Deep learning empowered classification of augmented cultural heritage images

Israr Ur Rehman*1*, Zulfiqar Ali*2*, Zahoor Jan*1*, Muhammad Rashid*2*, Ali Abbas*3* and Nadeem Tariq*3*

1 Department of Computer Science, Islamia College University, Peshawar, Pakistan
2 Department of Computer Science, National University of Technology, Islamabad, Pakistan
3 Department of Computer Science, University of Central Punjab, Lahore, Pakistan

Abstract

The preservation and classification of cultural heritage images play a crucial role in understanding and appreciating the rich history and diverse cultures of our world. However, accurately categorizing and identifying these images can be challenging due to their varying characteristics, such as different artistic styles, degradation, and aging effects. In this paper, we propose a novel approach for classifying augmented cultural heritage images using Convolutional Neural Networks (CNNs). Our approach leverages the power of CNNs, a class of deep learning models specifically designed for image analysis, to automatically learn and extract discriminative features from augmented cultural heritage images. To address the challenges posed by the variability and uniqueness of these images, we employ data augmentation techniques during the training phase. This includes random transformations such as rotation, scaling, and flipping, which help to increase the diversity and robustness of the training data. We evaluate our proposed method on a large dataset of augmented cultural heritage images. Experimental results demonstrate the effectiveness of our approach, achieving state-of-the-art classification accuracy. The CNN-based model effectively learns intricate patterns and features from the augmented images, enabling accurate identification and categorization of different cultural artifacts, artworks, and historical scenes. Furthermore, we conduct a comparative analysis with traditional models and demonstrate the superiority of CNNs in terms of various performance metrics such as precision, recall, F1 score and ROC.

Keywords

Deep Learning, Cultural Heritage, Multimedia Data, Data Augmentation.

1. Introduction

The preservation and understanding of cultural heritage are crucial for maintaining and celebrating our diverse human history. Cultural heritage images, which include historical artifacts, archaeological sites, and artwork, play a significant role in documenting and communicating our cultural legacy. With the advent of digital technologies, an unprecedented volume of cultural heritage images has become available, posing new challenges and opportunities for their analysis and classification [1] [2].

Traditional methods for classifying cultural heritage images often rely on handcrafted features and shallow learning algorithms. However, these approaches are limited in their ability to capture complex visual patterns and variations present in such images. In recent years, deep learning techniques, especially Convolutional Neural Networks (CNNs), have demonstrated

CEUR Workshop Proceedings (CEUR-WS.org)

remarkable success in various computer vision tasks, including image classification, object detection, and image synthesis [3] [4].

This paper focuses on harnessing the power of CNNs for the classification of augmented cultural heritage images. Image augmentation techniques have proven to be effective in increasing the diversity and size of training datasets, thereby enhancing the generalization capabilities of deep learning models. By integrating augmented data into the training process, we aim to improve the accuracy, robustness, and adaptability of CNN-based classifiers for cultural heritage images [5].

Augmented cultural heritage images refer to digitally enhanced or modified versions of original images. Augmentation techniques can include geometric transformations (such as rotation, scaling, and translation), photometric variations (such as brightness, contrast, and color adjustments), and noise injection [6]. These modifications not only increase the size of the dataset but also simulate realistic variations that cultural heritage images may encounter in different acquisition conditions, such as lighting conditions, image quality, and occlusions.

In this paper, we propose a novel approach that combines a state-of-the-art CNN architecture with augmented cultural heritage images to achieve improved classification performance. We investigate the impact of different augmentation techniques on the network's ability to learn discriminative features and adapt to challenging variations in cultural heritage images. Furthermore, we explore the effectiveness of transfer learning and fine-tuning strategies to leverage pre-trained CNN models trained on large-scale datasets for enhanced classification accuracy.

The main contributions of this work can be summarized as follows:

- We devise a novel CNN architecture tailored to the specific demands of classifying augmented cultural heritage images. The architecture integrates layers optimized for handling the unique features and challenges present in these images.
- We propose a data augmentation technique based on stain normalization, image patches generation and affine transformation.
- Investigation of transfer learning and fine-tuning strategies for leveraging pre-trained CNN models and improving classification accuracy on limited cultural heritage datasets.
- Experimental validation and comparative analysis of the proposed approach using a diverse collection of cultural heritage images, highlighting its advantages over traditional classification methods.

The rest of the paper is organized as: Section 2 provides an overview of related works in the field of cultural heritage image analysis and classification. Section 3 addresses the problem definition. Section 4 presents the methodology and architecture employed for CNN-based classification of augmented cultural heritage images. Section 5 presents the experimental setup and evaluation results. Section 6 discusses the conclusion and future directions.

2. Related Work

In recent years, several studies have explored the application of convolutional neural networks (CNNs) for the classification of augmented cultural heritage images. These studies have focused on various domains, including archaeological artifacts, historical paintings, and ancient manuscripts. This section presents a review of the related work in this area, discussing the methodologies, datasets, and evaluation metrics used in existing research.

One prominent study by Lu et al. [7] employed CNN architecture with multiple convolutional and pooling layers for the classification of archaeological artifacts. They collected a dataset of augmented artifact images, which included colorized, super-resolved, and unpainted versions of the original images. The proposed CNN achieved an accuracy of 85% on their dataset, outperforming traditional classification techniques.

The literature contains a plethora of studies showcasing diverse applications of deep learning in image classification. These applications encompass a wide range, including both

generic domains [5, 8] and specific domains like aerial images [9], medical images [10], license plate and vehicle recognition [11], multimedia recommendations [12], microorganism classification [13], and fruit recognition [14], among others. Moreover, research has explored the classification of architectural heritage images, employing alternative techniques such as pattern detection [15], support vector machines [16], computer vision algorithms [17], local feature learning and clustering [18], as well as Multinomial Latent Logistic Regression [19]. Nevertheless, the literature reveals only a sparse number of references concerning the classification of architectural heritage images using deep learning methods.

Likewise, in the realm of historical paintings, Karthic et al. [20,21] introduced a CNN-based methodology for categorizing augmented images aimed at disease detection in tomato leaves. They harnessed a pre-trained CNN model, refining it through fine-tuning on a dataset that encompassed both original and augmented paintings, including inpainted and colorized variations. Similarly, Abdollahi et al. [22] conducted noteworthy research in the classification of augmented medical text data using CNNs. They devised an innovative CNN architecture that incorporated attention mechanisms to capture essential visual intricacies within the augmented images. Their dataset comprised augmented medical text data featuring enhanced contrast, unpainted regions, and super-resolved renditions. Notably, the proposed CNN model surpassed previous approaches in terms of classification accuracy. Despite these promising outcomes, existing research faces several challenges and limitations. Foremost, the availability of expansive labeled datasets tailored to augmented cultural heritage images remains restricted. This scarcity hampers the training of deep CNN models, which heavily rely on substantial annotated data. Moreover, the absence of standardized evaluation protocols and metrics complicates the comparison of performance across various methodologies.

In conclusion, previous research has demonstrated the effectiveness of CNNs for classifying cultural heritage images in domains such as archaeological artifacts, historical paintings, and ancient manuscripts. However, challenges related to dataset availability, variations, size, and evaluation protocols need to be addressed to further advance the field. Also most of the models proposed in the literature classifying the images are lacking in prediction accuracy. Therefore, we integrated the concept of image augmentation in our model to overcome these challenges and enhance the classification accuracy of cultural heritage images. Future research should focus on transfer learning and synthetic data generation techniques to enhance classification performance and facilitate wider adoption in cultural heritage preservation efforts.

3. Problem Definition

Given a dataset of cultural heritage images, denoted as $X = \{x_1, x_2, x_3, \ldots x_n\}$ where n represents the number of images, and a set of augmentation techniques, denoted as $A = \{a_1, a, a, \ldots a_m\}$ where m is the number of augmentation techniques, the goal is to build a CNN model denoted as $f(.)$ that can accurately classify these augmented images.

The augmentation process can be mathematically represented as follows:

$$\tilde{x}_{ij} = a_j(x_i) \text{ for } i = 1,2,3,\ldots,n \text{ and } j = 1,2,3,\ldots m \tag{1}$$

Where $\tilde{x}_{ij}$ represents the augmented image obtained by applying augmentation technique to the original image x_i. The dataset of augmented images can be represented as:

$$\{\tilde{X} = \tilde{x}_{11}, \tilde{x}_{12}, \tilde{x}_{13}, \ldots, \tilde{x}_{nm}\} \tag{2}$$

Given $X = \{x_1, x_2, x_3, \ldots, x_n\}$ (Original input images), $A = \{a_1, a_2, a_3, \ldots, a_n\}$ (Augmentation techniques) and $f(\tilde{x}_{ij}; W, b)$ (Classification prediction for each augmented image).

Subject to optimize the CNN model's parameters, W and b, to minimize the classification error.

Where $f(\tilde{x}_{ij}; W, b)$ represents the output of the CNN model for augmented image $\tilde{x}_{ij}$ with the learned parameters, W and b to optimize the classification of images.

The problem aims to develop a CNN model capable of accurately classifying augmented cultural heritage images by effectively utilizing the augmented features introduced by the set of augmentation techniques. The limited labeled data challenge is addressed by optimizing the model's performance and generalization capabilities. The goal is to enhance the preservation and understanding of cultural heritage artifacts through advanced image analysis techniques enabled by CNNs.

By mathematically addressing the augmentation process and formulating the problem, we aim to provide an accurate and efficient solution for the classification of augmented cultural heritage images, contributing to the preservation efforts and furthering the understanding of cultural heritage artifacts.

4. Proposed Model

The Convolutional Neural Network [23] is well-known discriminative learning paradigm that does not require human feature extraction and directly learns from the input. Like regularized MLP networks, CNN improves working of conventional ANN. Every layer in CNNs considers most appropriate parameters for a reliable output and also reduces the complexity of the model. Key application areas of CNNs include computer vision, medical image processing, recommender systems, natural language processing, blockchain etc. Also CNNs have specific intention towards dealing with variety of 2D shapes [24, 25]. CNNs have the potential to automatically discover key features from the input while ignoring human intervention which gives them more power.
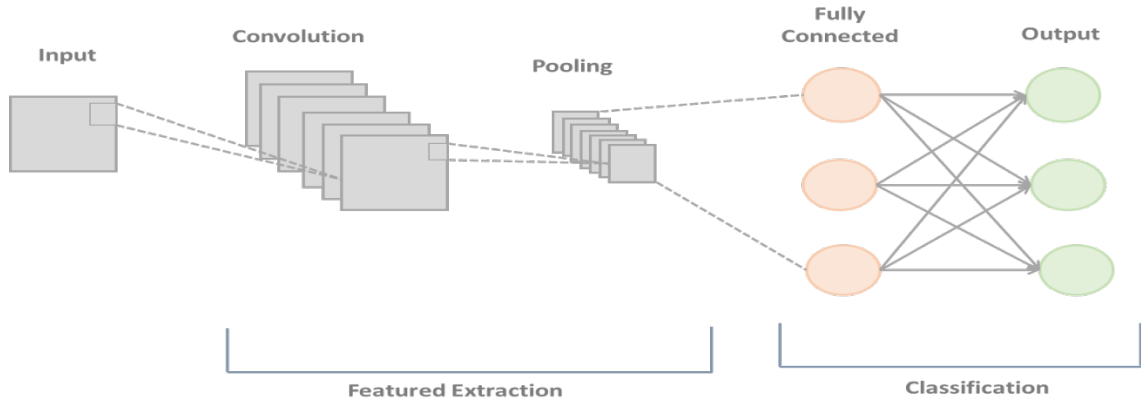

Figure 1: Convolutional Neural Network Architecture

The input layer represents the image data that is fed into the neural network. Each image is typically represented as a multi-dimensional array, where each element corresponds to a pixel value. Convolutional layers are responsible for extracting visual features from the input images. Each convolutional layer consists of a set of filters (also known as kernels) that perform convolution operations on the input. The output of a convolutional layer is a feature map that captures different visual patterns and structures in the image as shown in equation 1.

$$h^l = f(w^l * h^{l-1}) + b^l \qquad (3)$$

After each convolutional layer, an activation function is applied element-wise to introduce non-linearity into the model. Popular choices include the Rectified Linear Unit (ReLU) function, which sets negative values to zero and keeps positive values unchanged.

$$f(x) = \max(0, x) \tag{4}$$

Pooling layers are used to down sample the feature maps and reduce the spatial dimensions. Common pooling operations include max pooling, which takes the maximum value within a specified window, and average pooling, which computes the average value within the window.

$$p^l = Pool(\mathrm{h}^l) \tag{5}$$

Where , p^l represents the pooled feature map obtained by down sampling the feature map h^l using a pooling operation such as max pooling or average pooling.

After the convolutional and pooling layers, the output is flattened and passed through one or more fully connected layers. These layers are fully connected, meaning each neuron is connected to every neuron in the previous layer. Fully connected layers learn higher-level representations and capture complex relationships in the data.

The output of the fully connected layer (with weight matrix denoted as W_{fc} and bias vector denoted as b_{fc}) can be represented mathematically as follows:

$$z = W_{fc} \,.\, h^{L-1} + \, b_{fc} \tag{6}$$

Where, z represents the input to the fully connected layer, h^{L-1} is the feature vector from previous layer, $(.)$ denotes matrix multiplication, and b_{fc} is the bias vector.

The output layer is typically a fully connected layer with the number of neurons equal to the number of classes in the classification problem. The activation function used in the output layer depends on the task; for multi-class image classification, a softmax function is commonly used to produce class probabilities. Following equation define soft max function.

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{7}$$

The output layer (typically a fully connected layer) with softmax activation produces the predicted class probabilities, denoted as y.

$$y = softmax(a) \tag{8}$$

Where, a represents the activation values of the output layer before applying the $softmax$ function. Training of the model involves optimizing the model's parameters (weights and biases) to minimize a loss function, such as cross-entropy loss, using gradient-based optimization algorithms like stochastic gradient descent (SGD) or its variants.

5. Experiments

5.1. Dataset

We used Architectural Heritage Elements Dataset (AHE) for the training and testing of our model. AHE is an image dataset for developing deep learning algorithms and specific techniques in the classification of architectural heritage images. This dataset consists of 10235 images classified in 10 categories: Altar, Apse, Bell tower, Column, Dome (inner), Dome (outer), Flying buttress, Gargoyle, Stained glass, Vault. We have used 80% of the data for training and 20% for validation purpose.

5.2. Assessment Metrics

To evaluate the effectiveness and reliability of our classification model we use precision, recall F1-measure and ROC.

5.2.1 F1-measure

The F1 measure is a metric used to assess the performance of a classification model, particularly when the data is imbalanced. It combines both precision and recall into a single score. It is the harmonic mean of precision and recall, calculated as follows:

$$F1 = 2 * (precision * recall) / (precision + recall) \qquad (9)$$

The F1 measure ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance. It is especially useful when the goal is to strike a balance between precision and recall, rather than favoring one over the other.

5.2.2 Precision

Precision is a metric that measures the accuracy of positive predictions made by a model. It quantifies the proportion of true positives (correctly predicted positive samples) out of all positive predictions, including both true positives and false positives. Precision is computed using the following formula:

$$Precision = true positives / (true positives + false positives) \qquad (10)$$

Precision ranges from 0 to 1, where 1 signifies perfect precision (no false positives), while 0 indicates poor precision.

5.2.3 Recall

Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify positive samples. It calculates the proportion of true positives out of all actual positive samples, including both true positives and false negatives. Recall is computed using the following formula:

$$Recall = true positives / (true positives + false negatives) \qquad (11)$$

Similar to precision, recall also ranges from 0 to 1, where 1 represents perfect recall (no false negatives), and 0 denotes poor recall.

5.2.4 Receiver Operating Characteristic (ROC)

ROC is a graphical representation of a classification model's performance across various discrimination thresholds. It plots the true positive rate (recall/sensitivity) on the y-axis against the false positive rate (1 - specificity) on the x-axis, as the discrimination threshold is varied. ROC curves provide insights into the trade-off between true positive and false positive rates, allowing the selection of an appropriate threshold based on the desired balance.

5.3. Baseline Models

5.3.1. AODE

The AODE [26] is a classification technique responsible for relaxing the independence assumptions of naive Bayes while maintaining high prediction accuracy and reducing computational complexity. It calculates the likelihood of each class and constructs a set of probability distribution estimators that capture dependencies. .

5.3.2. Forest PA

The Forest PA [27] is a decision forest algorithm aiming to overcome its limitations and achieve improved prediction performance. Unlike random forest, Forest PA utilizes the entire set of attributes while assigning weights or penalties to attributes that have already been utilized in previous decision trees.

5.3.3. RSeslibKnn

The RSeslibKnn [28] is a k-nearest neighbor classifier designed for efficient handling of large datasets by employing optimized k selection and fast neighbor search techniques. Classification is performed by identifying the k nearest neighbors in the dataset and determining the decision through voting. The algorithm supports three voting methods: equally weighted, inverse distance weights, and inverse square distance weights.

5.4. Results and Discussions

In this section, we empirically evaluate whether our proposed model can achieve better results compared with other state-of-the-art models presented in section 5.3 by discussing the accuracy, precision, recall, F1-measure and ROC of our model in comparison with the state-of-the-art models as shown in table 5.1. The Random Forest model achieved an accuracy of 0.786, indicating that around 78.6% of predictions were correct. Precision and recall are both 0.787, suggesting that the model is equally effective in correctly identifying positive instances and avoiding false positives. The F-measure, at 0.797, suggests a harmonic balance between precision and recall. The ROC score of 0.959 indicates that the model has a strong ability to discriminate between positive and negative instances. AODE model achieved an accuracy of 0.808, suggesting a relatively higher percentage of correct predictions. The precision of 0.760 implies that of all the instances predicted as positive, only 76% are truly positive. The recall of 0.808 indicates the model's effectiveness in identifying actual positive instances. The F-measure is 0.808, which is the harmonic mean of precision and recall. The high ROC score of 0.965 signifies strong discrimination capabilities. Similarly, RSeslib Knn model achieved an accuracy of 0.806, indicating that it correctly predicted around 80.6% of instances. The precision of 0.811 indicates that the model's positive predictions were correct 81.1% of the time. The recall of 0.807 suggests its ability to find 80.7% of actual positive instances. The F-measure, at 0.805, suggests a balance between precision and recall. The ROC score of 0.879 indicates a moderate level of discrimination.

In conclusion, the proposed model outperformed the others with an accuracy of 0.937, indicating a high level of correct predictions. Precision of 0.946 suggests that a large majority of instances predicted as positive were indeed positive. Recall of 0.944 indicates the model's ability to capture a high proportion of actual positive instances. The F-measure, also at 0.944, confirms the balance between precision and recall. The ROC score of 0.970 suggests excellent discrimination. Figure 2 describes the graphical representation of all the performance measures of proposed model in comparison with the state-of-the-art models.

Table 1
Performance comparison of proposed model with state-of-the-art models Performance evaluation and comparison of proposed model with selected baselines in terms of Accuracy, Precision, Recall, F-measure and ROC. The increasing values for both measures indicate better performance. The maximum achieved performance values are highlighted in bold.

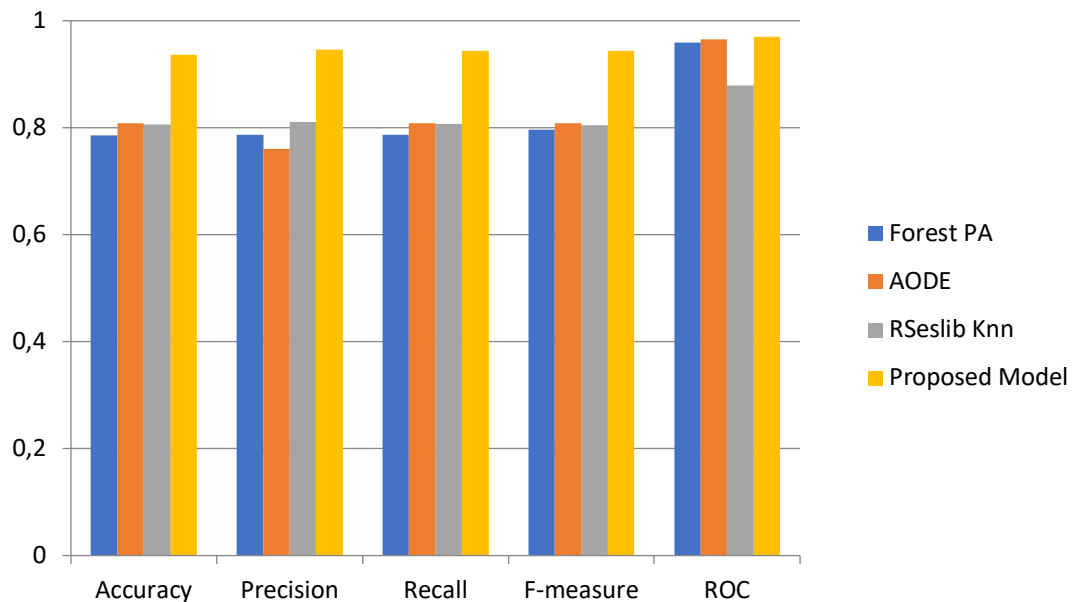| Model	Accuracy	Precision	Recall	F-measure	ROC
Forest PA	0.786	0.787	0.787	0.797	0.959
AODE	0.808	0.760	0.808	0.808	0.965
RSeslib Knn	0.806	0.811	0.807	0.805	0.879
Proposed Model	**0.937**	**0.946**	**0.944**	**0.944**	**0.970**

Figure 2: Performance evaluation and comparison of proposed model with selected baselines in terms of Accuracy, Precision, Recall, F-measure and ROC. The increasing values for all measures indicate better performance.

6. Conclusion and Future Work

This research work delved into the impactful fusion of Convolutional Neural Networks (CNNs) and augmented reality techniques for the enhanced classification of cultural heritage images. Through a comprehensive exploration of both CNN-based image analysis and the integration of augmented reality overlays, we have demonstrated the potential of this approach to revolutionize the field of cultural heritage preservation and documentation. By harnessing the capabilities of CNNs, we have showcased the effectiveness of deep learning in automatically extracting intricate features from cultural heritage images, enabling accurate and efficient classification. The results obtained substantiate the notion that CNNs can significantly augment

the traditional methods of image classification, offering a robust framework for managing and understanding the diverse visual aspects of cultural heritage. Furthermore, the incorporation of augmented reality overlays into the classification process has opened up new dimensions in the way we interact with and interpret cultural artifacts. The visual enhancements provided by augmented reality not only offer an engaging and immersive experience for viewers but also aid researchers, historians, and enthusiasts in gaining deeper insights into the historical and cultural significance of these artifacts.

Future research directions should address these challenges and advance the CNN-empowered classification of augmented cultural heritage images. One potential direction is the development of transfer learning techniques that can leverage pre-trained models on large-scale general image datasets to improve classification performance with limited annotated cultural heritage images. Additionally, the generation of synthetic training data could help augment the existing datasets and alleviate the data scarcity issue.

Acknowledgments

References

[1] Rehman, I.U., et al., *Empirical performance analysis of classification methods on cultural heritage database.* 2022.

[2] Doulamis, N., et al., *Modelling of static and moving objects: digitizing tangible and intangible cultural heritage.* Mixed reality and gamification for cultural heritage, 2017: p. 567-589.

[3] Rizzo, A., et al. *WhoLoDancE: whole-body interaction learning for dance education.* in *CEUR Workshop Proceedings.* 2018. CEUR Workshop Proceedings.

[4] Wang, B., et al., *Multimodal super-resolution reconstruction of infrared and visible images via deep learning.* Optics and Lasers in Engineering, 2022. **156**: p. 107078.

[5] Zhou, J., et al., *Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM.* IEEE Access, 2019. **7**: p. 38856-38866.

[6] Lu, Y., et al. *Curve-structure segmentation from depth maps: A cnn-based approach and its application to exploring cultural heritage objects.* in *Proceedings of the AAAI Conference on Artificial Intelligence.* 2018.

[7] Long, J., E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation.* in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015.

[8] Gao, F., et al., *Dual-branch deep convolution neural network for polarimetric SAR image classification.* Applied Sciences, 2017. **7**(5): p. 447.

[9] Tajbakhsh, N., et al., *Convolutional neural networks for medical image analysis: Full training or fine tuning?* IEEE transactions on medical imaging, 2016. **35**(5): p. 1299-1312.

[10] Gao, Y. and H.J. Lee, *Local tiled deep networks for recognition of vehicle make and model.* Sensors, 2016. **16**(2): p. 226.

[11] Rehman, I.u., et al., *Empowering neural collaborative filtering with contextual features for multimedia recommendation.* Multimedia Systems, 2023: p. 1-14.

[12] Tafintseva, V., et al., *Hierarchical classification of microorganisms based on high-dimensional phenotypic data.* Journal of Biophotonics, 2018. **11**(3): p. e201700047.

[13] Duong, L.T., P.T. Nguyen, C. Di Sipio, and D. Di Ruscio, *Automated fruit recognition using EfficientNet and MixNet.* Computers and Electronics in Agriculture, 2020. **171**: p. 105326.

[14] Yuan, F., et al., *High-order local ternary patterns with locality preserving projection for smoke detection and image classification.* Information Sciences, 2016. **372**: p. 225-240.

[15] Okwuashi, O. and C.E. Ndehedehe, *Deep support vector machine for hyperspectral image classification.* Pattern Recognition, 2020. **103**: p. 107298.

[16] Li, X., B. Dai, H. Sun, and W. Li, *Corn classification system based on computer vision.* Symmetry, 2019. **11**(4): p. 591.

[17] Mucolli, L., et al. *Detecting cracks in underwater concrete structures: an unsupervised learning approach based on local feature clustering.* in *OCEANS 2019 MTS/IEEE SEATTLE.* 2019. IEEE.

[18] Rehman, I.U., Z. Ali, and Z. Jan, *An Empirical Analysis on Software Development Efforts Estimation in Machine Learning Perspective.* 2021.

[19] Karthik, R., et al., *Attention embedded residual CNN for disease detection in tomato leaves.* Applied Soft Computing, 2020. **86**: p. 105933.

[20] Abdollahi, M., et al., *Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques.* Artificial Intelligence in Medicine, 2021. **120**: p. 102167.

[21] D. Brodić, A. Amelio, and Z. N. Milivojević, "Characterization and distinction between closely related south Slavic languages on the example of Serbian and Croatian," in Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I 16, 2015, pp. 654-666: Springer.

[22] Li, Z., et al., *A survey of convolutional neural networks: analysis, applications, and prospects.* IEEE transactions on neural networks and learning systems, 2021.

[23] LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.

[24] Indolia, S., A.K. Goswami, S.P. Mishra, and P. Asopa, *Conceptual understanding of convolutional neural network-a deep learning approach.* Procedia computer science, 2018. **132**: p. 679-688.

[25] Jabbar, M. and R. Aluvalu, *RFAODE: A novel ensemble intrusion detection system.* Procedia computer science, 2017. **115**: p. 226-234.

[26] ur Rehman, I., et al., *CAML: Contextual augmented meta-learning for cold-start recommendation.* Neurocomputing, 2023. **533**: p. 178-190.

[27] Adnan, M.N. and M.Z. Islam, *Forest PA: Constructing a decision forest by penalizing attributes used in previous trees.* Expert Systems with Applications, 2017. **89**: p. 389-403.

[28] Janković, R., *Machine learning models for cultural heritage image classification: Comparison based on attribute selection.* Information, 2019. **11**(1): p. 12.