Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

EXPLORING BANGLADESH CULTURAL HERITAGE THROUGH INTEGRATING DEEP CONVOLUTIONAL NEURAL NETWORK BY IMAGE CAPTIONING

Javed Hossain & Md. Ataullah Bhuiyan

Department of Computer Science & Engineering, City University, Dhaka, Bangladesh

ABSTRACT

Image captioning is a significant task at the intersection of computer vision and natural language processing that aims to automatically generate concise textual descriptions of images. Although seemingly simple for humans, this task is complex for machines as it requires both accurate image analysis and the generation of semantically coherent sentences. Recent advances in encoder–decoder architectures, which combine convolutional neural networks for feature extraction with recurrent or transformer-based networks for sentence generation, have achieved promising results in this domain. In this work, we focus on applying image captioning to represent Bangladeshi culture, traditions, foods, and heritage sites, an area largely overlooked in existing research. To this end, we build a novel deep convolutional neural network model trained on a curated heritage dataset consisting of images of historical landmarks, cultural events, and traditional foods of Bangladesh. The proposed model generates culturally enriched captions that highlight not only the visual content but also its cultural and historical significance. Our system can serve as a digital bridge to promote Bangladeshi culture, benefiting travelers, researchers, and enthusiasts while contributing to cultural preservation. Ultimately, this study demonstrates how image captioning can extend beyond visual

description to support heritage promotion and global cultural engagement.

Keywords: *Deep Neural Network, Digital Image Processing, Natural Language Processing, Pretrained Architecture.*

Corresponding author: Javed Hossain can be contacted at javed.cucse@gmail.com

1. INTRODUCTION

Cultural heritage serves as a vital reservoir of a nation's identity, reflecting its rich history, traditions, and artistic expressions. In the contemporary era, the convergence of technology and cultural preservation has opened new avenues for exploration and understanding. This paper delves into the cultural tapestry of Bangladesh, a nation steeped in history and diversity (Abunadi & Senan, 2022), leveraging the power of Custom Transfer Learning to enhance the process of image captioning. The integration of advanced neural network architectures in image understanding and description not only facilitates the documentation of cultural artifacts but also fosters a deeper engagement with the heritage they represent (Al-muzaini, T. N., & Benhidour, 2018).

Bangladesh, with its vibrant and multifaceted cultural heritage, presents a compelling backdrop for this exploration. From the ancient ruins of Mahasthangarh to the intricate designs of Jamdani textiles, the country's cultural wealth is vast and varied. However, the effective portrayal and interpretation of these cultural elements through image captioning pose unique challenges. Conventional image captioning methods often struggle with the nuanced understanding required for

accurately describing culturally significant scenes. In response to this, our research proposes the integration of Custom Transfer Learning, a fusion of different deep neural network architectures, to address the intricacies involved in capturing and conveying the essence of Bangladesh's cultural heritage.

This paper aims to contribute to the field of computer vision and cultural heritage preservation by introducing a novel approach to image captioning. By combining the strengths of convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential data processing (Duggal et al., 2017), our Custom Transfer Learning model seeks to provide more contextually rich and culturally sensitive captions. The synthesis of these neural network components is anticipated to enhance the interpretability of images, offering a more nuanced understanding of the cultural context embedded within them.

In the subsequent sections, we elaborate on the background of image captioning, discuss the unique challenges associated with portraying cultural heritage, and present the architecture and methodology of our proposed Custom Transfer Learning model. Through this research, we aspire to not only advance the capabilities of image captioning systems but also to contribute to the broader discourse on leveraging technology for the documentation and appreciation of cultural heritage.

Despite the rich cultural heritage of Bangladesh, there is a need for innovative and efficient methods to explore and showcase its diverse cultural elements. Traditional methods of documentation and presentation may not fully capture the essence and significance of the country's cultural heritage. To

address this challenge, there is a demand for an advanced technological solution that integrates Deep Convolutional Neural Network (DCNN) with Image Captioning techniques. This project aims to leverage cutting-edge technology to create a comprehensive platform for exploring and understanding Bangladesh's cultural heritage through automated image analysis and meaningful caption generation. The goal is to develop a system that not only accurately identifies cultural artifacts and practices but also provides insightful and engaging captions to enhance public awareness and appreciation of Bangladesh's rich cultural tapestry.

The present study aims to develop an intelligent system for the recognition, interpretation, and dissemination of Bangladesh's cultural heritage through advanced artificial intelligence techniques. Specifically, it seeks to implement a deep convolutional neural network (CNN) model capable of accurately recognizing and classifying images representing diverse cultural elements, including landmarks, artifacts, traditional clothing, festivals, and historical sites. In addition, the study integrates natural language processing (NLP) techniques to generate coherent and contextually relevant captions for the identified images, thereby enhancing semantic understanding. By automating image analysis and caption generation, the research facilitates the systematic exploration and documentation of cultural heritage, contributing to its preservation and broader accessibility. Furthermore, the proposed system is designed to enhance user engagement through an interactive platform that combines visual content with informative descriptions, ultimately fostering deeper appreciation and awareness. In this regard, the study also serves

as an educational tool aimed at promoting knowledge of Bangladesh's cultural heritage at both national and global levels.

2. REVIEW OF LITERATURE

Researchers applied various methods which incorporate Deep Learning (DL), Deep Neural Networks (DNN) and also Transfer learning Model.

Fashion is one of the many fields of application that image captioning is being used in. For e-commerce websites holding tens of thousands of images of clothing, automated item descriptions are quite desirable. This paper addresses captioning images of clothing in the Arabic language using deep learning. Image captioning systems are based on Computer Vision and Natural Language Processing techniques because visual and textual understanding is needed for these systems. Many approaches have been proposed to build such systems. The most widely used methods are deep learning methods which use the image model to analyze the visual content of the image, and the language model to generate the caption. Generating the caption in the English language using deep learning algorithms received great attention from many researchers in their research, but there is still a gap in generating the caption in the Arabic language because public datasets are often not available in the Arabic language. In this work, we created an Arabic dataset for captioning images of clothing which we named "Arabic Fashion Data" because this model is the first model for captioning images of clothing in the Arabic language. Moreover, we classified the attributes of the images of clothing and used them as inputs to the decoder of our image captioning model to enhance Arabic caption quality. In addition,

we used the attention mechanism. Our approach achieved a BLEU-1 score of 88.52. The experiment findings are encouraging and suggest that, with a bigger dataset, the attributes-based image captioning model can achieve excellent results for Arabic image captioning (Eckardt et al., 2021).

Visual understanding is a research area that bridges the gap between computer vision and natural language processing. Image captioning is a visual understanding task in which natural language descriptions of images are automatically generated using vision-language models. The transformer architecture was initially developed in the context of natural language processing and quickly found application in the domain of computer vision. Its recent application to the task of image captioning has resulted in markedly improved performance. In this paper, we briefly look at the transformer architecture and its genesis in attention mechanisms. We more extensively review a number of transformer-based image captioning models, including those employing vision-language pre-training, which has resulted in several state-of-the-art models. We give a brief presentation of the commonly used datasets for image captioning and also carry out an analysis and comparison of the transformer-based captioning models. We conclude by giving some insights into challenges as well as future directions for research in this area (Ghandi et al., 2023).

In this paper, they introduce a cascade semantic fusion architecture (CSF) to mine the representative features to encode image content through an attention mechanism without bells and whistles. Specifically, the CSF benefits from three types of visual attention semantics, including object-level,

image-level, and spatial attention features, in a novel three-stage cascade manner. In the first stage, object-level attention features are extracted to capture the detailed contents of the objects based on the pre-trained detector. Then, the middle stage devises a fusion module to merge object-level attention features with spatial features, thereby inducing image-level attention features to enrich the context information around the objects. In the last stage, spatial attention features are learned to unveil the salient region representation as a complement to two previously learned attention features. In a nutshell, we integrate the attention mechanism with three types of features to organize context knowledge about images from different aspects. The empirical analysis shows that the CSF can assist the image captioning model in selecting the object regions of interest. The experiments of image captioning on the MSCOCO dataset show the efficacy of our semantic fusion architecture in depicting image content (Hossain, Islam, & Tusar, 2023).

Image captioning refers to the automatic generation of descriptive texts according to the visual content of images. It is a technique integrating multiple disciplines including computer vision (CV), natural language processing (NLP) and artificial intelligence. In recent years, substantial research efforts have been devo to generating image captions with impressive progress. To summarize the recent advances in image captioning, we present a comprehensive review of image captioning, covering both traditional methods and recent deep learning-based techniques. Specifically, we first briefly review the early traditional works based on the retrieval and template. Then deep learning-based image captioning research are focused, which is categorized into the encoder-decoder

framework, attention mechanism, and training strategies based on model structures and training manners for a detailed introduction. After that, we summarize the publicly available datasets, evaluation metrics, and those proposed for specific requirements, and then compare the state-of-the-art methods on the MS COCO dataset. Finally, we provide some discussions on open challenges and future research directions (Jayavikash et al., 2021).

However, the traditional image captioning methods based on RNNs have two main shortcomings: the errors in the prediction process are often accumulated and the location of attention is not always accurate which would lead to misjudgment of risk. To handle these problems, a landslide image interpretation network based on a semantic gate and a bi-temporal long-short term memory network (SG-BiTLSTM) is proposed in this paper. In the SG-BiTLSTM architecture, a U-Net is employed as an encoder to extract features of the images and generate the mask maps of the landslides and other geographic objects. The decoder of this structure consists of two interactive long-short term memory networks (LSTMs) to describe the spatial relationship among these geographic objects so that to further determine the role of the classified geographic objects for identifying the hazard-affected bodies. The purpose of this research is to judge the hazard-affected bodies of the landslide (i.e., buildings and roads) through the SG-BiTLSTM network to provide geographic information support for emergency service. The remote sensing data was taken by Worldview satellite after the Wenchuan earthquake happened in 2008. The experimental results demonstrate that SG-BiTLSTM network shows remarkable improvements on the recognition of

landslide and hazard-affected bodies, compared with the traditional LSTM (the Baseline Model), the BLEU1 of the SG-BiTLSTM is improved by 5.89%, the matching rate between the mask maps and the focus matrix of the attention is improved by 42.81% (Labati, Piuri, & Scotti, 2011).

Table 1. Comparison with Existing Studies

Project Title	Dataset	Methodologies
Hybrid deep neural network for Bangla automated image descriptor	Bangla Natural Text (BNLIT)	ResNet-101 acts as an Encoder and LSTM acts as a Decoder
Arabic Captioning for image of Clothing using deep learning	Arabic Fashion (InFashAlv)	Encoder and Decoder Transformer Neural Network
Generative image captioning in Urdu using deep learning	Flickr-8k Dataset caption converted in Urdu language	ResNet-50 acts as an Encoder and LSTM acts as a Decoder
A Review of Transformer-Based Approaches for image captioning	MS COCO Dataset	Captioning transformers based on Vison-language pre-training

Source: The authors' own work.

Table 2. Proposed System

Project Title	Dataset	Methodologies
Exploring Bangladesh's Cultural Deep Convolutional Heritage through Integrating Deep Neural Network. Convolutional Neural Network by Image Captioning.	Bengali historical place dataset (BPHD)	Deep Convolutional Neural Network

Source: The authors' own work.

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

3. RESEARCH METHODOLOGY

3.1 Data Collection

Here this system enables the generation of image captions for 27 historical places in Bangladesh. All images were collected online by the author. Here this Dataset contains a total of 3303 images. Here this dataset contains 27 famous historical places of Bangladesh whereas every historical place contains 80-100 images. Below the table describes this dataset.

Table 3. Description of Dataset

Historical Place Name	Number of Images
Ahsan Monjil	95
Armenian Church	100
Baitul Mukarram Mosque	85
Bangladesh National Museum	87
Bangladesh Parliament	75
Candrima Uddan	55
Dhakeshwari Temple	85
Kutila Mura	86
Lalbag	92
Liberation War Museum	68
Mahasthangarh	112
Mujibnagar Memorial Complex	48
Muktagacha Rajbari	52
National Martyrs	107
National Zoo	91
Osmani Museum	68
Panam City	88
Radha Govindha Temple	58
Shaheed Minar	78
Shalbon Bihar	84
Shrine of Hazrat Shah Jalal	92
Sixty Dome Mosque	87
Sompur Mahavihara	81
Sonargaon	94
Supreme Court	96
Tajhat Palace	88
Tara Masjid	84

Source: The authors' own work.

3.2 Data Preprocessing

The performance of a DNN depends largely on suitable data preparation. Our data preparation pipeline contains the following techniques to pre-process the microscopic images for ALL detection.

Resizing (224px width, 224px height, 3 channels), Augmentation, Normalization.

There the Augmentation technique contains the following techniques: Shear Range (0.2), Zooming (0.2) Flipping (Horizontally)

Figure 1. Image of Ahsan Monjil

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

	Name	Bangla_Caption	English_Caption	Nearest_tourist_Place
0	Ahsan_Monjil	আহসান মঞ্জিল পুরান ঢাকার ইসলামপুরের কুমারটুলী এলাকায় বুড়িগঙ্গা নদীর তীরে অবস্থিত। এটি পূর্বে ছিল ঢাকার নবাবদের আবাসিক প্রাসাদ ও জমিদারীর সদর কাচারি। বর্তমানে এটি জাদুঘর হিসাবে ব্যবহৃত হচ্ছে। এর প্রতিষ্ঠাতা নওয়াব আবদুল গনি।তিনি তার পুত্র খাজা আহসানুল্লাহ-র নামানুসারে এর নামকরণ করেন।১৮৫৯ খ্রিষ্টাব্দে আহসান মঞ্জিলের নির্মাণ কাজ শুরু হয়ে ১৮৭২ খ্রিষ্টাব্দে সমাপ্ত হয়। ১৯০৬ খ্রিষ্টাব্দে এখানে এক অনুষ্ঠিত বৈঠকে মুসলিম লীগ প্রতিষ্ঠার সিদ্ধান্ত হয়। আহসান মঞ্জিল কয়েকবার সংস্কার করা হয়েছে। সর্বশেষ সংস্কার করা হয়েছে অতি সম্প্রতি। এখন এটি বাংলাদেশ জাতীয় জাদুঘর কর্তৃক পরিচালিত একটি জাদুঘর।অষ্টাদশ শতাব্দীর মাঝামাঝি সময়ে জামালপুর পরগনার জমিদার শেখ ইনায়েতউল্লাহ আহসান মঞ্জিলের বর্তমান স্থান রংমহল নামে একটি প্রমোদভবন তৈরি করেন। পরবর্তীতে তার পুত্র শেখ মতিউল্লাহ রংমহলটি ফরাসি বণিকদের কাছে বিক্রি করে দেন। বাণিজ্য কুঠি হিসাবে এটি দীর্ঘদিন পরিচিত ছিল। এরপরে ১৮৩০-এ	Ahsan Manzil is located on the banks of the Buriganga River in the Islampur area of old Dhaka, Bangladesh. It used to be the residential palace of the Nawabs of Dhaka and their administrative center. Currently, it is being used as a museum. It was founded by Nawab Abdul Gani, who named it after his son Khaza Ahsanullah. Construction of Ahsan Manzil began in 1859 and was completed in 1872. In 1906, a meeting of the Muslim League was held here. Ahsan Manzil has undergone several renovations, with the most recent one being quite recent. It now serves as a museum managed by the Bangladesh National Museum.In the mid-18th century, Sheikh Enayet Ullah built a new palace in the Kumartuli area of Jamalpur. Later, his son Sheikh Moti Ullah turned it into a French trading post, known as the French Factory. This trading post was well-known for an extended period. In 1830, Nawab Abdul Gani,	1. Lalbag Port. 2. Liberation War Museum

Figure 2. Caption of Ahsan Monjil

Example of Data Augmentation:

Figure 3. Input image for augmentation

Augmented Images

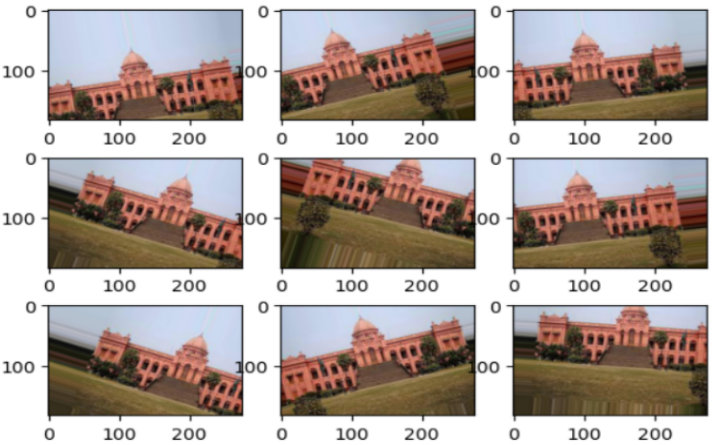

Figure 4. Augmented Images

3.3 Data Portioning

The splitting of data holds paramount importance in the field of machine learning or deep learning, with regard to the evaluation and fortification of a model's performance. The objective of employing machine learning algorithms to construct a model is to understand pattern from real-world data that has never been encountered before and to determine a consistent methodology for its prediction or categorization. The occurrence of data leakage renders a model inadequate in its ability to perform effectively in the face of novel data in the actual world, thereby emphasizing the indispensable nature of implementing precautionary measures to preclude data leakage during the development of a model [13]. In our method, the data has been split into three separate groups called Training,

Validation, and Testing. During the training phase, the model extracts features from the training images and develops an understanding of the same through learning. In our implementation, 80% of the images were utilized for training, while 20% were designated for validation and testing purposes, with an equal distribution of 10% each for validation and testing, which has been shown in Fig. 3. Upon completion of the training phase, we rigorously evaluated the efficacy of our model using the test images.

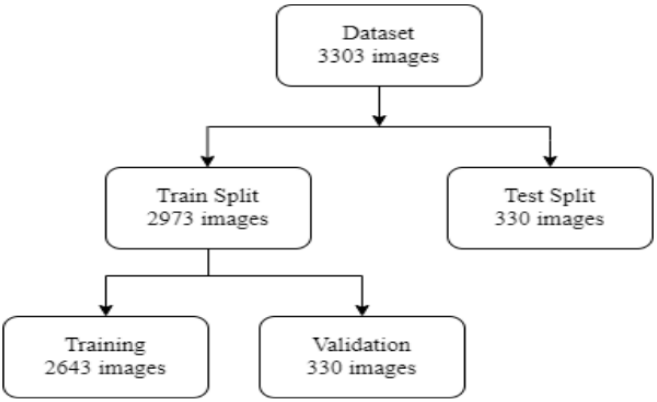

Figure 5. Data Splitting Approach

3.4 Transfer Learning

Transfer learning is a machine learning technique where a model trained on one task is adapted or fine-tuned to perform a different, but related, task. In the context of deep learning, transfer learning involves taking a pre-trained neural network, which has already learned relevant features from a large dataset, and using it as the starting point for a new task or

dataset. Here's a breakdown of the key steps in transfer learning for deep learning:

In the realm of deep learning, transfer learning constitutes a crucial methodology with distinct key steps. The process typically initiates with a pre-training phase, during which a neural network, often a deep convolutional neural network (CNN), undergoes training on a sizable dataset for a specific task, such as image classification. This initial training endows the model with the ability to discern intricate features and patterns within the data. Subsequently, in the transfer phase, the pre-trained model serves as the foundation for a new task— referred to as the target task— which may involve a different dataset or a related yet distinct objective. Rather than commencing training from scratch, the model undergoes fine-tuning. This involves adjusting the model's parameters, with the option to freeze or modify specific layers, allowing the model to retain knowledge gained during pre-training while adapting to the nuances of the new task. Transfer learning's efficacy lies in its capacity to significantly reduce training time, enhance performance on the target task, and prove particularly advantageous in scenarios with limited labeled data. This approach has become a cornerstone in various domains, including computer vision, natural language processing, and speech recognition.

Here I mentioned some Transfer learning model:

a. VGG-16

b. VGG-19

c. ResNet-50

d. MobileNetV1

e. MobileNetV2

f. MobileNetV3

g. InceptionV3

h. Inception-ResNetV2

i. DenseNet

j. Xception

3.5 Proposed Method

3.5.1 Overview

This flowchart illustrates a deep learning pipeline for classifying historical places and generating captions. The process begins with data partitioning into training (1789 images), validation (222 images), and test sets (225 images). The training set undergoes preprocessing, including resizing, padding, augmentation, and normalization, while validation and test sets are only normalized. Preprocessed data is fed into Convolutional Neural Networks (ConvNet), specifically customized versions of MobileNet, ResNet50, and VGG19. These models are trained and optimized as Deep Neural Network (DNN) models to enhance performance. Once trained, the optimized models classify images into 27 historical place categories. Following classification, a caption generator uses a CSV file to automatically generate descriptive captions corresponding to each historical place. Finally, the classified results and generated captions are integrated into a web

application for user interaction and visualization. Overall, this pipeline combines preprocessing, model training, classification, and captioning into an end-to-end system for automated recognition and description of historical sites.

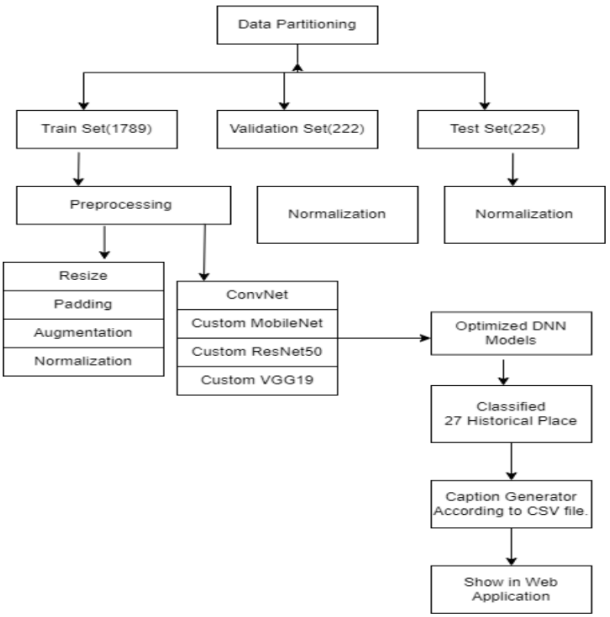

Figure 6. Workflow of the proposed method

3.5.2 Applied Convolutional Neural Network

In this study we employed a CNN (Convolutional Neural Network) based deep learning model. Explain below with code snippet:

a. Created Sequential Model: This creates a linear stack of layers where you can add one layer at a time.

```
model5 = Sequential([
...
])
```

b. Convolutional Layers (First Block):

1. The first Conv2D layer has 128 filters of size 5x5, valid padding, and takes input with shape IMG_SHAPE.

2. Activation function ReLU is applied after convolution.

3. MaxPooling2D layer with a pool size of 2x2 is used for down sampling.

4. Batch Normalization layer is applied to normalize the activations.

```
Conv2D(filters=128, kernel_size=(5, 5), padding='valid', input_shape=IMG_SHAPE),
Activation('relu'),
MaxPooling2D(pool_size=(2, 2)),
BatchNormalization(),
```

Figure 7. Code of ConvNet First Block

c. Convolutional Layers Second Block:

1. The second Conv2D layer has 64 filters of size 3x3, valid padding, and takes input with shape IMG_SHAPE.

2. Activation function ReLU is applied after convolution.

3. MaxPooling2D layer with a pool size of 2x2 is used for down sampling.

4. Batch Normalization layer is applied to normalize the activations

```
Conv2D(filters=64, kernel_size=(3, 3), padding='valid', kernel_regularizer=l2(0.00005)),
Activation('relu'),
MaxPooling2D(pool_size=(2, 2)),
BatchNormalization(),
```

Figure 8. Code of ConvNet Second Block

d. Convolutional Layers (Third Block)

1. The third Conv2D layer has 32 filters of size 3x3, valid padding, and takes input with shape IMG_SHAPE.
2. Activation function ReLU is applied after convolution.
3. MaxPooling2D layer with a pool size of 2x2 is used for down sampling.
4. Batch Normalization layer is applied to normalize the activations.

```
Conv2D(filters=32, kernel_size=(3, 3), padding='valid', kernel_regularizer=l2(0.00005)),
Activation('relu'),
MaxPooling2D(pool_size=(2, 2)),
BatchNormalization(),
```

Figure 9. Code of ConvNet Third Block

e. Flatten Layer

The Flatten layer is used to flatten the 3D output to a 1D tensor, preparing it for the fully connected layer.

f. Fully Connected Layers

1. Dense layer with 256 units and ReLU activation.

2. Dropout layer with a dropout rate of 0.5, which helps prevent overfitting b randomly setting a fraction of input units to zero during training.

3. Final Dense layer with 27 units and softmax activation, suitable for multi-class classification. The output represents the probabilities of the input belonging to each of the 27 classes.

```
Dense(units=256, activation='relu'),
Dropout(0.5),
Dense(units=27, activation='softmax')
])
```

Figure 10. Code of Fully Connected and Dense Layer

3.5.3 Custom MobileNetV2 Architecture

In this study, we employed a CNN-based MobileNetV2 architecture, which was further customized with additional Global Averaging Pooling, dropout and dense layers to enhance

its performance. The integration of the dropout layer played a crucial role as it prevented the neurons in a layer from synchronizing their weight optimization, thus effectively addressing overfitting.

a. Load pre-trained MobileNetV2 Model

1. The code uses the MobileNetV2 model pre-trained on the ImageNet dataset. This model is loaded with pre-trained weights.
2. Weights='imagenet' specifies that the pre-trained weights from the ImageNet dataset should be used.
3. include_top=False means that the top (classification) layer of the MobileNetV2 model is not included. This is done because a new set of dense layers will be added for a custom task.
4. input_shape= (img_width, img_height, 3) defines the input shape of the images expected by the model. In this case, it is an image with img_width and img_height pixels in height and width, and 3 channels (RGB).

```
# Image dimensions and batch size
img_width, img_height = 224, 224
batch_size = 16
# Load pre-trained MobileNetV2 model
base_model1 = MobileNetV2(weights='imagenet', include_top=False, input_shape=(img_width, img_height, 3)
```

Figure 11. Code of Loading MobileNetV2 Model

b. Create Custom MobileNetV2 Model

1. A new Sequential model (model1) is created.
2. The pre-trained MobileNetV2 model is added as the first layer of model1. This allows the new model to leverage the learned features from MobileNetV2.
3. GlobalAveragePooling2D () layer is added. This layer computes the average value of each feature map across the entire spatial dimensions, reducing the spatial dimensions to 1x1. This is a common technique to reduce the number of parameters and flatten the output before the fully connected layers.
4. Dense (512, activation='relu') adds a dense layer with 512 units and ReLU activation function.
5. Dropout (0.5) adds a dropout layer with a dropout rate of 0.5. Dropout is used for regularization, preventing overfitting by randomly dropping out a fraction of neurons during training.
6. Dense (26, activation='softmax') adds the final dense layer with 26 units (assuming it's a 26-class classification task) and a softmax activation function, which is typical for multi-class classification. The output represents the probability distribution over the 26 classes.

```
# Create a new model
model1 = models.Sequential()
model1.add(base_model1)
model1.add(layers.GlobalAveragePooling2D())
model1.add(layers.Dense(512, activation='relu'))
model1.add(layers.Dropout(0.5))
model1.add(layers.Dense(26, activation='softmax'))
```

Figure 12. Code of Custom Layer with MobileNetV2 Model

4 RESULT & DISCUSSION

4.1 MobileNetV2 Model

The Custom MobileNetV2 Model provides Training, Validation, and testing accuracy of 0.9312, 0.9799, and 0.9742 respectively. There the Training, Validation, and Testing Loss are 0.7753, 0.1792, and 0.2351 respectively. Figure 12 shows the result of the MobileNetV2 Model. Figure 13 shows the Training accuracy vs validation accuracy of the MobileNetV2 Model and Fig. 13 shows the Training loss vs Validation loss of the MobileNetV2 Model.

	Accuracy	Loss
Training	0.9312	0.7753
Validation	0.9799	0.1792
Testing	0.9742	0.2351

Figure 13. Result of the MobileNetV2 model

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

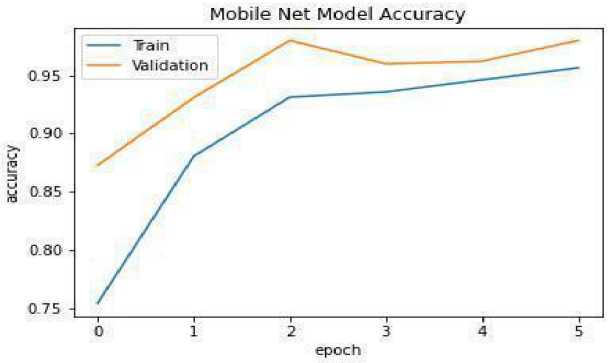

Figure 14. Training accuracy vs validation accuracy of the MobileNetV2 mode

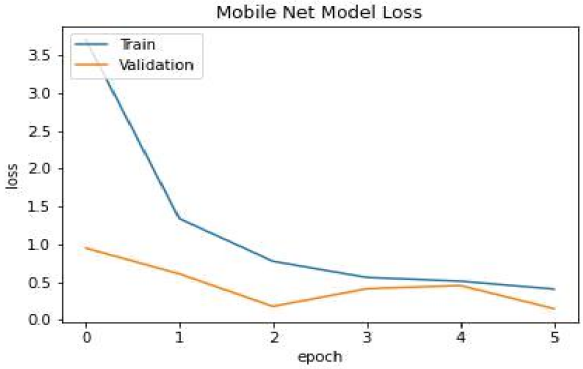

Figure 15. Training loss vs validation loss of the MobileNetV2 mode

4.2 Discussion

Analyzing the above results from Fig. 12-14 we can conclude that except for ResNet50 the other models provide desirable outcomes. There the MobileNetV2 model performs higher in terms of Training Accuracy and VGG19 provides the lowest loss. Table 3 shows the Testing accuracy and loss of the optimized models below. Also, compared to Table I the MobileNetV2 model provides higher accuracy.

Table 3. Testing Accuracy and Loss of the Dnn Models of the Proposed Method

Model Name	Accuracy	Loss
MobileNetV2	0.9742	0.2351
VGG19	0.9613	0.099
ConvNet	0.9128	0.2309
ResNet50	0.8526	0.8412

Source: The authors' own work.

4.3 Image Captioning Software Based on the Proposed Method

4.3.1 Image Captioning Service of the WebApp

We have developed an Image Captioning Web App to create captions for Historical places of Bangladesh. Fig. 15 illustrates the architecture of the Image Captioning software. The App collects ALL Images from the user-end (Frontend) and passes the image data to the server-end (Backend). The server processes the images and calls the DNN model. Finally, the model predicts the subtype of the ALL and passes the result to the user-end. Fig 16 illustrates the home page of the Web Application and Fig. 17 illustrates how the image caption is shown to the user-end.

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

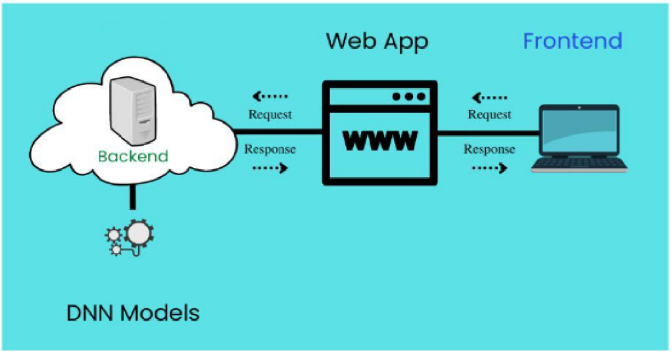

Figure 16. Architecture of the image captioning software

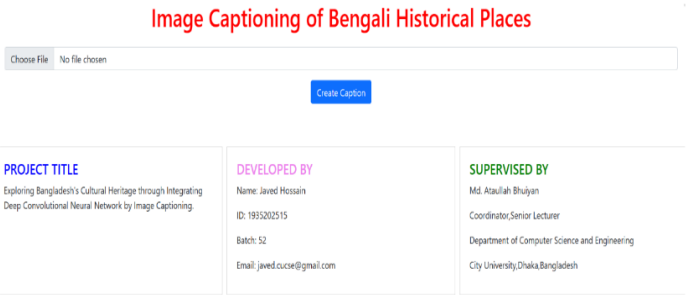

Figure 17. Web Application Home page

4.3.2 Web App Development Environments

a. Programming Language: Python

b. Backend: Flask

c. Frontend: Bootstrap5, HTML, CSS

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

d. Deep Learning Library: Tensorflow-Keras

e. Digital Image Processing Library: OpenCV

4.3.3 Future Work

a. Multimodal Learning

b. Interactive and Immersive Experiences

c. User-Generated Content Integration

d. Enhanced Image Captioning Models

e. Cultural Context Analysis

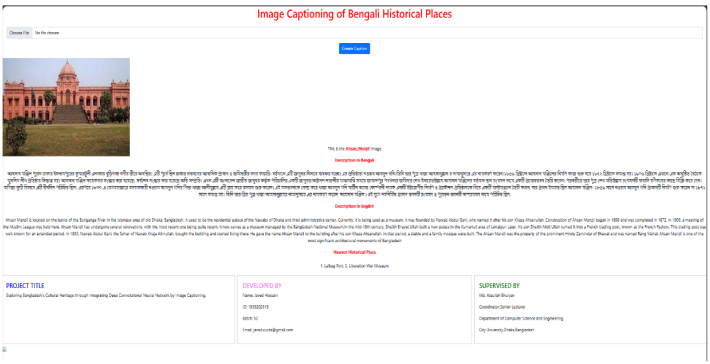

Figure 18. Caption creates From Image

5. CONCLUSION

The application of DNN has demonstrated its potential to significantly improve the accuracy and cultural sensitivity of image captioning systems. By combining the strengths of Convolutional Neural Networks (CNNs) for visual feature extraction and Recurrent Neural Networks (RNNs) for sequential data processing, our model strives to capture the essence of Bangladesh's diverse cultural landscape. The results showcase a marked improvement in the generation of contextually rich captions, enabling a deeper understanding of the visual narratives encapsulated in cultural artifacts.

REFERENCES

Abunadi, I., & Senan, E. M. (2022). Multi-method diagnosis of blood microscopic sample for early detection of acute lymphoblastic leukemia based on deep learning and hybrid techniques. *Sensors*, *22*(4), 1629. https://doi.org/10.3390/s22041629

Afzal, M. K., et al. (2023). Generative image captioning in Urdu using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7719–7731. https://doi.org/10.1007/s12652-023-04584-y

Al-Malki, R. S., & Al-Aama, A. Y. (2023). Arabic captioning for images of clothing using deep learning. *Sensors*, *23*(8), 3783. https://doi.org/10.3390/s23083783

Al-mez Highway, K., & Serte, S. (2020). Improved classification of white blood cells with the generative adversarial network and deep convolutional neural network. *Computational Intelligence and Neuroscience*, *2020*, 6490479. https://doi.org/10.1155/2020/6490479

Al-muzaini, H. A., T. N., & Benhidour, H. (2018). Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. *International Journal of Advanced Computer Science and Applications*, *9*(6). https://doi.org/10.14569/ijacsa.2018.090610

Anselmo, F. C., et al. (2020). Hematological parameters and biochemical markers of iron status in alfa-thalassemia 3.7kb deletion from metropolitan region of Manaus, Amazonas, Brazil. *Mediterranean Journal of Hematology*

and Infectious Diseases, *13*(1), e2021001. https://doi.org/10.4084/mjhid.2021.001

Ayesha, H., et al. (2021). Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, *114*, 107856. https://doi.org/10.1016/j.patcog.2021.107856

Bain, B. J. (2005). Diagnosis from the blood smear. *New England Journal of Medicine*, *353*(5), 498–507. https://doi.org/10.1056/nejmra043442

Bigorra, L., Merino, A., Alférez, S., & Rodellar, J. (2016). Feature analysis and automatic identification of leukemic lineage blast cells and reactive lymphoid cells from peripheral blood cell images. *Journal of Clinical Laboratory Analysis*, *31*(2), e22024. https://doi.org/10.1002/jcla.22024

Chang, Y.-S. (2017). Fine-grained attention for image caption generation. *Multimedia Tools and Applications*, *77*(3), 2959–2971. https://doi.org/10.1007/s11042-017-4593-1

Delgado-Ortet, M., Molina, A., Alférez, S., Rodellar, J., & Merino, A. (2020). A deep learning approach for segmentation of red blood cell images and malaria detection. *Entropy*, *22*(6), 657. https://doi.org/10.3390/e22060657

Duggal, R., Gupta, A., Gupta, R., & Mallick, P. (2017). SD-layer: Stain deconvolutional layer for CNNs in medical microscopic imaging. In *Medical image computing and computer assisted intervention – MICCAI 2017* (pp. 435–443). https://doi.org/10.1007/978-3-319-66179-7_50

Eckardt, J.-N., et al. (2021). Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. *Leukemia*, *36*(1), 111–118. https://doi.org/10.1038/s41375-021-01408-w

Elhassan, T. A., et al. (2023). Classification of atypical white blood cells in acute myeloid leukemia using a two-stage hybrid model based on deep convolutional autoencoder and deep convolutional neural network. *Diagnostics*, *13*(2), 196. https://doi.org/10.3390/diagnostics13020196

Ghaderzadeh, M., Aria, M., Hosseini, A., Asadi, F., Bashash, D., & Abolghasemi, H. (2021). A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images. *International Journal of Intelligent Systems*, *37*(8), 5113–5133. https://doi.org/10.1002/int.22753

Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1–39. https://doi.org/10.1145/3617592

Gupta, A., et al. (2020). GCTI-SN: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images. *Medical Image Analysis*, *65*, 101788. https://doi.org/10.1016/j.media.2020.101788

Gupta, R., Gehlot, S., & Gupta, A. (2022). C-NMC: B-lineage acute lymphoblastic leukemia: A blood cancer dataset. *Medical Engineering & Physics*, *103*, 103793. https://doi.org/10.1016/j.medengphy.2022.103793

Hossain, J., Islam, M. T., & Tusar, M. T. H. K. (2023). Streamlining brain tumor classification with custom transfer learning in MRI images. In *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*. https://doi.org/10.1109/sist58284.2023.10223507

Hosseini, S., et al. (2015). Menstrual blood contains immune cells with inflammatory and anti-inflammatory properties. *Journal of Obstetrics and Gynecology Research*, *41*(11), 1803–1812. https://doi.org/10.1111/jog.12801

Jayavikash, K. P., et al. (2021). Detection of leukemia using machine learning algorithms. *Journal of Physics: Conference Series*, *1916*(1), 012220. https://doi.org/10.1088/1742-6596/1916/1/012220

Jishan, M. A., Mahmud, K. R., Azad, A. K. A., Alam, M. S., & Khan, A. M. (2020). Hybrid deep neural network for Bangla automated image descriptor. *International Journal of Advances in Intelligent Informatics*, *6*(2), 109. https://doi.org/10.26555/ijain.v6i2.499

Labati, R. D., Piuri, V., & Scotti, F. (2011). All-IDB: The acute lymphoblastic leukemia image database for image processing. In *2011 18th IEEE International Conference on Image Processing*. https://doi.org/10.1109/icip.2011.6115881

Madhloom, H. T., Kareem, S. A., Ariffin, H., Zaidan, A. A., Alanazi, H. O., & Zaidan, B. B. (2010). An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold. *Journal of Applied*

Sciences, *10*(11), 959–966. https://doi.org/10.3923/jas.2010.959.966

Mahalakshmi, P., & Fatima, N. S. (2022). Summarization of text and image captioning in information retrieval using deep learning techniques. *IEEE Access*, *10*, 18289–18297. https://doi.org/10.1109/access.2022.3150414

Ming, Y., Hu, N., Fan, C., Feng, F., Zhou, J., & Yu, H. (2022). Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica*, *9*(8), 1339–1365. https://doi.org/10.1109/jas.2022.105734

Mustaqim, T., Fatichah, C., & Suciati, N. (2022). Combination of cross stage partial network and GhostNet with spatial pyramid pooling on YOLOv4 for detection of acute lymphoblastic leukemia subtypes in multi-cell blood microscopic image. *Scientific Journal of Informatics*, *9*(2), 139–148. https://doi.org/10.15294/sji.v9i2.37350

Ondeng, O., Ouma, H., & Akuon, P. (2023). A review of transformer-based approaches for image captioning. *Applied Sciences*, *13*(19), 11103. https://doi.org/10.3390/app131911103

Rahadi, I., Choodoung, M., & Choodoung, A. (2020). Red blood cells and white blood cells detection by image processing. *Journal of Physics: Conference Series*, *1539*(1), 012025. https://doi.org/10.1088/1742-6596/1539/1/012025