ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2026.0202

EMPOWERING TAMIL CHARACTERS RECOGNITION WITH ALEXNET AND VISION TRANSFORMER

A. Muthulakshmi, M. Ragavan and S. Siddarth
Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, India

Abstract

The process of recognizing handwritten Tamil char- acters through deep learning methodologies like AlexNet and Vision Transformer (ViT) is a comprehensive endeavor encompassing various stages seamlessly integrated into a cohesive framework. It begins with the acquisition and standardization of a diverse corpus of handwritten Tamil characters, followed by meticulous preprocessing steps such as resizing, grayscale conversion, and pixel normalization to ensure uniformity and enhance model compatibility. Subsequently, data augmentation techniques are employed to enrich dataset variability and mitigate overfitting, employing strategies like rotation, scaling, shearing, and flipping. Central to the recognition pipeline is the strategic selection of deep learning models, with AlexNet and ViT emerging as primary contenders. While AlexNet offers a classical convolutional neural network (CNN) architecture well-suited for image classification tasks, ViT presents a transformative approach leveraging transformer architectures, particularly adept at handling large-scale vision tasks. Training initiates with dataset partitioning into training, validation, and testing subsets, ensuring robust model evaluation. AlexNet is trained utilizing popular deep learning libraries such as PyTorch or TensorFlow, whereas ViT leverages implementations like Google's TensorFlow or Hugging Face's Transformers library. Throughout the training process, models undergo iterative optimization, finetuning hyperparameters, and architecture adjustments to maximize performance while guarding against overfitting. Evaluation metrics such as accuracy, precision, recall, and F1-score serve as benchmarks for model proficiency, with validation data acting as a litmus test for generalization. Rigorous testing on an independent test set solidifies performance assessments before transitioning to deployment. Deployment involves integrating the trained models into practical applications, whether web-based, mobile, or desktop, requiring efficient inference mechanisms and user-friendly interfaces. The real-world efficacy of the recognition system hinges on seamless integration, scalability, and optimization of the user experience, ensuring its practical utility and effectiveness.

Keywords:

Tamil Character, Handwritten, Character Recognition, Vision Transformer, AlexNet, Deep Learning, Neural Networks

1. INTRODUCTION

Handwritten character recognition plays a significant role in numerous applications, including digitization of historical documents, automated translation systems, and optical character recognition (OCR) in handheld devices. Tamil language is one of the ancient languages in the world, has a unique set of characters with intricate shapes and strokes, making handwritten Tamil character recognition a challenging task. Traditional approaches to handwritten character recognition [24] often rely on feature extraction techniques followed by classification algorithms. However, these methods may struggle to capture the complex patterns and variations present in handwritten Tamil characters. With the help of Deep learning techniques, particularly convolutional neural networks (CNNs) and transformer-based models, have shown remarkable success in image classification tasks, including handwritten character recognition. CNNs, such as AlexNet, have been widely adopted for image recognition tasks due to their ability to automatically learn hierarchical features from raw pixel data. On the other hand, transformer-based models, exemplified by the Vision Transformer architecture [28], have demonstrated superior performance in various computer vision tasks by leveraging self-attention mechanisms to capture long-range dependencies in images. In this paper, we propose a novel approach for recognizing Tamil handwritten characters by combining the strengths of AlexNet and Vision Transformer architectures. We aim to leverage the feature extraction capabilities of AlexNet with the self-attention mechanisms of Vision Transformer to improve the accuracy and robustness of Tamil character recognition. Our proposed method involves pretraining both AlexNet and Vision Transformer on large-scale datasets and fine-tuning them on a dataset of Tamil handwritten characters. We evaluate the performance of our approach on a benchmark dataset and compare it with existing methods to demonstrate its effectiveness.

Handwritten character recognition has been extensively studied in the literature, with various approaches proposed to tackle the challenges associated with different languages and scripts. Traditional methods often rely on handcrafted feature extraction techniques, such as histogram of oriented gradients (HOG) and scale-invariant feature transform (SIFT), followed by classifiers such as support vector machines (SVMs) or knearest neighbors (KNN). While these methods have achieved moderate success, they may struggle to generalize well to unseen data and exhibit limited scalability. With the advent of deep learning, there has been a paradigm shift in handwritten character recognition towards end-to-end trainable models that can automatically learn discriminative features from raw data. CNNs, in particular, have emerged as powerful tools for image classification tasks, including handwritten character recognition. Models such as LeNet [22], AlexNet [3], and VGG [24] have demonstrated state-of-the-art performance on benchmark datasets such as MNIST [21] and CIFAR [23].

Transformer-based models, originally introduced for natural language processing tasks, have also shown promise in computer vision tasks. The Vision Transformer architecture, in particular, has gained attention for its ability to capture long-range dependencies in images using self-attention mechanisms. By dividing images into patches and processing them through multiple transformer layers, Vision Transformer can effectively capture spatial relationships and global context information, making it well-suited for tasks such as image classification and object detection. Several recent studies have explored the use of deep learning techniques for handwritten character recognition in various languages. However, to the best of our knowledge, there has been limited research on recognizing Tamil handwritten characters using deep learning architectures, especially in the

context of combining multiple architectures for improved performance. In this paper, we bridge this gap by proposing a novel approach that combines AlexNet and Vision Transformer for Tamil handwritten character recognition.

We describe our proposed approach for recognizing Tamil handwritten characters using a combination of AlexNet and Vision Transformer architectures. The overall architecture of our method is illustrated. We begin by preprocessing the dataset of Tamil handwritten characters to extract relevant features and prepare them for training. This involves steps such as resizing images to a uniform size, normalization, and augmentation to increase the diversity of the training data. Next, we pretrain both AlexNet and Vision Transformer architectures on largescale datasets to initialize their parameters and enable them to capture generic features from images. AlexNet is pretrained on a dataset of diverse images to learn hierarchical features, while Vision Transformer is pretrained on a large-scale image dataset using self-supervised learning techniques to capture global context information. After pretraining, we fine-tune both architectures on the dataset of Tamil handwritten characters to adapt them to the specific characteristics of Tamil script. We employ transfer learning techniques to leverage the knowledge learned during pretraining while fine-tuning the models on the target task. To combine the features extracted by AlexNet and Vision Transformer, we employ fusion techniques such as concatenation or attention mechanism. This enables the models to leverage both local and global information for better character recognition performance. Finally, we perform classification using a softmax layer on top of the fused features to predict the class labels of the input Tamil handwritten characters. We employ standard optimization techniques such as stochastic gradient descent (SGD) with adaptive learning rate schedules to train the classification layer. In this section, we present the experimental results evaluating the performance of our proposed method for recognizing Tamil handwritten characters. We conduct experiments on a benchmark dataset of Tamil characters and compare the performance of our approach with existing methods. We describe the dataset of Tamil handwritten characters used in our experiments, including the number of classes, the distribution of samples per class, and any preprocessing steps applied to the data.

2. RELATED WORKS

Recent advancements in Optical Character Recognition (OCR) have been fueled by the rapid evolution of deep learning methodologies and innovative architectural designs. Among the prominent techniques explored, Convolutional Neural Networks (CNNs) have shown remarkable effectiveness in recognizing characters, especially in challenging scenarios such as cursive handwriting and handwritten Arabic script. Studies by Chandio et al. [1] and Sahlol et al. [2] have demonstrated the prowess of CNNs in recognizing cursive characters in natural scene images and handwritten Arabic characters, respectively. Furthermore, the integration of transfer learning and augmentation techniques has proven beneficial for OCR tasks with limited training data. Rasheed et al. [3] and Nasir et al. [5] leveraged transfer learning and augmentation with AlexNet to achieve promising results in recognizing handwritten Urdu characters and digits.

A notable trend in recent OCR research is the emergence of ViT architectures, which offer a unique approach to image understanding through self-attention mechanisms. Vision Transformers have been extensively explored across various computer vision tasks, including OCR. Han et al. [10] provided a comprehensive survey of ViT models, highlighting their potential for advancing computer vision capabilities. Moreover, hybrid architectures such as CMT, proposed by Guo et al. [11], combine the strengths of both CNNs and ViTs, showcasing superior performance in diverse computer vision tasks.

Scene text recognition and character segmentation remain challenging areas in OCR research. Park et al. [5] introduced a novel multi-lingual OCR system using reinforcement learning of a character segmented, demonstrating robust performance in recognizing text from multiple languages. Additionally, Chernyshova et al. [8] proposed a Two-Step CNN framework for text line recognition in camera-captured images, achieving accurate and efficient text line recognition. Despite these advancements, ongoing research efforts are directed towards addressing challenges such as handling diverse font styles, sizes, and lighting conditions in scene text recognition.

In the realm of scene text recognition, where variations in fonts, sizes, and lighting conditions present substantial challenges, innovative approaches have been devised to improve accuracy and robustness. Wan et al. [7] proposed a novel method for scene Chinese character recognition based on the similarities between Chinese characters, demonstrating superior performance in challenging scene environments. Additionally, Selvam et al. [19] introduced a Transformer-based framework for Scene Text Recognition (STR), leveraging the transformer architecture to capture both local and global contextual information from scene images, resulting in state of-the-art performance.

Recognizing characters in specialized contexts, such as ancient Chinese scripts and printed formulas, requires tailored approaches to address the unique characteristics of these domains. Wu et al. [17] introduced an attention mechanism-based method for recognizing ancient Chinese characters, showcasing superior performance in deciphering complex and stylized characters. Chen and Shi [18] proposed a "Encoder-Decoder" deep learning model for printed formula recognition, achieving accurate transcriptions of mathematical expressions through the extraction of intricate patterns and structures inherent in printed formulas.

Recognizing characters from handwritten texts in specific languages poses additional challenges due to variations in handwriting styles and script complexity. Samala et al. [20] presented a modified 25-layer AlexNet for classifying handwritten Telugu vowel characters, utilizing transfer learning techniques to adapt the model to Telugu script characteristics. Their approach achieved high accuracy in distinguishing Telugu vowel characters, showcasing the efficacy of deep learning and transfer learning in language-specific handwriting recognition tasks.

Recent studies have conducted performance analyses of Vision Transformer (ViT) architectures specifically tailored for character recognition tasks. Chowdhury et al. [16] evaluated ViT-based architectures for cursive handwritten text recognition, demonstrating competitive performance compared to traditional methods. Their analysis highlighted ViT's scalability and adaptability across various handwriting styles and languages,

suggesting its potential for advancing cursive handwritten text recognition.

3. PROPOSED WORK

As we are going to recognize the handwritten characters we want to get more accuracy. We are then ensembling all the models to predict the output. These two components are elaborated upon in the subsequent subsections.

3.1 IMAGE PREPROCESSING

3.1.1 Conversion to Grayscale:

$$C(x,y) = 0.2989\,R(x,y) + 0.5870\,G(x,y) + 0.1140\,B(x,y) \quad (1)$$

This equation converts a color image to grayscale, where R(x,y), G(x,y), and B(x,y) represent the red, green, and blue channels of the image at pixel location (x,y) respectively.

3.1.2 Inversion:

$$\text{Inverted}(x,y) = 255 - C(x,y) \quad (2)$$

This equation inverts the grayscale intensity values of the image.

3.1.3 Max Filtering:

Let *thick*(x,y) represent the grayscale image after max filtering. The max filter operation involves finding the maximum pixel value within a local neighborhood around each pixel.

3.1.4 Scaling:

$$\text{Ratio} = \frac{48.0}{\max(\text{ThickWidth}, \text{ThickHeight})}$$

This equation calculates the scaling ratio based on the maximum dimension (width or height) of the filtered image.

3.1.5 Resizing:

$$x_c = \frac{\sum_x \sum_y \text{res}(x,y)\,x}{\sum_x \sum_y \text{res}(x,y)}\;;\; y_c = \frac{\sum_x \sum_y \text{res}(x,y)\,y}{\sum_x \sum_y \text{res}(x,y)}$$

Let res(x,y) represent the resized image after Lanczos resampling. Resizing involves interpolating pixel values to fit the new dimensions calculated based on the scaling ratio.

3.1.6 Center of Mass Calculation:

$$\text{COM}(x_c, y_c) = \left(\frac{\sum_x \sum_y \text{res}(x,y)\,x}{\sum_x \sum_y \text{res}(x,y)}, \frac{\sum_x \sum_y \text{res}(x,y)\,y}{\sum_x \sum_y \text{res}(x,y)} \right)$$

This equation calculates the center of mass (centroid) of the resized image. This study pursues objective to classify the handwritten tamil character recognition. The Fig.1 can explain the flow of the study.

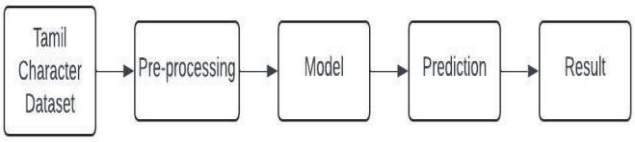

Fig.1. System Flow

- **Image Preprocessing module**: In this module we will first process the image for training because we can't get correct out if the images are in different format. So, we will do simple processing of the data like resizing, rotating, scaling, etc.
- **Training the model**: In this module we are going to use 4 models like Inception, VGG16, AlexNet, Vision
- **Image Pasting**: Let result (x,y) represent the final image. Pasting the resized image onto the blank canvas involves positioning the resized image based on the calculated center of mass.

3.2 MODEL TRAINING

Handwritten Tamil character recognition is a significant task in the domain of computer vision and natural language processing, particularly for applications such as document analysis, optical character recognition (OCR), and language understanding. In recent years, various deep learning architectures have been employed to tackle this challenge, including Inception, AlexNet, VGG16, and Vision Transformer models. Each of these models offers unique advantages and approaches to character recognition, leveraging different architectures and learning techniques.

- Inception Model: The Inception model [25], also known as GoogLeNet, is renowned for its deep architecture with computational efficiency. It consists of multiple inception modules that allow for the parallel operation of convolutional filters at different scales. In the context of handwritten Tamil character recognition, the Inception model can capture intricate details of characters at various resolutions, enabling it to recognize both subtle and prominent features. Moreover, its computational efficiency makes it suitable for real-time or resource-constrained applications.
- AlexNet: AlexNet [20] is one of the pioneering Convolutional Neural Networks (CNN) architectures that gained widespread attention after winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It comprises multiple convolutional and fully connected layers, followed by softmax for classification. In handwritten Tamil character recognition, AlexNet can effectively learn hierarchical features, starting from basic edges and textures to complex patterns specific to Tamil characters. Its ability to learn discriminative features through convolutional layers makes it well-suited for character recognition tasks.

Consider an input image $I \in \mathbb{R}^{W \times H \times C}$, $W \times H \times C$ of size $W \times H \times C$ is subjected to a CONV L_i with kernel size $K \times K$ and architecture allows VGG16 to learn more abstract features by stacking multiple convolutional layers. In handwritten Tamil character recognition, VGG16 can capture both local and global features of characters, leveraging its deep architecture to recognize patterns

at different levels of abstraction. However, its deeper architecture may lead to higher computational costs compared to shallower models like AlexNet.

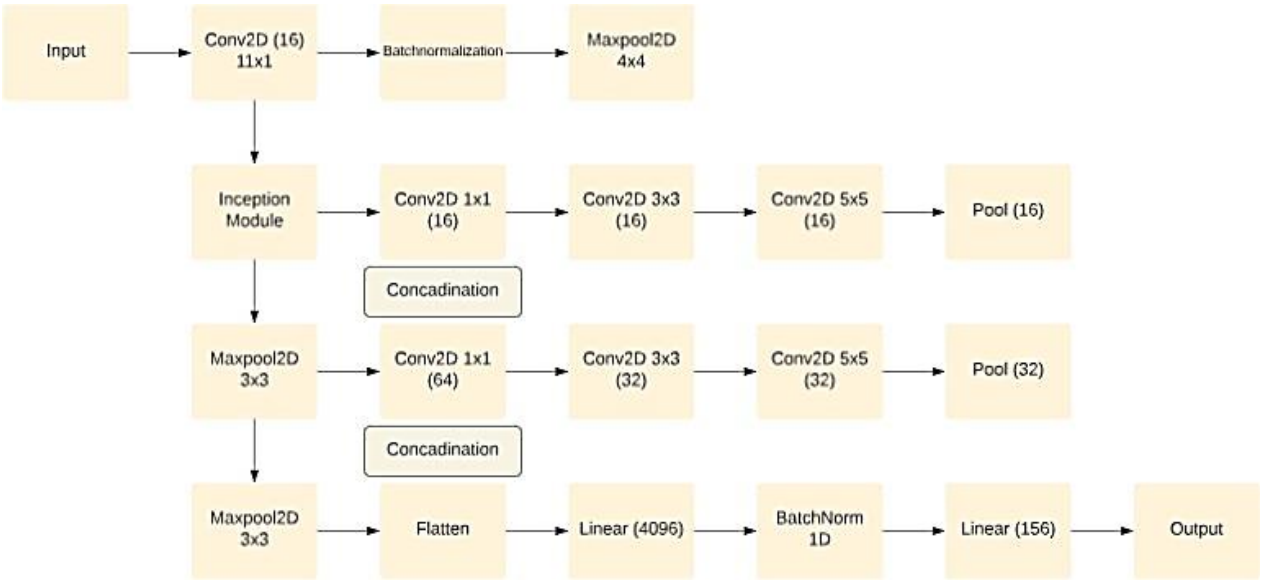

Fig.2. Inception Model

Vision Transformer (ViT): Vision Transformer [26] represents a paradigm shift in computer vision by introducing transformer architecture, originally proposed for natural language processing tasks, into the domain of image recognition. Instead *M* output maps. In the context of handwritten Tamil character recognition, ViT can effectively capture long-range dependencies and relationships between different parts of characters, facilitating accurate recognition even for complex or ambiguous cases.

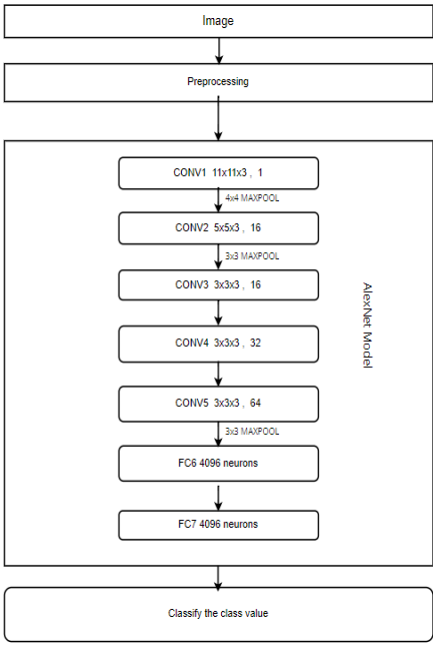

Fig.3. AlexNet Model

For CONV1 layer, there are 96 kernels (output channels) each of size 11x11x3 with Stride 4 where the input *W* and *H* shrink by a factor of 4. The convolutional operation is defined as represented in Eq.(1) for the image *i* with dimensions (*i,j*). Where G is the feature map and F is the convolution filter.

$$G(i,j) = \sum_x \sum_y I(i-x, j-y) F(x,y) \tag{1}$$

Its attention mechanism enables it to focus on relevant parts of the input, making it robust to variations in handwriting styles. Below is how attention [26] is used in processing image.

- Compute scores between different input vectors with $S = QK^T$
- Normalize the scores for the stability of gradient with $S_n = \frac{S}{\sqrt{d_k}}$
- Translate the scores into probabilities with softmax function $P = \text{softmax}(S_n)$
- Obtain the weighted value matrix with $Z=V \cdot P$.

VGG16: VGG16 [24] is characterized by its simplicity and uniform architecture, consisting of 16 convolutional and fully connected layers with small 3x3 convolutional filters. This Ui = WHK2CM are the number of output units, weights (parameters), and connections respectively in the case of CONV layers. When they are subjected to FC layers, $N_i = W \times H \times M$, $P_i = \frac{K^2 CM}{\text{Stride}^2}$, $P_i = K^2 H^2 CM$ and $U_i = W^2 H^2 CM$ are the number of output units, weights (parameters), and connections respectively in the case of CONV layers.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

To elaborate on this in greater detail, given an input vector and the number of heads *h*, the input vector is first transformed into three different groups of vectors: the query group, the key group,

and the value group. In each group, there are *h* vectors with dimension $d_q = d_k = d_v = d_{\text{model}} = 64$. The vectors derived from different inputs are then packed together into three different groups of matrices: $\{Q_i\}_{i=1}^{h}$, $\{K_i\}_{i=1}^{h}$, $\{V_i\}_{i=1}^{h}$. The multi-head [26] attention process is shown as follows:

- *Input*: Firstly, we have made a web application using Flask application. Flask is a lightweight web framework for Python. It's designed to be simple and easy to use, allowing developers to quickly build web applications. In that we will get an input from the canvas because we need to get the character that is drawn by the user. This Flask application acts like client and server.
- *Pre*-Process: After retrieving the image from the web we need to preprocess the image because we need to send the image same as that of model creation. So, we will process the image same as that we did for previous section.
- *Model*: Then we need to load the model in the flask application. As there are multiple characters in tamil language. So, we need to segment the character based on space because we have trained the model in single character.
- *Prediction*: Then in this process after sending the image to the model. Then this model compares with all the class value and it will display the output which has higher value.

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (3)$$

4. RESULTS AND DISCUSSION

Our study's primary goal was to improve odeep learning models' interpretability and comprehension in handwritten Tamil Charcter Recognition. Concluding the handwritten Tamil character recognition using various deep learning architectures like AlexNet, VGG16, Inception, and ViT reveals several insights and potential avenues for future research.

Firstly, the performance of each model can vary significantly based on factors such as dataset size, quality, preprocessing techniques, and hyperparameter tuning. While all architectures have demonstrated capabilities in image recognition tasks, their effectiveness may differ when applied to handwritten character recognition for Tamil script due to its unique characteristics such as complex ligatures and varying writing styles.

4.1 CLASSIFICATION

We are going to classify the Data by getting the real time input from the user. This flow can be seen below.

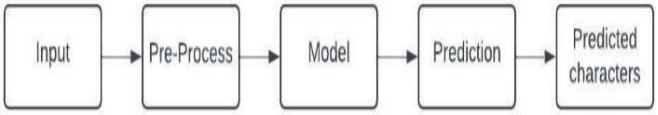

Fig.5. Flow Diagram

AlexNet, despite being one of the pioneering deep learning architectures, may not perform optimally for handwritten Tamil character recognition due to its shallower architecture compared to more modern alternatives. VGG16, with its deeper architecture, may capture more intricate features but could suffer from overfitting, especially with limited training data.

Inception, with its inception modules and computational efficiency, might offer better generalization and performance compared to AlexNet and VGG16. Its ability to capture features at multiple scales could be beneficial for recognizing the diverse range of characters and ligatures in Tamil script.

VisionTransformer, being a transformer-based model, could potentially outperform the other architectures by capturing long-range dependencies and contextual information effectively. However, its performance may heavily rely on the availability of large-scale datasets and extensive pretraining.

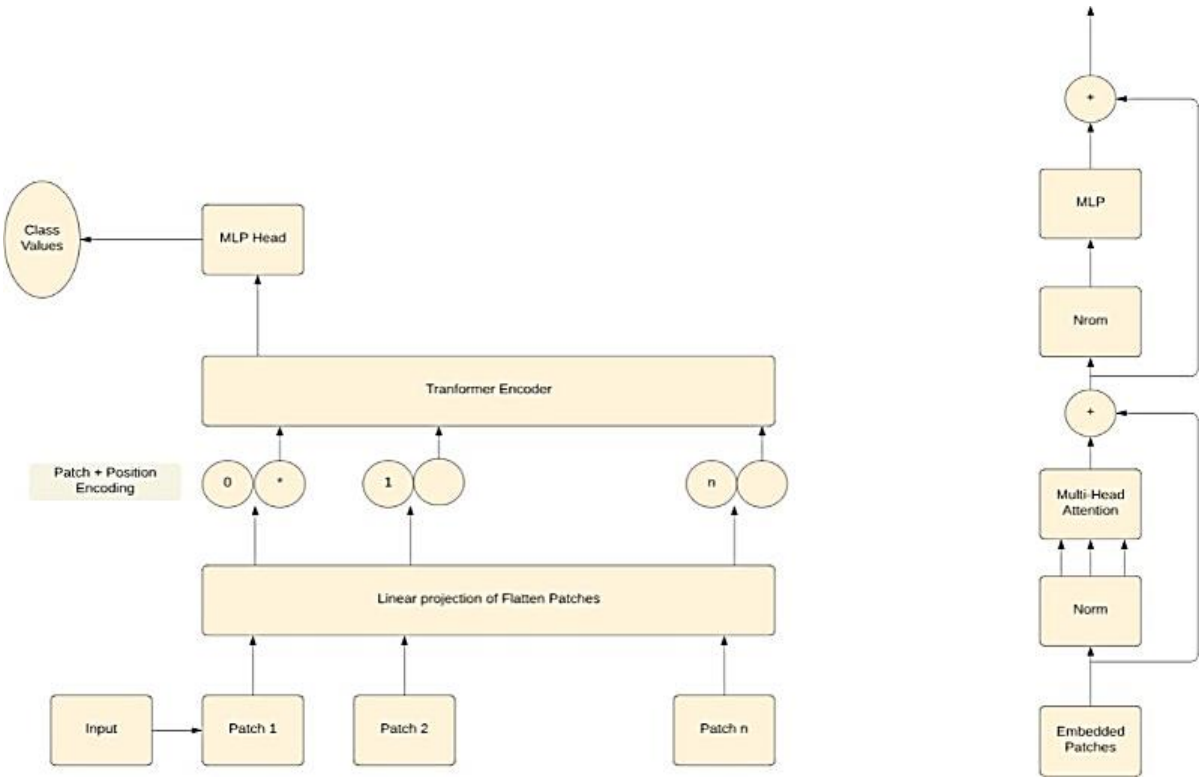

Fig.4. Vision Transformer

However, achieving superior performance with ViT necessitates adequate computational resources for training and finetuning, as well as access to sizable annotated datasets specific to handwritten Tamil characters. Despite these requirements, ViT's ability to learn from large-scale data and leverage pretraining on similar tasks or languages positions it as a promising candidate for achieving exceptional accuracy in handwritten Tamil character recognition.

In summary, while the choice of the best architecture for this task depends on several factors, Vision Transformer stands out as a compelling option due to its potential to deliver high accuracy by effectively leveraging the power of transformer-based architectures. Nevertheless, rigorous evaluation and comparison with other models are essential to validate its efficacy and identify the most suitable approach for the targeted application.

4.2 DATASET DESCRIPTION

The dataset of offline handwritten Tamil characters sourced from HP Labs India comprises around 500 samples for each of the 156 characters, scribed by native Tamil writers in Tamil Nadu, India. The entire dataset was bifurcated into distinct training (50,683 samples) and test sets (26,926 samples), which were utilized in subsequent experiments. The provided training set was further divided into a new training set and a validation set in an 80% to 20% split.

Initially provided as TIFF files of varying dimensions, the bi-level images underwent a transformation to the PNG format. During this conversion, the images were inverted to ensure the foreground appeared white against a black background, and a consistent thickening factor was applied. Subsequently, the images were resized such that the longer side measured 48 pixels, employing the Lanczos algorithm, which includes anti-aliasing, resulting in a transition from bi-level to grayscale images. Finally, the centers of mass of the resultant images were aligned onto a new 64 x 64 canvas. To normalize the images, each grayscale pixel value was transformed from the [0, 1] range to the [-1, 1] range. This preprocessing pipeline aimed to standardize the images and prepare them for subsequent analysis and machine learning tasks. The normalization step particularly ensured that the input data adhered to a consistent range, facilitating the training and convergence of machine learning models. Below we can see the sample dataset in Fig.7.

Fig.6. Sample Dataset

4.3 VISUALIZATION OF RESULTS

The presentation of results from our study is organized high accuracy deep learning model to recognize Tamil handwritten character form differnent model. So, we can visualize the accuracy to two different model AlexNet and ViT.

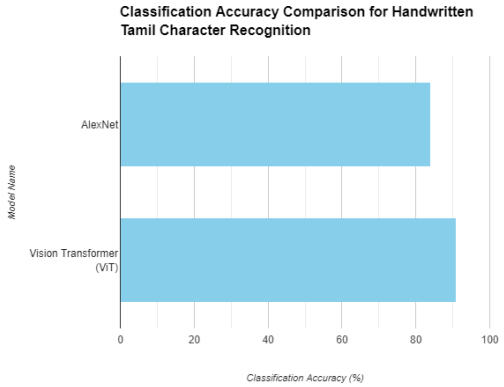

Fig.7. Visualization of AlexNet and ViT

4.4 ACCURACY COMPARISON OF DIFFERENT CLASSIFICATION MODELS

Upon training and evaluating all models, it was evident that each model achieved. The models employed for classification encompassed AlexNet, VGG16, ViT, Inception.

Table.1. Accuracy Comparison of Classification Models

Model	Accuracy (%)
AlexNet	84.27
VGG16	79.33
ViT	91.02
Inception	82.40

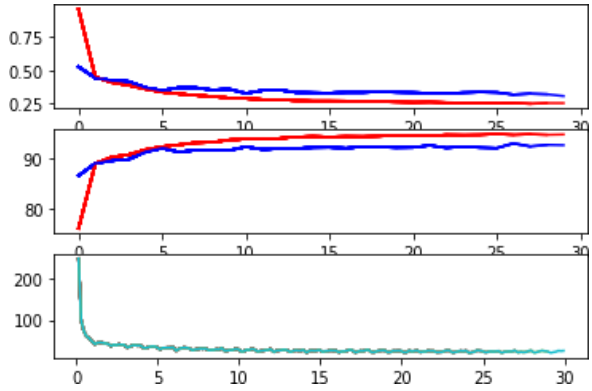

Fig.8. Loss Curve Comparison

In the above graph Fig represent that train loss and Validation loss for the model. In the next part we can infer that we can see training accuracy and validation accuracy of the model. Then finally we can see the loss data over each epoch it is keeping decreasing. So, that only we can get high accuracy of the data.

5. CONCLUSION AND FUTURE SCOPE

Finally, we present Vision Transformer (ViT) a novel pixel level classification model validated with real handeritten character from the page. This model exhibits remarkable performance metrics for classification handwritten character, with an average accuracy of 0.879 and noteworthy PP (0.915), Sensitivity (0.992), Specificity (0.998), and F1-score (0.989). Furthermore, this model will take more feature extraction form other model because of presence of Multi-Layer Attention. So, this model can provide high accuracy

The resulting website offers a versatile platform suitable for both children and adults to practice handwriting skills. However, numerous potential extensions could be implemented to enhance its functionality and user experience. For instance, integrating an audio tool to aid in pronunciation could complement handwriting practice, making the learning process more comprehensive. Expanding the optical character recognition (OCR) system to recognize entire words, involving character segmentation techniques, would greatly improve the platform's utility for language learners and educators alike. Additionally, introducing gamification elements such as progress tracking and interactive challenges could boost user engagement and motivation. Considering mobile compatibility and offline functionality would enhance accessibility, allowing users to practice handwriting across various devices and environments. Moreover, offering customization options such as adjustable difficulty levels and font styles would cater to diverse user preferences and learning objectives. Continuous updates and improvements based on user feedback and technological advancements would ensure the website remains relevant and beneficial. Overall, the potential for future enhancements and expansions of the project is vast, promising a more inclusive and interactive learning experience for users of all ages.

REFERENCES

[1] A.A. Chandio, M.A. Asikuzzaman and M.R. Pickering", Cursive Character Recognition in Natural Scene Images using a Multilevel Convolutional Neural Network Fusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 2, pp. 1-11, 2023.

[2] T. Sahlol, M.A. Abd Elaziz, M.A.A. Al-Qaness and S. Kim, "Handwritten Arabic Optical Character Recognition Approach based on Hybrid Whale Optimization Algorithm with Neighborhood Rough Set", *IEEE Transactions on Image Processing*, Vol. 2, No. 5, pp. 1-11, 2021.

[3] A. Rasheed, "Handwritten Urdu Characters and Digits Recognition using Transfer Learning and Augmentation with AlexNet", *IEEE Access*, Vol. 1, No. 6, pp. 1-9, 2023.

[4] T. Nasir, M.K. Malik and K. Shahzad, "MMU-OCR-21: Towards End-to-End Urdu Text Recognition using Deep Learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, pp. 1-21, 2022.

[5] J. Park, "Multi-Lingual Optical Character Recognition System using the Reinforcement Learning of Character Segmenter", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 5, pp. 1-13, 2013.

[6] M.I. Yap, "Optical Character Recognition with Chinese and Korean Character Decomposition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 3, pp. 1-21, 2022.

[7] Y. Wan, "Research on Scene Chinese Character Recognition Method based on Similar Chinese Characters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 1-11, 2023.

[8] Y.S. Chernyshova, "Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images", *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, No. 9, pp. 1-15, 2023.

[9] S.Y. Arafat and M.J. Iqbal, "Urdu-Text Detection and Recognition in Natural Scene Images using Deep Learning", *IEEE Transactions on Image Processing*, Vol. 25, No. 8, pp. 1-7, 2019.

[10] K. Han, "A Survey on Vision Transformer", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 7, pp. 1-11, 2023.

[11] J. Guo, "CMT: Convolutional Neural Networks Meet Vision Transformers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, pp. 1-15, 2023.

[12] A. Azadbakht, "MultiPath ViT OCR: A Lightweight Visual Transformer-based License Plate Optical Character Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 78, pp. 1-7, 2021.

[13] E. Ibrahimovic, "Optimizing Vision Transformer Performance with Customizable Parameters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 56, pp. 1-5, 2020.

[14] M.A. Zaryab and C.R. Ng, "Optical Character Recognition for Medical Records Digitization with Deep Learning", *IEEE Transactions on Medical Imaging*, Vol. 4, pp. 1-12, 2023.

[15] A. Xue, "Image-to-Character-to-Word Transformers for Accurate Scene Text Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, pp. 1-10, 2023.

[16] A. Chowdhury, "Performance Analysis of Vision Transformer based Architecture for Cursive Handwritten Text Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, pp. 1-11, 2021.

[17] L. Wu, "Ancient Chinese Recognition Method based on Attention Mechanism", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, pp. 1-7, 2019.

[18] Y. Chen and Z. Shi, "A Printed Formula Recognition Method based on the 'Encoder-Decoder' Deep Learning Model", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 1-14, 2023.

[19] P. Selvam, "A Transformer-based Framework for Scene Text Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, pp. 2-9, 2023.

[20] S. Samala, "Handwritten Telugu Vowel Character Classification using Modified 25-Layer AlexNET with Transfer Learning", *IEEE Transactions on Image Processing*, Vol. 30, pp. 1-9, 2023.

[21] G. Elizabeth Rani, M. Sakthimohan, G. Abhigna Reddy, D. Selvalakshmi, Thomalika Keerthi and R. Raja Sekar, "MNIST Handwritten Digit Recognition using Machine Learning", *Proceedings of International Conference on Advance Computing and Innovative Technologies in Engineering*, Vol. 2, pp. 1-11, 2022.

[22] Shuai Tan and Zhi Tan, "Improved LeNet-5 Model based on Handwritten Numeral Recognition", *Chinese Control and Decision Conference*, Vol. 9, pp. 1-7, 2019.

[23] S. Sobana Mari and G. Raju, "Modified View based Approaches for handwritten Tamil Character Recognition", *ICTACT Journal on Image and Video Processing*, Vol. 6, No. 1, pp. 1076-1085, 2015.

[24] S. MadhanMohan and E. Karthikeyan, "Classification of Image using Deep Neural Networks and SoftMax Classifier with CIFAR Datasets", *Proceedings of International Conference on Intelligent Computing and Control Systems*, Vol. 26, pp. 1-11, 2022.

[25] Raghunath Dey, Rakesh Chandra Balabantaray, Jayashree Piri and Debabrata Singh, "Offline Natural Scene Character Recognition using VGG16 Neural Networks", *Proceedings of International Conference on Inventive Research in Computing Applications*, Vol. 2, pp. 1-9, 2021.

[26] Jinhu Sun, Peng Li and Xiaojun Wu, "Handwritten Ancient Chinese Character Recognition Algorithm based on Improved Inception-ResNet and Attention Mechanism", *Proceedings of International Conference on Software Engineering and Artificial Intelligence*, Vol. 22, pp. 1-6, 2022.

[27] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang and Dacheng Tao, "A Survey on Vision Transformer", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 1, pp. 87-110, 2023.

[28] P. Revathy and R. Jayaprakash, "A Hybrid Bidirectional LSTM with Enhanced Vision Transformer Approach for Glaucoma Disease Prediction", *ICTACT Journal on Image and Video Processing*, Vol. 16, No. 1, pp. 3671-3677, 2025.