**DECIPHERING TAMIL CHARACTERS FROM ANCIENT PALM LEAF MANUSCRIPTS**

## Ms.J.Juslin Sega1, Dr.J.Shiny Duela2

## 1 Assistant Professor, Computer Science and Engineering, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,RAMAPURAM,CHENNAI.

#### 2 Associate Professor, Computer Science and Engineering, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,RAMAPURAM,CHENNAI.

#### **1**[***rjsega263@gmail.com***](mailto:rjsega263@gmail.com)**,** [**2shinyduj@srmist.edu.in**](mailto:2shinyduj@srmist.edu.in)**,**

ABSTRACT

Tamil is one of the oldest languages in the world, with a rich cultural and literary heritage that is conserved in both written and oral forms. Palm leaves were used to write manuscripts and they have been used as a medium to conserve knowledge such as in the field of astronomy, medicine, art, literature etc. There remains a growing need for transcription and digitizing palm leaf manuscripts and recognizing Tamil cursive characters remains a challenging task. This study uses Convolutional Neural Network (CNN) technique to train the characters and characteristics present in the palm leaf, allowing it to significantly perform classification of palm leaf characters in the training session. Background noise were removed during preprocessing phase. Connected component Analysis is then used for segmentation which is a technique used in image processing to identify connected regions in a binary image. Then we perform feature processing that includes text line spacing, spacing without obstacle, and spacing with an obstacle. Finally, datasets is fitted into the model for final classification of the characters. Finally, experiments performed on the collected characters is used to obtain the overall performance of the CNN model and it is calculated with the help of accuracy, recall, precision etc.

1. **INTRODUCTION**

Tamil is a language with a rich literary heritage and is recognized as one of the oldest in the world. In ancient times, poets used Palm leaves, particularly in Tamil Nadu, to conceal information. The ancient literature includes masterpieces, such as Sangam literature, Vaishnava, Saiva, medicinal works, gastronomy, astrology, Vaastu, gems, music, dance, and theatre, as well as Siddha. There has been growing interest among academics in the past decade to preserve ancient medical texts in Tamil and their value. To preserve medical materials, numerous scholars have generated conserved old medical writings in Tamil, such as those by saints like Agathiyar, which have undergone the first phase of a digitalization process. Nearly 10,000 manuscripts have been successfully scanned. For digitizing historical documents, apps that identify handwritten characters have used three key methods: statistical, structural or syntactic, and neural network-based techniques

**2.LITERATURE REVIEW**

[1] A novel ETEDL-THDR technique was developed specifically for the identification of tropical hazardous chemicals (THCs). This approach integrates various essential elements, such as MobileNet feature extraction, BiGRU (Bidirectional Gated Recurrent Unit) recognition, and WSO (Weighted Sum Optimization) for hyperparameter tuning. The WSO algorithm effectively adjusts the hyperparameters of the BiGRU model to optimize recognition accuracy. Rigorous experimental studies were conducted to validate the enhanced performance of the ETEDL-THDR model. Comparative analysis clearly illustrates the superior performance of the ETEDL-THDR approach compared to recent methodologies. Future research directions may involve exploring an ensemble approach with three deep learning-based fusion models to further elevate recognition outcomes.

[2] Character classification, particularly for identifying Old English characters in the Beowulf manuscript, was the focus of our study. We developed a convolutional neural network (CNN) model specifically tailored for this task. Our approach involved training and testing the model using the dataset from the Beowulf manuscript. Additionally, we conducted comparative analyses with various machine learning (ML) models including support vector machines (SVM), nearest neighbors (KNN), decision trees (DT), random forests (RF), and XGBoost. To extract features from the Beowulf manuscript character images, we utilized pretrained models such as VGG16, MobileNet, and ResNet50. These features were then used to train the SVM, KNN, DT, RF, and XGBoost models, followed by testing on our Beowulf test dataset. Recognition accuracy results were recorded, and model performance was evaluated based on recall, precision, and F1 score metrics. Each model's classification performance was visualized using ROC curves (Figure 9). To enhance the dataset, we augmented the Beowulf manuscript character images once, twice, and three times. Our proposed CNN model exhibited superior performance compared to other ML models, achieving the highest accuracy of 98.86% with threefold resampling. Furthermore, we evaluated the CNN model using the MNIST handwritten digits dataset, achieving a benchmark recognition accuracy of 99.03%

[3] The median filter is utilized to reduce noise by replacing each pixel's value with the median value within a specified window. One key limitation of using a 2D adaptive median filter is the loss of edge information that occurs when images undergo this filtering process. Similarly, the primary drawback of employing a 2D adaptive Wiener filter is its effectiveness being limited to scenarios where the variance of noise components in the image is relatively low. Another significant issue arises with the 2D adaptive log filter, as it tends to degrade the image quality. As noise components are minimized through this filter, the image quality diminishes progressively.

[4] This paper presents machine learning techniques for recognizing Urdu characters. Initially, datasets were collected for Urdu characters. Basic techniques such as the confusion matrix, ROC curve, and K-fold cross-validation are performed initially. It is followed by preprocessing and segmentation, which provide the character skeleton.

Among SVM, SMO, MLP, and simple logistic classifiers, SVM (Support Vector Machine) provides a high accuracy of over 98%.

[5] This paper introduces a novel approach to character recognition using feature fusion within a machine learning framework. The primary emphasis of this study revolves around two specific classifiers: neural networks and least squares support vector machines (LSSVM). Additionally, the paper delves into the structural aspects of a character recognition system grounded in machine learning principles. Furthermore, the document explores the merits and drawbacks associated with character recognition technologies and proposes a method that integrates feature fusion and machine learning for enhanced performance.

[6] This paper describes the process of building a Gujarati poetry corpus by collecting and curating poems from various poets in Gujarati literature. Each word within the corpus is then cross-referenced with a manually annotated metadata set that categorizes words according to the concept of "Rasa," aligning with the traditional Indian theory of "Navarasa." By applying this method to each poem, a counter value is computed, representing the predominant emotional theme ranging from 0 to 8. This value is then compared against a dictionary named 'ras' to determine the specific 'Rasa' associated with the poem, classifying it according to the principles of "Navarasa." This methodology efficiently extracts metadata from Gujarati poetry, encompassing a wide spectrum of emotional expressions inherent in the poems. The authors have endeavored to develop an automated system capable of categorizing poems based on their emotional content using the traditional Indian framework of 'Navarasa.'

[9] This study presents the challenges of character recognition in palm leaf manuscripts and their solutions. Convolutional Neural Network (CNN) is typically addressed, which is then followed by pattern recognition using OCR and textual processing analysis. Various optical character recognition (OCR) techniques are employed and compared, out of which the Tamil optical character recognition system and the recognition of CAPTCHA characters by supervised machine learning algorithms have an efficiency of 99%. These findings can be applied to palm leaf manuscripts of different languages as a kickstart to the complex character and text recognition problems.

[10] The solution for Tamil character recognition involves the implementation of a ResNet-50- based system, encompassing a database, an algorithm, and an application. The dataset comprises over 15,000 images, each with dimensions of 128x128 pixels. Notably, this proposed Tamil character recognition system attains an impressive accuracy rate of 96%.

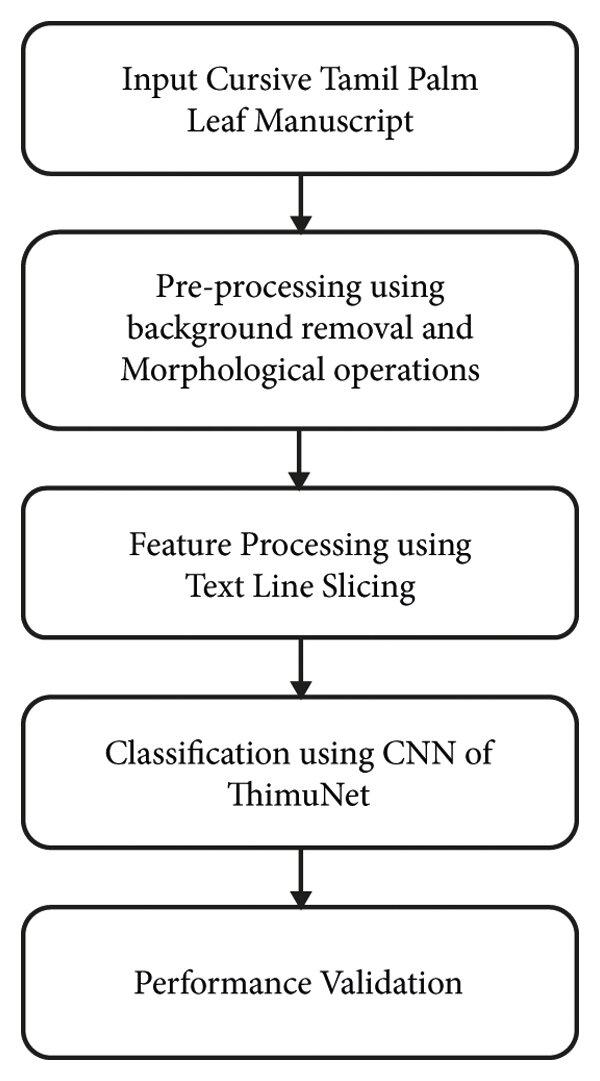
The system's potential lies in establishing a robust handwritten documentation and digitization platform. Moreover, there's potential to expand into Tamil word recognition, integrating with an open-source database for all 256 Tamil characters. ResNet incorporates an "identity shortcut connection" layer, allowing it to bypass redundant layers and reuse activation layers, thus preventing excessive complexity. Rigorous analysis methods ensure the system's efficiency and accuracy. By effectively addressing the limitations of previous methods, this working model consistently achieves a remarkable accuracy rate of 96%, surpassing the performance of alternative systems.

Ali and Joseph [8] developed a CNN perfect for dispensing real-time input pictures including Malayalam characters and the job of segmenting words and typescripts from an image and attractiveness prediction using the CNN model. The model's feature extraction process for digitizing the Malayalam script seamlessly incorporates convolutional neural network (CNN) techniques. This intricate procedure, encompassing recognition of 36 consonants and 13 vowels, unfolds progressively, achieving an outstanding accuracy of 97.26% on the training dataset.

In alignment with the insights shared by Narenthiran and Ravichandran [13], historical India bore witness to the documentation of diverse knowledge systems on palm leaf scrolls. The repositories of ancient wisdom predominantly expressed their contents through an array of characters from the Sanskrit language. Traditional knowledge, in accordance with Devika and Vijayakumar [12], aids in establishing lasting relationships between people and nature. This study aimed to distinguish between the contents of palm leaf documents that can be digitised, the specifics for digitization, and the various methods of palm leaf document scanning

**3. PROPOSED METHODOLOGY**

* 1. **Dataset:**

Datasets for Tamil leaf manuscripts were took from online and stored to fetch the characters. Totally, 100 samples of palm leaf manuscripts were collected. Each bundle of leaves is usually tied together with cord threads through two holes pierced through the entire manuscript by the insertion of bamboo strips. The resultant bundle is completed by adding the heavy wooden covers on either side of the leaves, also tied by the cords or wrapped with a soft textile cloth.

**3.2 Background normalization / Background removal:**

Pre-processing using background removal and Morphological operations

Historical documents typically encounter two types of issues. The first problem arises when the original document is in a state of decay or deterioration, and the second issue occurs when the document is converted into a digital format and has an uneven background. However, there are enhancement techniques available that can help to improve the quality of the image, particularly for low-contrast images. These techniques are effective in reducing uneven backgrounds and can facilitate the extraction of text from historical documents.

The process of pre-enhancement is applied to an input image I by using a linear function that stretches the image's grey level to its full dynamic range through contrast stretching. The output of this pre-enhancement process is depicted in the figure. The brightness of the input image is modified through this process, resulting in the increased distinction between the textual pixels and the background, which helps restore slightly faded text sections. Pre-processing transforms brighter pixels into even brighter ones, which may increase noise slightly, but it is necessary for preserving slightly faded text in both DIBCO and palm leaf documents. The proposed method effectively removes additional noise that may arise due to this change. The enhanced image IE is utilized to calculate the gradient image Gd in subsequent steps.

[Start] --> [Input Image I] --> [Pre-enhancement using contrast stretching] --> [Enhanced Image IE] --> [Calculate Gradient Image Gd] --> [Noise Removal] --> [Binarization] --> [Output Result]

**3.3 Morphological Operations**

Opening is commonly used in image processing applications such as object recognition and feature extraction, where it can help to eliminate unwanted background noise and improve the accuracy of the analysis. It can also be useful in preprocessing steps for optical character recognition (OCR), where it can help to improve the quality and legibility of the text.

In mathematical notation, opening can be expressed as:

A ⊖ B ⊕ B

where A is the original image, ⊖ denotes erosion, ⊕ denotes dilation, and B is a structuring element that defines the shape and size of the objects to be eroded and dilated.

Grayscale morphology is a variant of morphological image processing that works with grayscale images, rather than binary images. It involves the same basic operations as binary morphology, including erosion, dilation, opening, and closing, but the structuring elements used in these operations are grayscale, rather than binary.

In grayscale morphology, the value of each pixel in the structuring element determines its weight or contribution to the output pixel value. This allows for more precise control over the shape and size of the structuring element, and enables more accurate processing of images with complex or continuous intensity variations.

Grayscale morphology is commonly used in image processing applications such as edge detection, image segmentation, and feature extraction. It can also be useful in applications such as medical imaging and remote sensing, where it is important to analyse images with a high degree of accuracy and detail.

**4. FEATURE PROCESSING**

The connected element Analysis in cursive Tamil win flake scripts is a Herculean task, and it influences till the end of the character recognition process. Connected element Analysis is applied to the pre-processed double palm flake text images to member the text lines. e new way of approach, in- text line segmentation of Tamil win flake images is to determine whether the handicap is present between the text lines. Whenever the strokes of the character exceed from the text zone and extend in the space between the lines, also it's considered to be an handicap in this case.

**4.1 Connected Component Analysis**

Connected element analysis is a fashion used in image processing to identify and label the individual connected regions, or factors, in a double image. In a double image, the focus pixels (generally representing the object of interest) are set to one, and the background snaps- rails are set to zero. In connected element analysis, the algorithm identifies all the groups of focus pixels that are connected to each other, and assigns them a unique marker or identifier. The performing labelled image can also be used for further processing, analogous as character recognition or object discovery. Connected element analysis can be per- formed using various algorithms, analogous as the two- pass algorithm or the depth-first quest algo- rithm. It's generally used in operations analogous as OCR (optical character recognition), where it can be used to member individual characters in a text image

⮚ First, the image is binarized by applying a thresholding operation to convert it into a double image. Let I(x, y) be the input image and B( x, y) be the binarized image, also the thresholding operation can be represented as B( x, y) = { 1 if I( x, y)> T{ 0 if I( x, y) ≤ T

⮚ Next, the connected factors in the double image are linked. This is done by assigning a unique marker to each group of connected pixels in the double image. The labels are supporter integers starting from 1 and incremented for each new connected element. The process can be represented as L( x, y) = { 0 if B( x, y) = 0{ L( p) if B( x, y) = 1 and p is a neighbour of( x, y){ new marker if B( x, y) = 1 and p is not a neighbour of( x, y) where L( x, y) is the marker assigned to the pixel( x, y), p is a neighbour of( x, y), and new marker is a unique marker not previously assigned.

⮚ ultimately, the connected factors are segmented by lodging the pixels corresponding to each marker. This can be represented as = {(x, y)| L( x, y) = k}where S\_k is the set of snaps- rails belonging to the k- th connected element.

**5. CONVOLUTIONAL NEURAL NETWORKS**

A common Deep Learning framework used for image recognition and classification tasks is the Convolutional Neural Network (CNN). It consists of several layers, including convolutional, pooling, and fully connected layers. The Convolutional layer uses filters to extract features from the input image, while the Pooling layer reduces computation by taking the image. Finally, the fully connected layer makes the final prediction. The network finds the best filter using multi-scattering and gradient descent.

**5.1 CNN architecture**

Convolutional Layer (CL) The output of the convolution is computed by a convolution operation involving a filter on the input image and computing the dot product between the filter and the pixels present in the input. The formula for calculating the convolution result of a single filter applied to the input image can be expressed as follows:

Output argument = (Filter \* Input Image) + Binary

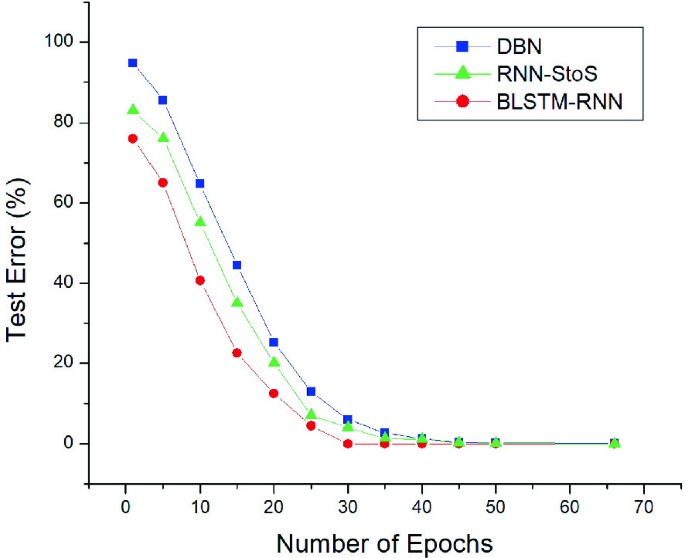
where '\*' stands for the convolution operation, 'Filter' is the CL learning parameter, 'Input Image' is the input image, 'Bias' is the scalar value added to the bias of each output element. The output produced by one filter is a 2D feature map, and the convolutional output of all CLs is the 2D feature map produced by all filters applied to the input image.

The convolutional layer in CNN uses a set of filters (also called kernel or feature map) that can be learned in the input image to extract local features or patterns. Sliding filter on the input image in sliding window mode, do the dot product between their weight and the corresponding pixel value in the input image. The result, called a feature map, shows the presence of local features such as edges, corners, and parts in the input image. The mathematical operations performed by the friction layer can be described as follows:

**y(i,j,k) = ∑∑∑ x(m,n,l) \* w(i-m+1,j-n+1,l,k)**

where y (i, j, k) is the output feature map at location (i, j) and pixel (m, n) at location x (m, n, l) for the whale filter and channel for lath, and w (i-m ). + 1, j-n + 1, l, k) is the weight of the filter at the location (i-m + 1, j-n + 1) and the lath channel and whale filter.

Pooling can be done using different methods such as peak pooling or average pooling where the maximum or average value is stored in a sliding window. Pooling helps reduce the number of parameters in the set, making changes to the input image more efficient and reliable. The mathematical operations performed by the pool layer can be defined as:

**y(i,j,k) = f({x(i',j',k) | i'** ∈ **[i*S, i*S+H), j'** ∈ **[j*S, j*S+W)})**

where y(i,j,k) is the output value at location (i,j) and x(i',j',k) is the input value at location (i',j') for the whale channel and S is the step for whale channel, H and W are the height and width of the pool window, respectively. The function f (.) can be either a maximum or an average pool.

The output of a convolutional layer in a neural network (CNN) can be expressed mathematically as:

**h(i,j,k) = f(∑∑∑ x(m,n,l) \* w(i-m+1,j-n+1,l,k) + b(k))**

where h(i,j,k) is the output feature map atposition (i,j) and for the whale filter x(m,n,l) is the input value at position (m,n) and lath is the channel. , w(i-m+1, j-n + 1, l, k) is the condition of the filter (i-m + 1, j-n + 1) and for lath channels and whale filters, b(k) is the bias term. for the whale filter, f (.) Activation function.

**6. Results and Discussion**

As mentioned earlier, the goal of this work is to create a CNN model that is more efficient than existing models for the received course data set. Proposal for neural construction is

PyTorch7 is used as a Python-based framework for network architecture. Several state-of-the-art models such as LeNet5, ResNet (18/34/50), Alex Net, DenseNet121, Inception Net v3 and others have been evaluated in this database to conduct comparative studies. For all these tests, an Intel Core i3 processor, 16 GB of RAM and 4 GB of internal memory and an NVIDIA graphics card use 768 CUBA cores.

**References**

[1] D. Ganapathy, “Preserving India’s palm Leaf Manuscripts for the Future,” WAGLOBAL, Kerala, India, 2016.

[2] N. S. Panyam, V. L. T.R., R. Krishnan, and K. R. N.V., “Modeling of palm leaf character recognition system using transform based techniques,” Pattern Recognition Letters, vol. 84, pp. 29–34, 2016.

[3] K. P. Geena and G. Raju, “View-based feature extraction and classification approach to Malayalam palm leaf document image,” International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 5, pp. 264–267, 2014.

[4] R. Chamchong and C. C. Fung, “A framework for the selection of binarization techniques on palm leaf manuscripts using support vector machine,” Advances in Decision Sciences, vol. 2015, Article ID 925935, 7 pages, 2015.

[5] N. P. Challa and R. V. K. Mehta, “Applications of image processing techniques on palm leaf manuscripts-A survey,” in Proceedings of the Conference on Cognitive Science and Artificial Intelligence, CA, USA, February 2017.

[6] S. Athisayamani, A. R. Singh, and A. S. Kumar, “Recurrent neural network-based character recognition system for Tamil palm leaf manuscript using stroke zoning,” in Inventive Communication and Computational Technologies, pp. 165– 176, Springer, Singapore, 2021.

[7] R. S. Ratheash and M. M. Sathik, “A detailed survey of text line segmentation methods in handwritten historical documents and palm leaf manuscripts,” International Journal Of Computer Sciences And Engineering, vol. 7, p. 99, 2019.

[8] J. Ali and J. T. Joseph, “A convolution neural network-based approach for recognizing Malayalam handwritten characters,” Malayalam Handwritten character recognition using cnn, vol. 9, no. 12, 2018.

[9] P. K. S. Balakrishnan and L. Pavithira, “Multi-font optical character recognition using Deep Learning,” International Journal of Recent Technology and Engineering, vol. 8, 2019.

[10] M. A. Hossain and S. Afrin, “Optical character recognition based on template matching,” Global Journal of Computer Science and Technology: C Software & Data Engineering, vol. 19, 2019.

[11] K. Baskar, “Manuscript Libraries in Tamil Nadu: a study of their organisation and preservation in the digital environment,” Doctoral Dissertation in Information Science, Madras University, Tamilnadu, 2018.

[12] S. Devika and K. Vijayakumar, “Digitization of palm leaf manuscripts in Tamil Nadu (India): a study,” Journal of Library Science and Research (JLSR), vol. 2, no. 1, pp. 1–10, 2016.

[13] R. Narenthiran and P. Ravichandran, “Cataloguing and digitization of multilingual manuscript libraries in Tamil Nadu: an evaluative study,” Journal of Advances in Library and Information Sciences, vol. 5, no. 3, pp. 248–253, 2016.

[14] T. K. M. Sageer and A. T. Francis, “Analysis of the palm leaves manuscripts collection for digital archiving: a case study of Sree Sankaracharya University of Sanskrit, Kalady,” Journal Impact Factor, vol. 5, no. 1389, pp. 90–99, 2016.

[15] R. S. Sabeenian, M. E. Paramasivam, P. M. Dinesh, R. Adarsh, and G. R. Kumar, “Classification of handwritten Tamil characters in palm leaf manuscripts using SVM based smart zoning strategies,” in Proceedings of the 2nd International Conference on Biomedical Signal and Image Processing, pp. 18–21, Kitakyushu, Japan, August 2017.

[16] B. Kiruba, A. Nivethitha, and M. Vimaladevi, “Segmentation of handwritten Tamil character from palm script using histogram program approach,” International Journal of Informative and Futuristic Research, vol. 4, no. 5, pp. 6418–6424, 2017.

[17] S. Ghosh, A. Mahajan, and S. Banerjee, “Palm leaf manuscript conservation, the process of seasoning with special reference to Saraswati Mahal library, Tamilnadu in India: some techniques,” International Journal of Information Movement, vol. 2, pp. 122–128, 2017.

[18] N. P. Challa and R. V. K. Mehta, “Applications of image processing techniques on palm-leaf manuscripts—a survey,” Helix: e Scientific Explorer, vol. 7, no. 5, pp. 2013–2017, 2017.

[19] R. Vinoth, R. Rajesh, and P. Yoganandhan, “Intelligence system for Tamil Vattezhuttuoptical character recognition,” International Journal of Computer Science Engineering and Technology, vol. 8, no. 4, pp. 22–26, 2017.

[20] M. Sornam and M. D. Poornima, “Tamil palm-leaf manuscript character segmentation using GLCM feature extraction,” International Journal of Computer Science and Engineering, vol. 6, no. 4, pp. 167–172, 2018.