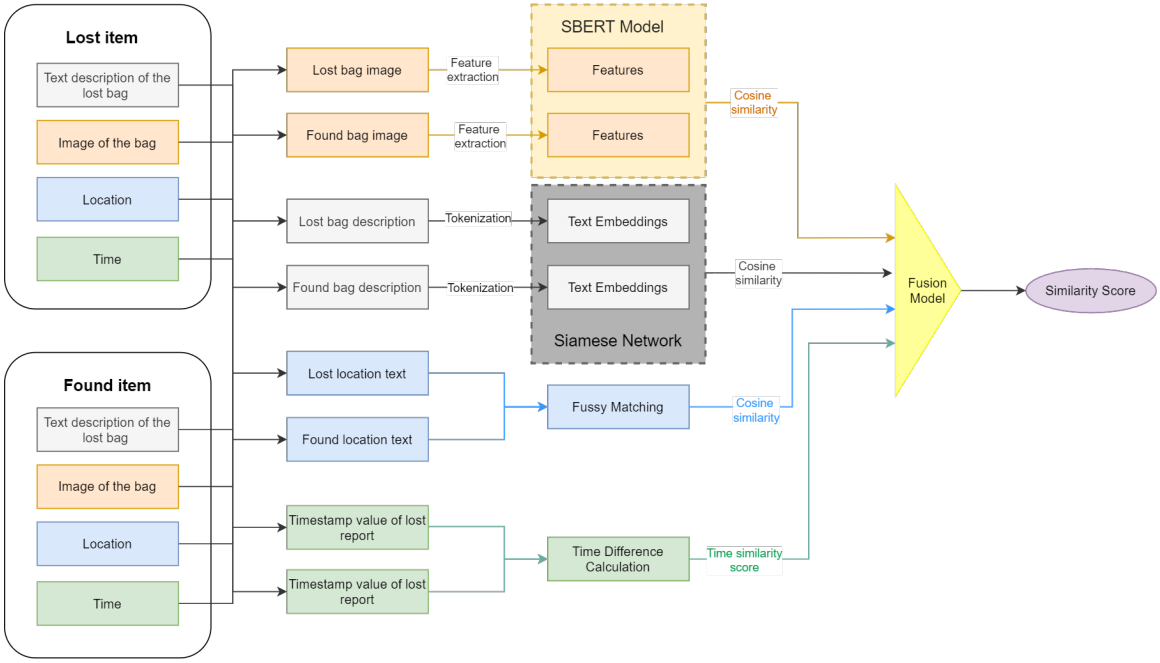

Figure 1: Design diagram

The network model architecture consists of a multi-input fusion network that captures semantic similarity from SBERT, visual similarity from a Siamese CNN, fuzzy-matched location similarity, and normalized time difference. These feature vectors are concatenated and passed through several dense layers that compute non-linear interactions between inputs from different modalities. The architecture consists of two fully connected hidden layers (with 32 and 16 units) along with ReLU activations, batch normalization, and dropout (rate = 0.3) for regularization. The final output layer is

a single neuron with a sigmoid activation function, which outputs a scalar similarity score between 0 and 1, corresponding to the probability of a match.

The system begins with two primary inputs: the lost item and the found item. Each item consists of an image and a text description, which are processed through different neural network models to extract meaningful features. The textual descriptions are fed into the SBERT (Sentence-BERT) model, a state-of-the-art transformer-based network designed for text similarity tasks. The SBERT model is applied to textual descriptions, capturing sentence-level semantic meanings. It converts textual descriptions into dense vector embedding, so much more of the meaning between sentences is compared more subtly than the syntactic structure in descriptions of items. Simultaneously, the images of the lost and found items are processed through a Siamese network, which possesses two identical branches and shares weights, which allows for the parallel processing of features extracted from lost and found images ensuring that both images are processed identically to produce comparable feature embedding. This neural network has been trained to reduce dimensionality and transform image data into a compact representation that captures key visual characteristics needed for item matching. To enhance the matching process further, our system incorporates additional features such as location and time. These features are critical for prioritizing recent activities in the same geographical area, increasing the likelihood of accurate matches. The extracted features from the SBERT model, Siamese network, location, and time are combined to form a comprehensive dataset, which serves as the input for the fusion model. Cosine similarity is used on the features that were extracted between two pairs of items to find out how similar they are. This is done for both text embedding and image feature vectors separately. The outputs, namely the text comparison similarity score and image comparison similarity score, the fuzzy matching score of location data, and the time difference between the lost and found reports, are fused to denote a total similarity score. This approach ensures that all modalities, especially image embedding data and text embedding data have a role in the final decision; resulting in a comprehensive multimodal assessment. This cross-modal total represents the probability of a correct match between the lost item and the found item, with a higher score corresponding to a higher probability of a match. Hence, this synergy that aligns the strength of the textual semantic analysis with the evaluation of visual similarity somehow has the potential to outperform traditional methods that use only one or the other modality. The last and final piece is a web application that incorporates this integrated system: users can enter descriptions and images of lost or found bags, and the system, using the developed models, will produce the most semantically matching item from the database. This practical application will lead to a community-driven platform for recovering lost belongings.

4.2 Data Preparation

Three main datasets were created for developing and training this research project: a dataset of textual descriptions of lost and found bags, a dataset of bag images, and a combined dataset for the fusion model.

4.2.1 Text dataset

We have created 2 test datasets consisting of 2 languages: English and Sinhala. The dataset comprises 2,600 pairs of textual descriptions of lost and found bags, with 1,799 entries allocated for

training and 801 for testing. Due to the difficulty of acquiring such diverse data, the descriptions were generated using a Python script that created a large number of unique and specific descriptions while checking all entries to avoid redundancy. The script employed pre-defined attributes such as colour, type of bag, material, and location. It employed various templates to create realistic and contextually rich descriptions for both lost and found items. For instance, the script would generate a description like, ”I lost a black leather handbag with a golden chain. It was last seen near the main entrance of the university”, ensuring diversity and specificity in the data. This same synthetic method was applied to create the Sinhala dataset as well.

To prepare these datasets for training, several preprocessing steps were undertaken. Textual data was tokenized and transformed into embedding using the SBERT model, preserving semantic information. Each description typically includes details such as colour, brand, size, distinctive marks, locations where the bags were lost or found, and the contents of the bags.

4.2.2 Image dataset

A total of 848 bag images were taken from the sources available, which included images taken by the research group and bag images extracted from the web. The variability in image sources was considered extremely crucial for developing a model that generalized across not only different types of images but also across a wide spectrum of image problems, such as varying backgrounds, light, or colour settings.

Considering that users may upload similar pictures of their bags available over the internet, the dataset was broken down to 283 different classes of bags containing almost all varied types of bags, including school bags, backpacks, women’s handbags, duffle bags, laptop bags, travel bags, and luggage bags. Then, each unique bag was placed in an individual folder and labelled with a corresponding numerical-related label, with consecutive numbering starting from 001. In every folder, at least two images of one bag are placed but captured from different angles and different lighting conditions over different backgrounds. This was very helpful in exposing that particular model to many visual inputs from that specific bag so that better feature extraction and similarity assessments could be made. Each directory was assigned to a specific shopping bag, and images in that directory were related to that shopping bag for training the Siamese network to identify and compare visual similarities. Images were preprocessed to maintain a consistent size and format. Photos were specifically resized to 224 x 224 px and normalized to have a similar input for its Siamese network model. To enhance the robustness, generalizability and number of images in the dataset, data augmentation was employed. Data augmentation was applied through rotating, shifting in terms of width and height, shearing, zooming, and horizontal flipping, which increased the variability of training data similar to previous studies [26, 27]. Some sample images of the bags used for training the model are presented in FIGURE 2.

4.2.3 Fusion model dataset

There are four main features in this dataset: textual similarity, image similarity, location similarity, and time similarity. Textual similarity is calculated using the cosine similarity of embedding generated by the fine-tuned SBERT (Sentence-BERT) model, which processes the descriptions of

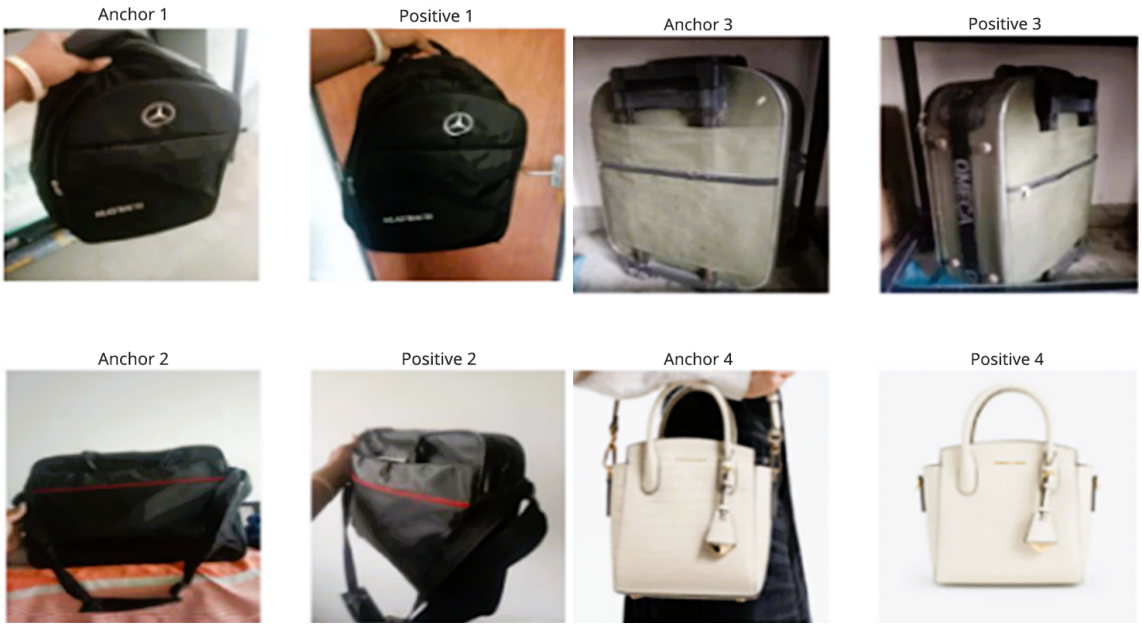

Figure 2: Sample images from the training dataset

the lost and found items. Image similarity is derived from the cosine similarity outputs of the Siamese network. For location and timestamp features, we employed fuzzy matching techniques and normalization. Location similarity is calculated using fuzzy matching, which provides a normalized similarity score between location descriptions. Timestamp similarity is determined based on the difference in days between the lost and found reports, normalized to a number between 0 and 1, which gives higher scores to more recent matches. But in this dataset, we simulated it with the random value generated. To enhance the robustness of the dataset, we introduced two target variables: 0 and 1. These targets help the model distinguish between similar and dissimilar items. A target value of 1 indicates that the lost and found items are similar (a match), while a target value of 0 indicates dissimilarity. Each entry in the dataset consists of the text similarity, corresponding image similarity, location, and timestamp, along with the calculated similarities and the target variable. In total, the dataset comprises 170 data entries, each representing a unique lost or found case. This dataset includes both positive examples (target value 1) where the lost and found items match, and negative examples (target value 0) where they do not. This balanced representation ensures that the fusion model can learn effectively from both similar and dissimilar pairs, improving its ability to accurately identify matches.

4.3 Model Development

4.3.1 SBERT model for text semantic analysis

The SBERT architecture modifies the standard BERT through the addition of a pooling operation over token-based representations to derive a fixed-sized sentence embedding. At a higher level, SBERT does not change the core architecture of BERT. Every encoder layer consists of multi-

head self-attention, followed by the position-wise fully connected feed-forward network. In the SBERT model, the pooling layer after the final BERT layer effectively transforms embedding of varying lengths into a fixed-size dense vector. This vector allows for straightforward computation of semantic similarity between two sentences. Generally, cosine similarity is calculated over sentence-pair similarity, which is quite effective and efficient for semantic comparison. The SBERT model used in this research is based on a pre-trained model 'paraphrase-xlm-r-multilingual-v1'. During fine-tuning, we used the pre-trained Sentence-BERT model with the paraphrase-xlm-r-multilingual-v1 model, applying Multiple Negatives Ranking Loss. The process of fine-tuning was done with a batch size of 16 over 4 epochs.

4.3.2 Siamese network model for image similarity comparison

The Siamese network architecture applied in this work is developed for the assessment of visual similarity between images of lost and found bags. The presented architecture employs a lightweight, efficient version of the MobileNetV2 architecture in feature extraction by using CNNs. Using the MobileNetV2 model enables the efficient processing of high-dimensional image data. It is a Siamese network composed of three parallel branches, each of which processes one of the three input images for one triplet: the anchor image, the positive image, and the negative image. The architecture was created by learning a mapping from the image space to an embedding space, such that the distance between embedding reflects the visual similarity of the images.

The approach used a custom data generator that produces anchor-positive-negative image batches for the training process. This makes sure that the triplets within each batch are varied and augmented, hence enabling learning about proper visual similarity and dissimilarity. The system defines three input layers to accept the anchor, positive, and negative images, each with a resolution of 224x224x3 pixels, ensuring compatibility with the pre-trained MobileNetV2 model. For feature extraction, the first 60 layers of MobileNetV2 are fine-tuned. The extracted features are then flattened to 1D and passed through two dense layers, each with 64 units and an L2 regularization rate of 0.01, followed by a dropout layer with a rate of 0.6 to prevent overfitting. Batch normalization is applied after the dense layers to stabilize and accelerate training through normalization. Additional dropout layers with rates between 0.5 and 0.7 are included to further regularize against overfitting by randomly dropping a fraction of input units to zero. The final dense layer produces a 64-dimensional space embedding for each image, which is concatenated into the Siamese network output and used for computing triplet loss. Even more fine-tuning of the model was achieved by implementing hyperparameter optimization through Keras Tuner. This involved the search for the best choice for hyperparameters, such as how many stackable layers in MobileNetV2, how many units in Dense layers, dropout rates, and L2 regularization rates. A similar triplet loss function was used to train the custom Siamese network model. This meant that the triplet loss function was used to make the distance between the anchor and positive embedding as small as possible and the distance between the anchor and negative embedding as large as possible. The loss is computed as follows, as introduced by Mustapha et al. (2021) [28]:

$$\text{loss} = \sum \max(\text{distance}(\text{anchor}, \text{positive}) - \text{distance}(\text{anchor}, \text{negative}) + \alpha, 0) \tag{1}$$

where 'α' is a margin parameter to ensure the network learns meaningful separations between similar, dissimilar images. It indeed guarantees that the model is going to learn effectively how to differentiate between similar and dissimilar images. The total training images are 848, out of

which data augmentation is done using ImageDataGenerator. Next, the training goes through a stratified k-fold cross-validation process for robustness. This architecture has successfully captured and matched features, including the colour, shape, and texture of bags.

4.3.3 Data fusion model for combining text and image similarity outputs

The development of the fusion model for matching lost and found items involves a series of meticulously designed steps, each chosen to enhance the model's ability to integrate multi-modal data accurately and efficiently. The SBERT outputs the semantic similarity between textual descriptions and a Siamese network model that measures the visual similarity of the images. The fusion model simply combines the two similarity scores to obtain one overall similarity score representing textual and visual information. The two similarity scores that were derived are then combined using a fully connected neural network in the following manner. This neural network is able to learn how to combine these textual and visual similarity scores in the most effective manner in order to arrive at a final decision.

The model begins with an input layer that takes four key features: textual similarity from SBERT, image similarity from the Siamese network, fuzzy matching score of the locations, and timestamp similarity, representing different aspects of the lost and found items. The following dense layers learn a non-linear combination of these input characteristics, enabling the model to understand complex relationships between the similarities of text, image, and other features. The Rectified Linear Unit (ReLU) activation in the dense layers introduces nonlinearity, allowing the model to learn more complex patterns. Dropout is applied to prevent overfitting by randomly setting a fraction of input units to zero during training, thereby enhancing the model's robustness. Batch normalization stabilizes the inputs for each layer, speeding up the training process. Finally, the output layer produces a similarity score for each text-image pair, combining the textual and visual similarity scores to determine an overall match.

5. RESULTS

5.1 Image Similarity Matching Model

The Siamese network model was trained using a triplet loss function, which ensured that the distance between the anchor and positive images (same bag) was minimized while the distance between the anchor and negative images (different bags) was maximized. We used a pre-trained model to identify and compare the visual features of lost and found bag images. To determine the best pre-trained network VGG16, Resnet50 and MobileNetV2 models are tested and compared. Initially, the training process was done with 50 defined epochs. The ResNet50 accuracy from training and validation is very high and almost equal to each other. Thus, this shows the network has over-fit. More regularization, more data augmentation, dropout, and an early stopping mechanism were some of the model adjustments made to address the over-fitting issue by adjusting the learning rate batch size and Reduced Model Complexity by changing the pre-train model to VGG16. Increased L2 regularization and dropout rates for further addressing the overfitting issue, but still got high overfitting and high fluctuations. After conducting an analysis, the switch was made to a more com-

pact base model (MobileNetV2), stronger regularization was applied, balanced data augmentation was utilized, the learning rate was decreased even further, and gradient clipping was implemented. According to the results, this model switch has a significant improvement, with validation accuracy that is closer to training accuracy and less divergence in loss.

5.2 Textual Similarity Matching Model

The fine-tuning of the SBERT (paraphrase-xlm-r-multilingual-v1) for lost and found reports resulted in 0.9625, indicating that approximately 96.25% of the lost and found item pairs were correctly matched. The balanced F1 Score, in terms of precision and recall, was 0.9513; as such, there is a high effectiveness in the identification of correct matching by the model. The Pearson Correlation coefficient was estimated at 0.9787, suggesting quite a strong linear relationship between predicted and true labels. To study the embedding generated by the SBERT model, t-SNE as a dimensionality reduction and visualization technique was applied.

5.3 Data Fusion Model

The fusion model, integrating both text and image similarity scores, exhibited promising performance in the task of identifying similar bag pairs. The model is first pre-trained on the dataset with pre-computed cosine similarity scores from SBERT to get textual embedding and from CLIP SEEN to get image embedding. The final validation accuracy of the fusion model was 0.7630, corresponding to a validation loss that steadily decreased to below 0.25. The training accuracy improved significantly over 50 epochs, starting from around 0.5 and reaching 0.87, with the training loss decreasing consistently from around 2.0 to below 1.0. The validation accuracy exhibited an optimistic upward trend, stabilizing at 100% closely tracking the training accuracy, indicating very little overfitting. Further, the accuracy of the model was tested in a confusion matrix to confirm the model's effectiveness, showing that all 17 instances of class 0 (non-matching pairs) and all 23 instances of class 1 (matching pairs) were correctly identified, resulting in perfect classification with no false positives or false negatives. The classification report supported these findings, with the model achieving perfect precision, recall, and F1 scores of 1.0 for both classes. The overall accuracy of the model was 0.87, thus proving the model's effectiveness and reliability in efficiently linking lost and found items through the integrated multimodal data. These overall evaluation metrics show that the model was adequately prepared for the generalization of the new data and provided the best performance over the validation set.

The confusion matrix further confirmed the model's effectiveness (FIGURE 3).

The confusion matrix shows that out of 38 instances of class 0 (non-matching pairs), 35 were correctly identified, with 3 false positives. Similarly, out of 37 instances of class 1 (matching pairs), 35 were correctly identified, with 2 false negatives. This resulted in a high level of classification accuracy, demonstrating the model's strong performance in distinguishing between matching and non-matching pairs. The classification report supported these findings, with the model achieving an F1 score of approximately 0.93. The overall accuracy of the model was also high, reflecting its robustness in accurately matching lost and found items based on the integrated multi-modal data.

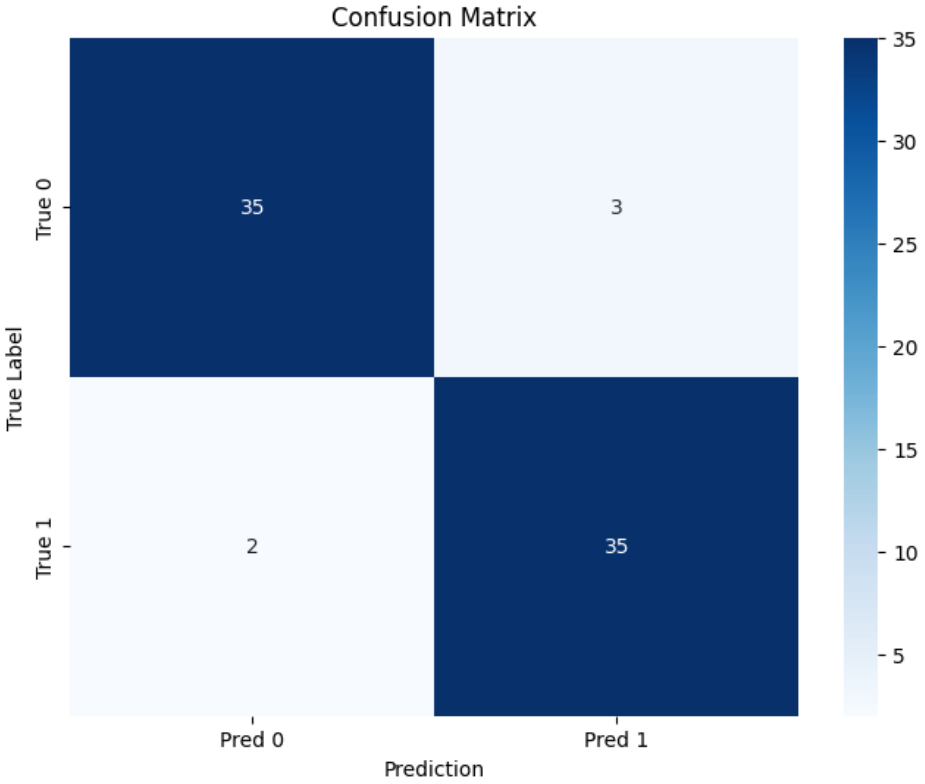

Figure 3: Confusion matrix

These comprehensive evaluation metrics indicate that the model was well-trained for generalisation on new data, achieving optimal performance over the validation set.

5.4 User-friendly GUIs

The key interfaces of the system are shown under FIGURE 4. The first input screen (FIGURE 4(A)) is used to enter the specific information about the lost or found bag. Users can also upload the image of the bag, simplifying the identification process. They also provide a textual description where they describe the object in detail, including the type of bag, any characteristics that can help in identification, and the place where the object was lost or found. This full information guarantees that the system is provided with sufficient information to make accurate matches. Details about the user's personal identification, such as a phone number and address are entered in the second input screen. These include their first and last name, email address, phone number, and physical address. The details section is important as this helps the person who has lost an item to be in touch with the one who has found it with a view to recovering it (FIGURE 4(B)). The output screen (FIGURE 4(C)) displays the results of the match made by the application. It shows an image of a bag, the description of which matches the one given by the user, and details the place where it was found. A percentage match designation means the likelihood of the match. The users can then

provide feedback and specify if the match is correct or incorrect so as to contribute to enhancing the efficiency of the system in the future.

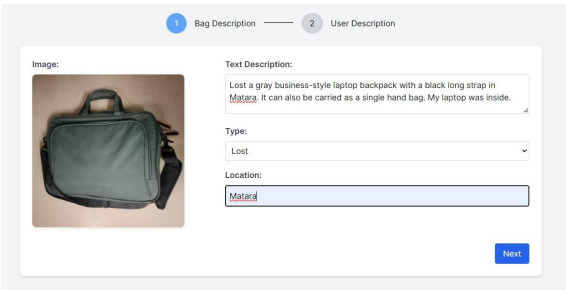

Figure 4(A)

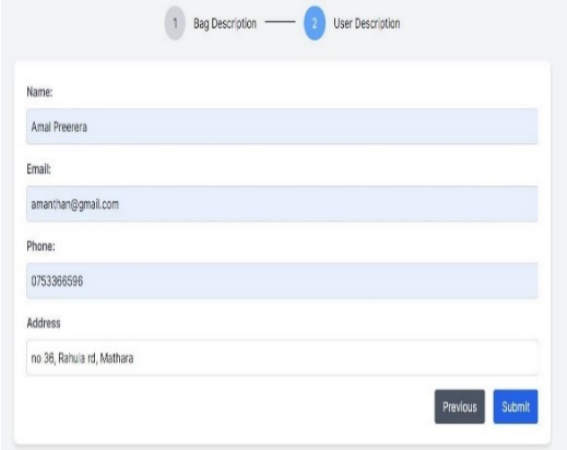

Figure 4(B)

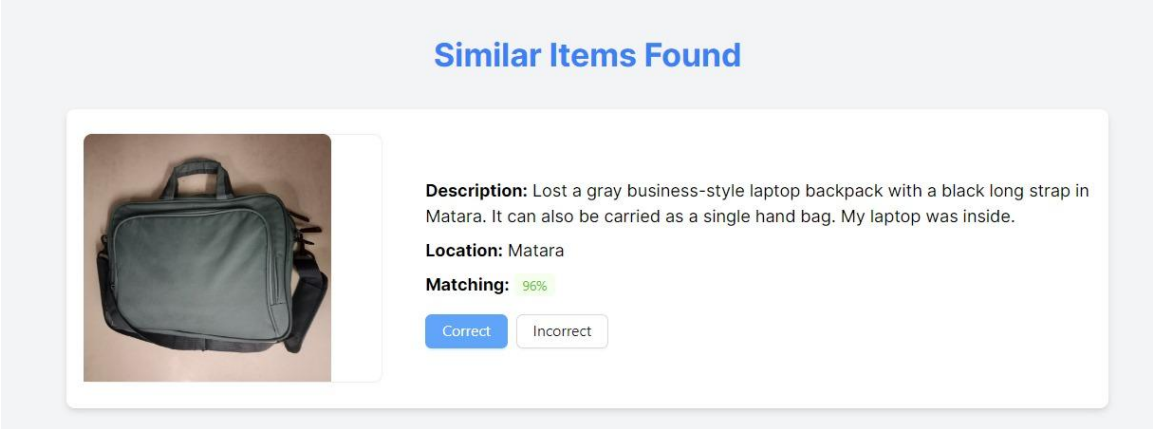

Figure 4(c)

Figure 4: Key interfaces of the application

6. DISCUSSION

Our approach addresses the limitations of traditional methods that rely solely on text-based or image-based matching by combining both modalities to leverage their complementary strengths. The SBERT model, which we fine-tuned on a dataset of 2,600 lost and found description pairs, achieved a high accuracy of 0.9526 and an F1 score of 0.9405 and the Siamese network, trained on 848 bag images using the MobileNetV2 architecture, attained an accuracy of 0.75, indicating its capability to discern visual similarities between images. Combining these two models into a fusion model that also incorporates location and time features further improved the system's performance, achieving an overall accuracy of 0.87 and an impressive F1 score of 0.98. This multi-modal approach not only enhances the matching process by considering both text and image data but also prioritizes matches based on recent activities and geographical proximity, thus increasing the

likelihood of accurate matches. Previous research employed simple sentence comparison methods that relied on query search as in Salman and Athab (2022) [1], who designed systems that enabled users to enter item descriptions and images, enabling keyword-based searches. While effective, these systems did not incorporate advanced matching algorithms, resulting in limitations regarding matching precision. Similarly, BM25 and TF-IDF sentence-matching models are contingent on lexical similarity, thereby failing to capture semantic similarities. Studies that relied solely on textual or visual information. For instance, Suryani and Edy (2020) [4], achieved an 88% accuracy rate using Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity methods in an Android ”Lost & Found” application. However, their reliance on textual descriptions posed a challenge due to the variability in user inputs, and the SBERT model applied in this research is among the state-of-the-art models. It proved to be best at capturing the semantic details and had a high correlation coefficient in our study, giving a high accuracy rate of 95%.

One of the main problems in text semantic analysis and sentence matching is that traditional models are not quite effective with multilingual data. The already mentioned models include TF-IDF, BM25, and those early ones based on neural networks Word2Vec and GloVe fundamentally designed for perplexing holding structures at best, so their performances are brought down within a multilingual context. We used fine-tuned paraphrase-xlm-r-multilingual-v1, on a multilingual dataset composed of English and Sinhala descriptions, which attained an accuracy of 96.25% with an F1 score of 0.9513. Zhou et al. (2023) [5], introduced ”LostNet,” which utilized MobileNetv2 combined with the Convolutional Block Attention Module (CBAM) to enhance item identification by comparing user-provided images with those in the system, achieving a testing accuracy of 96.8%. Despite its high accuracy, the system was limited to laptop use, making it less practical for broader applications. Our system, implemented as a web application, offers greater accessibility and practicality, showing its potential for real-world application in diverse environments.

Huang et al. (2023) [29], proposed a method utilizing ResNet for feature extraction to improve text detection accuracy in images. Their approach demonstrated strong performance in extracting deep visual features but also highlighted challenges related to computational efficiency. In comparison, our Siamese network, trained using MobileNetV2, achieves a 75% accuracy rate while offering improved computational efficiency. Unlike traditional feature extraction methods that require heavier models, our approach learns compact embeddings during training, leading to faster comparisons during prediction. While KNN depends on distance calculations that are computationally expensive, our model learns embedding during the training process, which take less time to compare during prediction. The efficiency was enhanced by dividing computational load across both text and image by combining SBERT and Siamese networks. It was further refined by performing the computations on both text and image modalities using SBERT and Siamese networks, respectively. In a different approach, Prawira and Saputri (2023) [6], came up with a lost item detection model that can compare images by using string matching techniques, drawing on NLP and the ResNet-50 architecture. Although this study achieved an impressive string-matching accuracy of 97.92%, the image comparison accuracy was only 29.96%, which highlights the need for more effective feature extraction techniques. The project mainly adopted an agile approach due to limited time availability and the nature of the application, which demands flexibility to address evolving user needs and data types. Further, the development was guided by key aspects relevant to software evolution and maintenance [30–32].

The following TABLE 1, provides a summary of the advantages and features of our model in comparison to existing models and research. It highlights the key advancements our approach offers across text analysis, image similarity, and multimodal fusion, emphasizing improvements in multilingual and noise-tolerant performance, efficient feature extraction, and dynamic fusion techniques. In contrast to traditional and existing models, our solution addresses common limitations in accuracy, computational efficiency, and scalability, offering a robust, real-world applicable alternative

Table 1: Comparison of existing lost and found systems and the proposed model.

Aspect	**Existing models**	**Our study**
Text semantic analysis	Traditional models like BM25 [33], and TF-IDF focus on keyword matching, not semantics. Word2Vec [10], and GloVe [11], capture word-level semantics but lack deep context understanding. BERT [14], requires extensive fine-tuning for multilingual tasks.	SBERT (paraphrase-xlm-r-multilingual-v1) captures deep semantic nuances at sentence level, performs robustly across multilingual data, achieving high Pearson/Spearman correlation (0.9787). Tuned for noise tolerance and real-world adaptability.
Image similarity	Handcrafted feature models like SIFT [34], and HOG [35], capture only local features. Deep ranking models [36], offer better semantic understanding but are computationally expensive. Models like KNN and early CNNs suffer from scalability and efficiency issues.	Siamese Network with MobileNetV2 captures both local and global features, achieving triplet accuracy of 0.8194. Robust to variations (lighting, orientation) via heavy data augmentation. Stratified K-fold validation improves generalization, while maintaining computational efficiency.
Fusion strategy	Early fusion methods [37], often naïvely combine modalities without optimal weighting, leading to poor performance. Static heuristics limit adaptability.	Neural fusion model combines text and image similarity optimally via learned weights using a deep neural network, offering adaptability, flexibility, and better generalization across tasks.
Lost and found systems	LostNet [5]: MobileNetV2 + CBAM, high accuracy (96.8%) but laptop-restricted. Huang et al. (2023) [29]: KNN + ResNet, high on small datasets, but computationally heavy. Prawira & Saputri (2023) [6]: ResNet-50 + NLP, only 29.96% image comparison accuracy.	Web-based solution using SBERT + Siamese Network. Achieves 75%+ accuracy efficiently. Designed for broad accessibility and scalability. Balances text and image modalities, optimized for real-world noisy, multilingual data.

The limitations of our system include the lack of language translation capabilities, which restricts its ability to handle multilingual inputs seamlessly. Although we have created our text-matching algorithm for both English and Sinhala languages, the system does not support automatic translation between these languages, potentially reducing its effectiveness in a diverse linguistic environment.

Another limitation is the location searching functionality, which could be significantly improved by incorporating precise longitude and latitude data for more accurate geospatial matching. Additionally, the system's comparison speed is hampered by the use of serial matching, resulting in slower performance when processing large datasets.

For the initial phase, most of the text data was created through data synthesis. To develop a more robust, generalizable, and scalable system, future research will expand the dataset to include a broader range of real-life lost and found item data from actual contexts, settings, and real-world scenarios, which would yield stronger representations. Incorporating increasingly heterogeneous datasets would allow models to learn broader representations and ultimately achieve greater robustness in non-stationary conditions. Additionally, since the initial research involved training only with Sinhala and English contexts, expanding the training to include datasets in other languages will improve the project's global applicability. Furthermore, system performance in large-scale deployments can be enhanced by using techniques such as parallel processing and incorporating datasets with real geolocation metadata (e.g., GPS coordinates) instead of relying solely on fuzzy string-matched location names. This would improve location-based filtering, leading to more accurate and context-aware item matching.

7. CONCLUSION

The research aimed to develop a robust system for matching lost and found items, particularly bags, by leveraging advanced deep-learning techniques. The developed predictive model for identifying similarities in lost item reports has demonstrated its effectiveness by utilizing multimodal DL techniques for optimized semantic matching. The research focused on a combination of textual and visual data, with textual features extracted using SBERT (Sentence-BERT) models and visual features analyzed using a Siamese neural network. These approaches have been integrated into a unified system to create a comprehensive similarity score that helps in comparing lost items with found items, especially for bags. The final fusion model achieved an accuracy of 0.87, demonstrating the effectiveness of the proposed system. It has achieved a remarkable degree of accuracy in recognizing and matching lost objects based on both textual descriptions and images.

8. CONFLICTS OF INTEREST

The authors declare that there are no financial or non-financial conflicts of interest related to this research.

9. FUNDING

No funding was received for this study.

References

[1] Abdul-Jalil Salman ZA, Athab OA. Smartphone Application for Managing Missed and Found Belongings. MEST J. 2022;10:66-71.

[2] Peter OS, Roseline OO, Oluwakemi CA. Ifound-an Online Lost Item Recovery Application. i-Manager's Journal on Information Technology. 2019;8:1.

[3] Suchana K, Alam SM, Meem AT, Turjo MD, Khan MM. Development of User-Friendly Web-Based Lost and Found System. J Softw Eng Appl. 2021;14:575-590.

[4] Suryani L, Edy K. PENGEMBANGAN APLIKASI" LOST & FOUND" BERBASIS ANDROID DENGAN MENGGUNAKAN METODE TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY. Electro Luceat. 2020;6:190-204.

[5] Zhou M, Fung I, Yang L, Di Wan N, K Wang T. LostNet: A Smart Way for Lost and Find. 2023. Arxiv preprint: https://arxiv.org/pdf/2301.02277

[6] Prawira J, Saputri TR. Lost Item Identification Model Development Using Similarity Prediction Method With CNN Resnet Algorithm. J Autonom Intell. 2023;7.

[7] Salton G. Introduction to Modern Information Retrieval. McGraw-Hill. 1983

[8] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41:391-407.

[9] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993-1022.

[10] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. ArXiv preprint: https://arxiv.org/pdf/1301.3781

[11] Pennington J, Socher R, Manning CD. Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Stroudsburg PA USA: Association for Computational Linguistics. 2014:1532-1543.

[12] Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. 2017. ArXiv preprint: https://arxiv.org/pdf/1705.02364

[13] Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, et al. Universal sentence encoder. 2018. ArXiv preprint: https://arxiv.org/pdf/1803.11175

[14] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pretraining of deep bidirectional transformers for language understanding. 2018. ArXiv preprint: https://arxiv.org/pdf/1810.04805

[15] Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. Roberta: A Robustly Optimized Bert Pretraining Approach. 2019. ArXiv preprint: https://arxiv.org/pdf/1907.11692

[16] Reimers N, Gurevych I. Sentence-Bert: Sentence Embeddings Using Siamese BERT-Networks. 2019. ArXiv preprint: https://arxiv.org/pdf/1908.10084

[17] Krizhevsky A, Sutskever I, Hinton GE. Imagenet Classification With Deep Convolutional Neural Networks. Adv Neural Inf Process Syst. 2012.

[18] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. ArXiv preprint: https://arxiv.org/pdf/1409.1556

[19] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. New York: IEEE. 2016:770-778.

[20] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE. 2005;1:539-556.

[21] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun. ACM. 202022;63:139-144.

[22] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: A survey. Multimedia Syst. 2010;16:345-379.

[23] Snoek CG, Worring M, Smeulders AW. Early Versus Late Fusion in Semantic Video Analysis. In: Proceedings of the 13th annual ACM international conference on multimedia. ACM. 2005:399-402.

[24] Ngiam J, Khosla A, Kim M, Nam J, Lee H, et al. Multimodal Deep Learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). 2011:689-696.

[25] Neverova N, Wolf C, Lacey G, Fridman L, Chandra D, et al. Learning Human Identity From Motion Patterns. IEEE Access. 2016;4:1810-1820.

[26] Paul H, Udayangani H, Umesha K, Lankasena N, Liyanage C, et. al. Maize Leaf Disease Detection Using Convolutional Neural Network: A Mobile Application Based on Pre-trained VGG16 Architecture. N Z J Crop Hortic Sci. 2024:1-17.

[27] Lankasena N, Nugara RN, Wisumperuma D, Seneviratne B, Chandranimal D, et al. Misidentifications in Ayurvedic Medicinal Plants: Convolutional Neural Network (CNN) to Overcome Identification Confusions. Comput Biol Med. 2024;183:109349.

[28] Mustapha MF, Mohamad NM, Osman G, Ab Hamid SH. Age Group Classification Using Convolutional Neural Network (CNN). J Phys Conf S. 2021;2084:012028.

[29] Huang LK, Tseng HT, Hsieh CC, Yang CS. Deep Learning-Based Text Detection Using Resnet for Feature Extraction. Multimedia Tool Appl. 2023;82:46871-46903.

[30] http://ir.kdu.ac.lk/handle/345/7407

[31] http://ir.kdu.ac.lk/handle/345/7425

[32] http://ir.kdu.ac.lk/handle/345/7418

[33] Robertson SE, Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. London: Springer London. 1994:232-241.

[34] Lowe DG. Distinctive Image Features From Scale-Invariant Keypoints. Int J Comput Vis. 2004;60:91-110.

[35] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005;1:886-893.

[36] Wang J, Song Y, Leung T, Rosenberg C, Wang J, et al. Learning Fine-Grained Image Similarity With Deep Ranking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014:1386-1393.

[37] Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41:423-443.