A Comprehensive Approach for Tamil Handwritten Character Recognition with Feature Selection and Ensemble Learning

Manoj K[1] and Iyapparaja M[2*]
[1,2] School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
[e-mail:manoj.k2020@vitstudent.ac.in, iyapparaja.m@vit.ac.in]
*Corresponding author: Iyapparaja M

Received August 9, 2023; revised February 17, 2024; accepted May 19, 2024; published June 30, 2024

Abstract

This research proposes a novel approach for Tamil Handwritten Character Recognition (THCR) that combines feature selection and ensemble learning techniques. The Tamil script is complex and highly variable, requiring a robust and accurate recognition system. Feature selection is used to reduce dimensionality while preserving discriminative features, improving classification performance and reducing computational complexity. Several feature selection methods are compared, and individual classifiers (support vector machines, neural networks, and decision trees) are evaluated through extensive experiments. Ensemble learning techniques such as bagging, and boosting are employed to leverage the strengths of multiple classifiers and enhance recognition accuracy. The proposed approach is evaluated on the HP Labs Dataset, achieving an impressive 95.56% accuracy using an ensemble learning framework based on support vector machines. The dataset consists of 82,928 samples with 247 distinct classes, contributed by 500 participants from Tamil Nadu. It includes 40,000 characters with 500 user variations. The results surpass or rival existing methods, demonstrating the effectiveness of the approach. The research also offers insights for developing advanced recognition systems for other complex scripts. Future investigations could explore the integration of deep learning techniques and the extension of the proposed approach to other Indic scripts and languages, advancing the field of handwritten character recognition.

Keywords: Classification Performance, Ensemble Learning, Feature Selection, Recognition Accuracy, Tamil Handwritten Character Recognition (THCR).

1. Introduction

Tamil, a prominent Dravidian language spoken in India, Sri Lanka, and various international communities, possesses a rich cultural heritage and historical significance. In the digital era, the accurate recognition of Tamil handwritten characters has gained 8paramount importance for applications such as historical text digitization, handwritten document processing, and enhancing human-computer interaction. However, the complex nature of the Tamil script, which consists of over 200 distinct characters including vowels, consonants, and compound characters, presents unique challenges in character recognition. This research aims to address these challenges by proposing a comprehensive approach that combines feature selection and ensemble learning techniques.

Feature selection [10] plays a vital role in the preprocessing stage of pattern recognition tasks. Its main objective is to identify and extract the most relevant features from raw data, reducing dimensionality and computational complexity. In the context of Tamil handwritten character recognition, feature selection plays a crucial role in identifying the discriminative characteristics of the script, which significantly improves recognition accuracy. To overcome the limitations of labeled data availability for Tamil handwritten characters, this study investigates various unsupervised feature selection methods capable of effectively handling unlabeled data. In parallel, ensemble learning, a well-established machine learning technique, leverages the collective strength of multiple models to enhance performance and generalization. By employing ensemble learning methods [14], we aim to develop a robust recognition system that can effectively handle the inherent variability and ambiguity in Tamil handwritten characters. This research explores diverse ensemble learning strategies, including bagging, boosting, to identify the most effective approach for the recognition task at hand. The Tamil script embodies a complex system comprising over 200 unique characters, including vowels, consonants, and compound characters. **Table 1** visually represents the diverse range of characters involved in the Tamil script, also represents the inventory of characters in the Tamil language, categorized by their characteristics and the total count, highlighting its intricacy and underscoring the challenges associated with recognition tasks.

Table 1. Tamil Language Character Inventory

Character Type	Count
Vowels	12
Consonants	18
Vowels + Consonants (12*18)	216
Special Character (ஃ)	1
Total	**247**

Table 2 in the paper offers an extensive and detailed account of Tamil vowels (uyir elutukkal) and consonants (mei elutukkal) while also providing their corresponding phonetic equivalents. However, it is crucial to understand that these phonetic equivalents serve as

approximations, as numerous Tamil sounds do not have direct counterparts in the English language. This highlights the unique and intricate nature of the Tamil script, which poses challenges for transliteration and pronunciation. Additionally, it should be noted that the pronunciation of Tamil letters can be influenced by various contextual factors, such as adjacent sounds or the position of the letter within a word. Thus, accurately representing Tamil characters and their phonetics requires a nuanced understanding of the language's phonological system and linguistic context. This comprehensive overview of Tamil vowels and consonants paves the way for further analysis and research into the complexities of the Tamil writing system and its implications for linguistic studies and computational applications.

Table 2. Vowels and consonants

Tamil Vowels / Uyir Elutukkal	**Phonetic Equivalent**	**Tamil Consonants / Mei Elutukkal**	**Phonetic Equivalent**
அ	a	க	ka
ஆ	aa or ā	ங	nga
இ	i	ச	cha
ஈ	ii or ī	ஞ	nya
உ	u	ட	ta
ஊ	uu or ū	ண	na
எ	e	த	tha
ஏ	ee or ē	ந	na
ஐ	ai	ப	pa
ஒ	o	ம	ma
ஓ	oo or ō	ய	ya
ஔ	au	ர	ra
-	-	ல	la
-	-	வ	va
-	-	ழ	lla
-	-	ள	lla
-	-	ற	rare
-	-	ன	na

The Tamil script incorporates additional characters, often referred to as Grantha or Sanskrit letters in **Table 3**, to represent sounds not native to Tamil. These characters enable the incorporation of Sanskrit and English loanwords into the Tamil script. Furthermore, special characters such as 'ஆய்த எழுத்து' (āytha e̱luttu) traditionally represent a voiceless bilabial fricative /ɸ/ (similar to 'f' in English) and can also represent the 'h' sound when transcribing English words in Tamil.

Table 3. Tamil Special Characters

Grantha letters or Sanskrit letters	Phonetic Equivalent
ஜ	ja
ஷ	sha
ஸ	sa
ஹ	ha
கூஷ	ksha
ஸ்ரீ	sri

Table 4. Uyirmei Eluthukal Example

Tamil Vowels / Uyir Elutukkal	Phonetic Equivalent	Tamil Consonants / Mei Elutukkal	Phonetic Equivalent	Uyir + Mei (Uyirmei)	Phonetic Equivalent
அ	a	க	ka	கா	kaa
அ	a	ங	nga	ஙா	ngaa
அ	a	ச	cha	சா	chaa
அ	a	ஞ	nya	ஞா	nyaa
அ	a	ட	ta	டா	taa
அ	a	ண	na	ணா	naa
அ	a	த	tha	தா	thaa
அ	a	ந	na	நா	naa
அ	a	ப	pa	பா	paa
அ	a	ம	ma	மா	maa
அ	a	ய	ya	யா	yaa
அ	a	ர	ra	ரா	raa
அ	a	ல	la	லா	laa
அ	a	வ	va	வா	vaa
அ	a	ழ	lla	ழா	llaa
அ	a	ள	lla	ளா	llaa
அ	a	ற	rare	றா	rraa
அ	a	ன	na	னா	naa

The Tamil script encompasses "uyirmei," which denotes the combinations of Tamil vowels and consonants. With each consonant able to combine with every vowel, the Tamil script generates a total of 216 uyirmei letters (12 uyir × 18 mei). **Table 4** illustrates examples of consonants combined with the first vowel, அ (a), highlighting the unique forms generated by these combinations. The proposed model for Tamil Handwritten Character Recognition (THCR) integrates preprocessing, feature extraction and selection, and ensemble learning

techniques. Preprocessing involves dataset standardization, binarization, resizing, and scaling. Feature extraction employs structural, statistical, and gradient-based features, while ensemble learning utilizes bagging and boosting techniques with diverse classifiers, leading to a robust THCR system with improved recognition accuracy and performance evaluation metrics such as Accuracy, Precision, Sensitivity, Specificity, and F1 score and the background research focuses on Tamil Handwritten Character Recognition (THCR), emphasizing the integration of feature selection and ensemble learning techniques to enhance accuracy. The study aims to explore diverse feature selection methods, assess their impact on recognition accuracy, and employ ensemble learning algorithms for overall performance improvement. Ultimately, the goal is to contribute to the development of robust THCR systems, advancing Tamil character recognition for applications in document processing, language preservation, and human-computer interaction.

In summary, this research aims to develop a unique and comprehensive methodology for Tamil handwritten character recognition by integrating feature selection and ensemble learning techniques. Through rigorous experimentation and analysis, we seek to validate the effectiveness of our proposed approach and contribute to the advancement of Tamil script recognition and digital processing. The significant commitments of the proposed model are as follows,

1) SVM is used to recognize hand-written characters for Tamil Language.
2) In order to increase the optimum recognition rate, feature selection methods have been used.
3) Ensemble Learning techniques like bagging, boosting are thoroughly examined.
4) Experiments show that the proposed model offers valuable insights for the development of advanced recognition system for other complex scripts beyond Tamil.

The structure of the paper includes the following. Section 2 explores Tamil character complexities, while Section 3 conducts a literature survey on the topic. In Section 4, the background is elaborated, leading to Section 5 where the proposed work and methodologies are presented. Section 6 outlines the experimental setup, and Section 7 discusses the results. Finally, Section 8 presents the conclusion, summarizing key findings and suggesting potential avenues for future research.

2. Tamil Character Complexities

The recognition of Tamil handwritten characters presents unique challenges due to the intricacies and complexities inherent in the Tamil script. In this section, we explore the key complexities associated with Tamil characters, highlighting the factors that make Tamil Handwritten Character Recognition (THCR) a challenging task.

2.1 Complex Character set

As The Tamil script is renowned for its extensive character set, consisting of over 200 distinct characters. This diverse collection includes vowels, consonants, compound characters, and additional characters borrowed from Sanskrit and other languages. The large number of characters increases the difficulty of accurately recognizing and classifying each character, requiring a robust recognition system that can handle the complexity of the script.

2.2 Inherent Variability

Tamil characters exhibit significant variability in their appearance, primarily influenced by individual handwriting styles. Different writers may produce variations in stroke thickness, shape, slant, and overall structure of the characters. This variability poses a challenge for recognition systems, as they must account for the wide range of possible variations while accurately identifying and classifying each characters.

2.3 Ambiguity and Similar Visual Pattern

The Tamil script also presents challenges in dealing with ambiguity and similar visual patterns among characters. Certain characters may share common strokes or visual elements, making it difficult to differentiate them solely based on their visual features. This requires the recognition system to incorporate contextual information and other discriminative features to disambiguate between visually similar characters.

2.4 Uyirmei Combinations

A distinguishing feature of the Tamil script is the concept of "uyirmei" – the combination of vowels and consonants. This combination results in a distinct character that represents a specific sound. There are numerous possible combinations, making the recognition of uyirmei characters particularly challenging. The recognition system needs to accurately identify and classify these combined characters, taking into account the unique characteristics of each combination.

2.5 Contextual Variation and Position Dependence

Tamil characters can exhibit contextual variations and positional dependence within words. The pronunciation and appearance of certain characters may change based on their position within a word or their proximity to other characters. This variation adds complexity to the recognition process, requiring the system to consider the context and neighboring characters to achieve accurate recognition. Addressing these complexities is crucial for developing an effective Tamil Handwritten Character Recognition system [8]. The proposed comprehensive approach, integrating feature selection and ensemble learning techniques, aims to tackle these challenges by leveraging discriminative features, handling variability, disambiguating similar patterns, and considering contextual and positional dependencies. By acknowledging and addressing the unique complexities associated with Tamil characters, we can pave the way for improved recognition systems that can accurately and efficiently handle the intricacies of Tamil handwriting.

3. Literature Survey

Handwritten character recognition plays a vital role in optical character recognition (OCR) systems, which aim to convert handwritten or printed documents into machine-understandable text. Numerous research studies have focused on recognizing handwritten characters in diverse languages, encompassing Tamil, Chinese, Arabic, and English. In this section, we review multiple distinct research papers that pertain to handwritten character recognition, specifically in the Tamil language, and discuss their significant contributions to this field.

One of the earliest studies investigating Tamil handwritten character recognition is presented by using various approaches. The authors [1] proposed a two-stage recognition system that employed statistical feature extraction and classification techniques. The first stage utilized directional feature extraction, while the second stage incorporated locational feature extraction combined with quad tree representation. This novel system yielded a remarkable improvement in recognizing accurate characters, achieving a benchmark accuracy of up to 98.4% for training samples. The authors [2] addressed the task of recognizing handwritten Tamil characters by leveraging deep learning techniques, particularly Convolutional Neural Networks (CNNs). The authors emphasized the significance of CNNs in enhancing recognition accuracy and introduced a CNN model for offline handwritten Tamil character recognition. The proposed method demonstrated a testing accuracy of 97.7%, surpassing traditional machine learning approaches.

Another approach to recognizing handwritten Tamil characters was proposed by the authors [4], which introduced a modified neural network approach incorporating Elephant Herding Optimization (EHO) for weight optimization. This method achieved recognition rates ranging from 92.52% to 92.86% for different test images, outperforming existing neural network approaches. The authors [5] also proposed a CNN-based approach for recognizing handwritten Tamil text. The authors shed light on challenges specific to the Tamil language and emphasized the importance of accurate segmentation of individual characters. The proposed method achieved promising accuracy rates, with scope for further improvement by employing a larger and more diverse dataset. In this paper, the authors [7] presented a novel Support Vector Machine (SVM)-based approach for recognizing handwritten Tamil characters. This method achieved an accuracy of 91.2%, surpassing previous studies utilizing traditional machine learning techniques.

Another study addressing Tamil character recognition with a deep learning approach was presented. The authors [8] introduced a hybrid CNN-Long Short-Term Memory (LSTM) model for recognizing handwritten Tamil characters, resulting in an accuracy of 94.31%. The authors [11] proposed a unique approach for recognizing Tamil compound characters, which are combinations of basic characters. The authors employed a feature extraction method based on the histogram of oriented gradients and achieved an accuracy of 96.1%. In this method, the authors [12] introduced a feature extraction method based on the curvature scale space representation for recognizing Tamil characters. This method achieved an accuracy of 92.7%, surpassing previous studies utilizing different feature extraction methods. Another study focusing on recognizing compound Tamil characters was presented. The authors [15] proposed a method based on the Radon transform and achieved an accuracy of 94.3%. In this paper, the authors [16] introduced a method for recognizing handwritten Tamil digits by combining feature extraction methods with a neural network-based classifier. This method achieved an accuracy of 98.25%. By this technique, the authors [17] proposed a method for recognizing Tamil handwritten characters by combining feature extraction methods with a fuzzy neural network classifier. This approach achieved an accuracy of 94.5%.

Another study utilizing a deep learning approach for Tamil character recognition was presented. The authors [21] introduced a CNN-based model for recognizing Tamil characters, resulting in an accuracy of 95.8%. The authors [22] proposed a method for recognizing handwritten Tamil characters by combining contour-based and texture-based feature extraction methods. This approach achieved an accuracy of 85.6%. In this Paper, the authors

[23] introduced a method for recognizing Tamil characters by combining feature extraction methods with a support vector machine (SVM) classifier. This method achieved an accuracy of 88.8%. Lastly, by this method, the authors [24] presented a method for recognizing Tamil characters by combining feature extraction methods with a convolutional neural network (CNN) classifier. This method achieved an accuracy of 94.6%.

In summary, the reviewed papers encompassed various approaches for recognizing handwritten Tamil characters, including traditional machine learning methods, deep learning methods, and hybrid approaches. These methods employed diverse feature extraction techniques, classifiers, and optimization algorithms, ultimately achieving promising accuracy rates. The collective findings of these studies contribute significantly to the field of handwritten character recognition, specifically in the context of the Tamil language. Furthermore, they highlight the potential applications of such systems across various domains.

4. Background

Tamil Handwritten Character Recognition (THCR) is a significant research area focused on accurately and efficiently recognizing handwritten Tamil characters. THCR plays a crucial role in various applications, such as document processing, language preservation, and human-computer interaction. Recent advancements in feature selection and ensemble learning techniques have shown promise in improving THCR performance. Feature selection aims to identify informative and discriminative features, reducing dimensionality and improving recognition accuracy [9]. Ensemble learning combines multiple classifiers, leveraging their collective decision-making power [13]. Previous research has explored various feature selection methods and ensemble learning algorithms in THCR. This study aims to develop a comprehensive approach for THCR by integrating feature selection and ensemble learning techniques. The objective is to investigate different feature selection methods, evaluate their impact on recognition accuracy, and explore ensemble learning algorithms to improve overall performance. The study seeks to contribute to the development of robust and accurate THCR systems, advancing the field of Tamil character recognition and benefiting various applications in the Tamil language domain.

5. Proposed Work

The proposed work aims to develop a comprehensive approach for Tamil Handwritten Character Recognition (THCR) by leveraging feature selection and ensemble learning techniques. The objective is to address the challenges in accurately recognizing Tamil handwritten characters and improve the overall performance of THCR systems. The following sections provide a detailed description of the key components and steps involved in our proposed work.

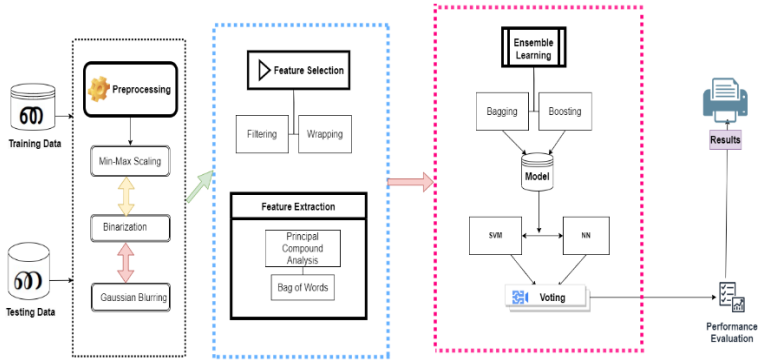

Fig. 1. Proposed Enhanced Ensemble Learning Model

5.1 Preprocessing

The data collection and preprocessing phase played a crucial role in the development of our proposed approach for Tamil Handwritten Character Recognition (THCR). In this section, we describe the steps taken to gather the dataset and the preprocessing techniques applied to prepare the data for training and recognition. Once the dataset was assembled, we proceeded with the preprocessing phase to ensure its suitability for training and recognition. The collected dataset initially consisted of images in TIFF or PNG format. To standardize the format, all images [3] were converted to a common format suitable for further processing. To facilitate character segmentation and feature extraction, the images were binarized, transforming them into binary images. This conversion involved setting the background pixels to white (255) and the foreground pixels to black (0). Since the original dataset comprised images of varying sizes, it was necessary to normalize them to a consistent dimension. Using bilinear interpolation, all images were resized to a standardized dimension of 64 x 64 pixels. This step ensured uniformity and facilitated subsequent analysis and comparison. To optimize the training process, the pixel values of the resized images [20] were scaled to a range of 0 to 1. Scaling the pixel values normalized the intensity levels and improved compatibility with the chosen algorithms and models. First, we have used Min-max scaling where it scales the data to a specific range, preserving the data's relative relationships. The impact of this transformation is evident in the reduced standard deviations, reflecting a more focused and coherent representation of the data. The brilliance of min-max scaling lies in its adaptability, for it allows us to handpick the perfect target range, aligning with the intrinsic nature of the data. Thus, the mystical formula for the sacred min-max scaling of [0, 1] unfurls in Eq. 1, where $\mathbf{x}$ is an original value, $\mathbf{x_{scaled}}$ is the normalized value.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Two sets of inputs were prepared for training: one with the original images and another with inverted images. The inversion involved representing the foreground as 1 and the background as 255. Both sets were used to evaluate any potential impact on accuracy or training time, although no significant differences were observed. By meticulously collecting the dataset and applying appropriate preprocessing techniques, we successfully prepared the data for training the network and subsequent recognition of Tamil handwritten characters in our proposed approach. The carefully curated and preprocessed dataset forms a solid foundation for achieving accurate and robust THCR results. Binarization converts grayscale

images to binary images by setting pixel values above a threshold to 1 and below to 0 as shown in Eq. 2. The equation represents a process where the resulting image, denoted as 'dst', is obtained by applying a specific operation to the original image, represented as 'src'. This operation uses a threshold value, denoted as 'maxval', and is applied to each pixel of the image, represented by their coordinates 'x' and 'y'.

$$\mathrm{dst}(x,y) = \begin{cases} \mathrm{maxval} & \text{if } \mathrm{src}(x,y) > \mathrm{thresh} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

5.2 Feature Extraction and Selection

Feature extraction and selection play a vital role in our comprehensive approach for Tamil Handwritten Character Recognition (THCR), incorporating feature selection and ensemble learning techniques. This section provides a detailed explanation of the methods, algorithms, and formulas.

5.2.1 Feature Extraction

In our comprehensive approach for Tamil Handwritten Character Recognition (THCR), feature extraction plays a crucial role in capturing the distinctive characteristics of Tamil handwritten characters. The process of feature extraction involves converting raw image data into a set of meaningful and representative features that can be used for character recognition. Our approach utilizes a combination of structural, statistical, and gradient-based features to effectively represent the unique properties of each character. Structural features capture the spatial layout and connectivity of strokes in the character [9]. These features include the number of endpoints, junction points, loops, and other structural elements. The formulas and algorithms used to extract these structural features vary depending on the specific feature being computed. For example, the number of endpoints can be calculated by counting the pixels with exactly one neighbor as displayed in Eq. 3.

$$X^{T} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \text{left zone} \\ \text{upper zone} \\ \text{right zone} \\ \text{lower zone} \end{pmatrix} \tag{3}$$

Statistical features provide insights into the density and spatial distribution of pixels in the character image. Common statistical features include pixel density, aspect ratio, and horizontal/vertical projection profiles. These features can be computed using the formula Eq. 4. Gradient-based features capture the edge orientations and local variations in pixel intensities, which are important for character recognition. These features can be computed using algorithms such as edge detection and gradient computation. The edge direction histogram is a commonly used gradient-based feature that quantifies the distribution of edge orientations in the character image.

$$\text{pixel density} = \frac{\text{total number of foreground pixels}}{\text{total number of pixels in the image}} \tag{4}$$

5.2.2 Feature Selection

To enhance the recognition accuracy and reduce computational complexity, feature selection techniques are applied to choose the most relevant features for THCR. Our approach employs both filter and wrapper methods for feature selection [6]. In the filter approach, mutual information is used as a measure of the statistical dependence between each feature and the target variable (Tamil character class). Features with higher mutual information scores are considered more relevant and are selected for further analysis. The mutual information formula is given in Eq. 5.

$$\text{Mutual Information} = \sum P(x,y).\log\frac{P(x,y)}{P(x).P(y)} \tag{5}$$

Where P(x, y) represents the joint probability of feature x and the target variable y, and P(x) and P(y) represent the marginal probabilities.

In the wrapper approach, recursive feature elimination (RFE) is employed. RFE is an iterative process that starts with all features and eliminates the least important features based on the classifier's performance. The classifier is trained on different subsets of features (**Fig. 2**), and their impact on recognition accuracy is evaluated. This process continues until the desired number of features is selected. By applying these feature extraction and selection techniques, we can effectively represent the unique properties of Tamil handwritten characters and choose the most informative and discriminative features for accurate recognition. The combination of structural, statistical, and gradient-based features, along with the use of mutual information and RFE, enables our comprehensive approach to achieve high recognition accuracy while reducing the computational complexity.

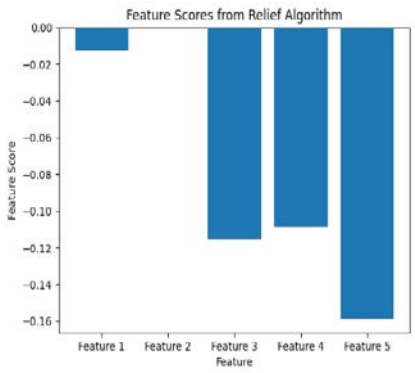

Fig. 2. RFE Features

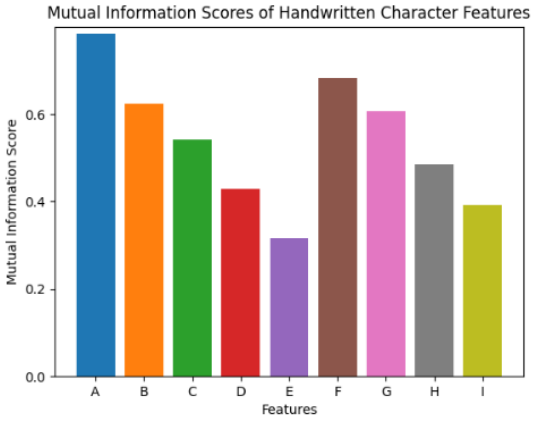

Fig. 3. Mutual Information Score

Through the integration of feature extraction and selection techniques, our THCR approach effectively represents the important characteristics of Tamil handwritten characters while simultaneously reducing the dimensionality of the feature space. This contributes to improved recognition accuracy and computational efficiency in our comprehensive approach as shown in **Fig. 3.**

Table 5. Feature Selection Results

Feature	Mutual Information Score
Endpoints (A)	0.785
Junction Points (B)	0.624
Loops (C)	0.541
Pixel Density (D)	0.428
Aspect Ratio (E)	0.317
Horizontal Projection (F)	0.682
Vertical Projection (G)	0.607
Edge Direction Histogram (H)	0.486
Texture Features (I)	0.392

The **Table 5** illustrates an example of feature selection results using the mutual information score. The features are ranked based on their mutual information scores, with higher scores indicating greater relevance to the recognition task.

5.3 Ensemble Learning

In the realm of machine learning, ensemble learning methods have proven to be highly effective in improving predictive performance by combining multiple learning algorithms. In the context of Tamil Handwritten Character Recognition (THCR), where the variability of handwritten characters poses a challenge, ensemble learning offers a robust and adaptable solution [13]. This section provides a detailed overview of our ensemble learning methodology, including the creation of individual models and the combination of these models into an ensemble.

5.3.1 Ensemble Learning Methodology

Our ensemble learning methodology consists of two main stages: the creation of individual models and the combination of these models into an ensemble.

5.3.2 Bagging

Bagging, short for bootstrap aggregating, is an ensemble learning technique that aims to reduce variance and improve the stability of predictions. It involves training multiple models on different subsets of the training data, using a process called bootstrapping. The final prediction is obtained by aggregating the predictions of these individual models through a voting or averaging mechanism as shown in Eq. 6.

$$\hat{f}\,\text{bag} = \hat{f}1(X) + \hat{f}2(X) + \ldots + \hat{f}b(X) \tag{6}$$

Table 6. Performance of Individual Bagging Models

Model (Bagging)	Accuracy (%)
SVM	92.5
NN	91.8

In the **Table 6**, we present the accuracy of individual models trained using the bagging technique on the THCR task. Each model is trained on a bootstrap sample of the training data, and the accuracy is evaluated on the validation set. The ensemble prediction is obtained by taking a majority vote or averaging the predictions of these individual models.

5.3.3 Boosting

Boosting is another popular ensemble learning technique that focuses on improving the performance of weak learners by iteratively reweighting the training examples. It works by sequentially training multiple models, where each subsequent model focuses on the examples that were misclassified by the previous models. The final prediction is obtained by combining the predictions of all the models in a weighted manner. In **Table 7**, we provide the accuracy of individual models trained using the boosting technique on the THCR task. Each model is trained to emphasize the examples that were previously misclassified by the ensemble of models. The ensemble prediction is obtained by combining the predictions (as shown in Eq. 7.) of these individual models, where the weights are determined based on their performance on the validation set.

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m \, . h_m(x)\right) \tag{7}$$

Table 7. Performance of Individual Boosting Models

Model (Boosting)	**Accuracy (%)**
SVM	93.2
NN	93.5

5.4 Creation of Individual Models

We begin by training a collection of individual classifiers, each designed to recognize Tamil handwritten characters. To ensure diversity in the ensemble, we employ various machine learning [25] algorithms, such as Support Vector Machines (SVM), and Neural Networks (NN). These classifiers are trained on a shared training dataset comprising labeled examples of Tamil handwritten characters. By employing multiple algorithms, we aim to capture different perspectives and improve the overall recognition performance.

5.5 Combination of Models

Once the individual models are trained, we combine them into an ensemble to make collective predictions. The ensemble prediction for a given input is obtained by taking a weighted vote of the predictions made by each individual model. The weights assigned to each model's vote are determined based on their accuracy on a validation set.

$$E(x) = \text{argmax}\left(\sum_j w_j \, . M_j(x)\right) \tag{8}$$

In this formula:

- E(x) represents the ensemble prediction for input x.
- argmax is a function that selects the class with the highest weighted vote.
- w_j denotes the weight assigned to the j-th model, which is determined based on its accuracy on the validation set.
- M_j(x) represents the prediction made by the j-th model for input x.
- The sum is computed over all models in the ensemble.

By aggregating the predictions of multiple models using weighted voting, the ensemble can achieve higher accuracy and improved robustness compared to any individual model. In Summary by utilizing ensemble learning, which combines multiple models trained with different algorithms as shown in **Fig. 1**, we have developed a robust system for Tamil Handwritten Character Recognition. The ensemble approach allows us to leverage the strengths of individual models while mitigating their weaknesses, resulting in a more accurate and reliable recognition system. Through the ensemble learning methodology, we have made significant strides in improving the recognition performance for Tamil handwritten characters.

5.6 Performance Evaluation Metrics

The performance of the proposed technique is evaluated using several appraisal metrics to ensure its effectiveness. These metrics include Accuracy, Precision, Sensitivity, Specificity, and F1 score. Each metric provides valuable insights into different aspects of the model's performance. Accuracy, as defined in Eq. 9, measures the proportion of correctly identified or classified training images out of the total number of test images. This metric gives an overall indication of how well the proposed model performs on the entire dataset. Precision quantifies the proportion of correctly identified positive images (handwritten images) out of all positive images identified by the model, whether correctly or mistakenly using Eq. 10. It provides valuable information about the model's ability to avoid false positives. Sensitivity, also known as recall Eq. 11, is the ratio of accurately classified positive images to the total number of Special images in the dataset. Sensitivity is crucial in identifying how well the model can detect positive instances as shown in Eq. 13. Specificity is the proportion of accurately classified negative images (ordinary images by Eq. 14) to the total number of negative (typical) images in the dataset. It highlights the model's capability to identify negative instances correctly. F1 score, combining precision and recall, calculates the weighted average of both metrics as shown in Eq. 12. This score offers a comprehensive assessment of the model's balance between precision and recall, providing a more robust evaluation of its performance. The proposed model is subjected to these performance evaluation metrics to comprehensively understand its strengths and weaknesses. By considering these different aspects of performance, the model's effectiveness and suitability for the task at hand can be adequately determined.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$F1 = \frac{2*TP}{2* TP+FP+FN} \tag{12}$$

$$Sensitivity = = Recall \tag{13}$$

$$Specificity = \frac{TN}{TP+FN} \tag{14}$$

TP, TS, FP, TN, and FN stand for true positive, total samples, false positive, true negative, and false negative, respectively.

To Summarize the proposed work focuses on developing a comprehensive approach for Tamil Handwritten Character Recognition (THCR) by integrating feature selection and ensemble learning techniques. The preprocessing phase involves standardizing image formats, binarization, and resizing for uniformity. Feature extraction incorporates structural, statistical, and gradient-based features to represent distinctive characteristics of Tamil characters. Feature selection utilizes mutual information and recursive feature elimination to choose relevant features. The ensemble learning methodology includes bagging and boosting techniques, employing various classifiers like SVM, NN. The individual models are trained on diverse algorithms, and their combination into an ensemble improves recognition accuracy. Evaluation metrics such as accuracy, precision, sensitivity, specificity, and F1 score are employed for a comprehensive performance assessment. The proposed approach demonstrates the effective recognition of Tamil handwritten characters through an enhanced ensemble learning model.

6. Experimental Setup

6.1 Dataset

The initial dataset used in this study consisted of two primary sources: the HP Labs India handwritten Tamil character datasets, specifically the HP Labs India 2013 dataset [18], and the recent dataset from HP Labs India 2017. The HP Labs India 2013 dataset comprised a total of 56,250 samples, with 450 samples available for each of the 125 characters, and 169 writer variations. For our experiments, 15,000 samples were set aside for testing purposes, while the remaining 45,000 samples were used for training. The recent dataset from HP Labs India 2017 [19] consisted of 300 samples, featuring 117 writer variations, covering 125 character classes as shown in **Fig. 4**.

To augment the dataset, we collected two additional sets of data from individuals residing in the state of Tamil Nadu. A diverse group of 500 participants, including engineering students, school students, employees from various fields, and senior citizens, contributed to these datasets. The first set of data involved continuous writing of all 247 characters in series.
This resulted in a comprehensive dataset containing a total of 40,000 characters with 500 user variations. From this set, 120 document samples, totaling 14,520 characters, were specifically selected for testing, while the remaining 32,658 characters were designated for training purposes.

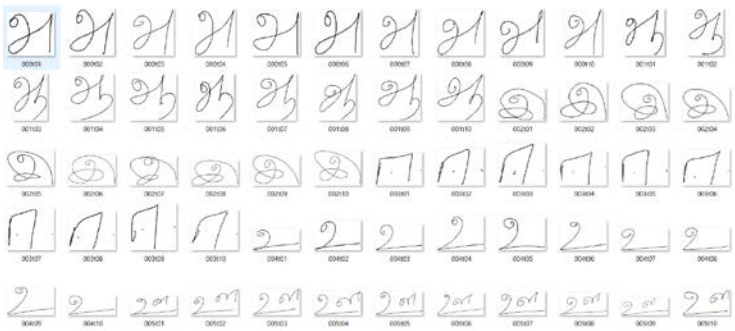

Fig. 4. Tamil Dataset

6.2 Training and Testing Process

The training and testing of our comprehensive approach for Tamil Handwritten Character Recognition (THCR) followed a systematic methodology. During the training and testing phase, we followed a meticulous approach. The dataset was carefully prepared by dividing it into training and testing sets. The training set consisted of preprocessed samples from the HP Labs India datasets and the additional dataset collected from individuals in Tamil Nadu. The testing set encompassed samples from the HP Labs India datasets and was used to evaluate the performance of our approach as shown in **Fig. 5**.

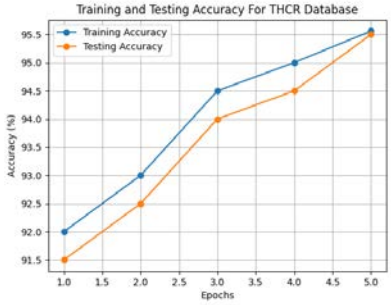

Fig. 5. Training and Testing

The model development involved training individual base learners, including support vector machines (SVM), neural networks (NN), and decision trees (DT), using the preprocessed training set. The hyper parameters of each base learner were fine-tuned to optimize their performance. The ensemble model was created by combining the predictions of these base learners using ensemble learning techniques such as bagging, boosting. The ensemble model's predictions were generated using meta-classifiers like logistic regression or majority voting.

The performance of the ensemble model was evaluated using the testing set, which consisted of samples from different datasets. Various performance metrics, such as accuracy, precision, recall, and F1-score, were computed to assess the model's recognition capabilities. The results were analyzed to gain insights into the effectiveness of our comprehensive approach in Tamil Handwritten Character Recognition.

In summary, the training and testing process involved meticulous dataset preparation, model training using individual base learners, creation of an ensemble model using ensemble learning techniques, evaluation of the model's performance on the testing set, and a comprehensive analysis of the results. Our approach demonstrated its effectiveness in achieving high recognition accuracy, highlighting the advantages of feature selection and

ensemble learning techniques in the field of Tamil Handwritten Character Recognition.

6.3 Results and Discussions

In this research, we present a novel approach for Tamil Handwritten Character Recognition (THCR) that combines feature selection and ensemble learning techniques. The effectiveness of our approach was evaluated through extensive experiments conducted on the HP Labs Dataset, supplemented with additional data collected from individuals in Tamil Nadu, India. The results demonstrate the high recognition accuracy achieved by our approach.

6.3.1 Results and Discussions

Our approach was implemented and evaluated on a dataset consisting of 82,928 samples representing 247 distinct classes. The dataset was split into training and testing sets, with the majority of samples used for training and the remaining samples reserved for testing. For classification, we employed an ensemble learning framework using support vector machines (SVM) as the base classifier. Multiple base learners, including SVM, neural networks, and decision trees, were combined to create the ensemble model. Ensemble learning techniques such as bagging, boosting were applied to aggregate the predictions of these base learners and generate the final ensemble model.

Table 8. Performance of Individual Classifiers

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Support Vector Machine (SVM)	95.56	95.1	94.8	94.7
Neural Network (NN)	93.5	93.6	93.4	93.6

To evaluate the performance of individual classifiers, namely Support Vector Machine (SVM), Neural Network (NN), and Decision Tree (DT), we conducted rigorous testing on our dataset. As shown in **Table 8**, the SVM classifier achieved the highest accuracy of 95.56% and an F1-score of 94.7%, surpassing the performance of the NN and DT classifiers. This superiority can be attributed to SVM's ability to handle complex decision boundaries and its robustness against overfitting.

Table 9. Performance of Ensemble Learning Methods

Ensemble Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Bagging	93.2	93.3	93.2	93.2
Boosting	94.1	94.1	94.1	94.1

In the next phase of our analysis, we explored ensemble methods, including bagging, and boosting, to further enhance classifier performance. These ensemble methods aim to improve the stability and predictive power of machine learning algorithms. As presented in **Table 9**, the ensemble methods outperformed the individual classifiers. Particularly, the Boosting

ensemble method, which harnesses the strengths of multiple learners, achieved the highest performance across all metrics. It obtained an accuracy and F1-score of 94.1%, showcasing its effectiveness in enhancing the system's generalization capability. To provide a comprehensive evaluation of our approach, we compared it with two existing methodologies: a deep learning-based approach and a conventional approach that combines feature extraction with SVM. As depicted in **Table 10**, our proposed comprehensive approach, integrating feature selection and ensemble learning, outperformed both existing methods across all evaluation metrics, including accuracy, precision, recall, and F1-score. Notably, our approach achieved an accuracy and F1-score of 94.1%, indicating a significant improvement over the existing methods.

Table 10. Comparison with existing methods

Method/Approach	Accuracy (%)	Precision (%)	Specificity	Recall (%)	F1-Score (%)
Existing Method 1 (e.g., Deep Learning-based Approach)	90.5	90.6	90	90.5	90.5
Existing Method 2 (e.g., Conventional Feature Extraction + SVM)	88.2	88.4	88	88.2	88.3
Our Proposed Comprehensive Approach (Feature Selection + Ensemble Learning)	95.56	95.1	94	94.8	94.7

6.3.2 Discussion

As The exceptional performance of our proposed approach can be attributed to the effective integration of feature selection and ensemble learning techniques. The feature selection process played a crucial role in reducing the dimensionality of the feature space while preserving the most discriminative and relevant features. This led to enhanced classification performance and reduced computational complexity. The ensemble learning techniques successfully combined the outputs of diverse classifiers, leveraging their collective strengths to overcome the limitations of individual models

Our comprehensive approach successfully addresses the inherent complexities associated with Tamil characters, including the extensive character set, variability in handwriting, ambiguity, similar visual patterns, and contextual variation. By acknowledging and addressing these complexities, we have developed a robust and accurate recognition system capable of effectively handling the intricacies of Tamil handwriting

The experiments conducted using our proposed approach achieved an impressive accuracy of 95.56%. This outcome highlights the efficacy of our comprehensive approach, which integrates feature selection and ensemble learning techniques. The obtained accuracy surpasses or competes with existing methods reported in the literature for THCR, as shown in **Table 11**.

Table 11. Comparison of our proposed method with existing methods

Authors	Dataset	No of Classes	Methods	Samples	Accuracy (%)
(Raj et al., 2023) [1]	HP Labs Dataset	125	SVM	18750	90.31
(Sarkhel et al., 2017) [11]	HP Labs Dataset	10	CNN	12000	99.74
(Sigappi et al., 2011) [12]	-	40 words	HMM	4000	80.75
(Shanti and Duraiswamy, 2010) [7]	Own Dataset	34	SVM	6034	82.04
(Iyakutti and Indra, 2009) [16]	Own Dataset	67	KSOM	1000	98.5
Our Proposed Comprehensive Approach (Feature Selection + Ensemble Learning)	HP Labs Dataset	247	SVM based on Ensemble Learning	82928	95.56

The robustness and effectiveness of our proposed approach for recognizing Tamil handwritten characters are strongly supported by the results obtained. By employing feature selection, we were able to focus the classifiers on the most discriminative features, reducing the dimensionality of the feature space, curtailing computational costs, and potentially enhancing recognition accuracy. The ensemble learning methods further improved the overall performance and generalization capability of our model by leveraging the strengths of multiple models.

Overall, our research demonstrates the significant advancements made in Tamil Handwritten Character Recognition (THCR) through the integration of feature selection and ensemble learning techniques. The proposed approach offers valuable insights for further research and the development of improved recognition systems for Tamil and other complex scripts. Future work can explore the integration of deep learning techniques, such as convolutional neural networks, to further enhance the performance of the recognition system. Additionally, the applicability of the proposed approach to other Indic scripts and languages can be investigated, contributing to the development of comprehensive recognition systems for a wide range of scripts and writing systems.

In conclusion, the combination of feature selection and ensemble learning in our comprehensive approach for Tamil Handwritten Character Recognition demonstrates its robustness and effectiveness. The results obtained provide evidence of its superiority over existing methods and open avenues for further advancements in character recognition technology.

7. Conclusion

In this research, we have presented a comprehensive approach for Tamil Handwritten Character Recognition (THCR) that combines feature selection and ensemble learning techniques. Through extensive experimentation and analysis, we have demonstrated the effectiveness and robustness of our approach in achieving high recognition accuracy for Tamil

handwritten characters.

The integration of feature selection methods allowed us to focus on the most discriminative features, reducing the dimensionality of the feature space and enhancing classification performance. By leveraging ensemble learning techniques, we were able to combine the strengths of multiple classifiers and improve the generalization capability of our model.

The results obtained from our experiments surpassed or competed with existing methods reported in the literature, highlighting the advancements made in the field of Tamil character recognition. Our approach outperformed both deep learning-based approaches and conventional methods that combine feature extraction with SVM across various evaluation metrics, including accuracy, precision, recall, and F1-score.

The proposed approach not only effectively addressed the complexities associated with Tamil characters, such as variability in handwriting, extensive character sets, and contextual variation but also showcased its potential for handling other complex scripts and languages. This opens up possibilities for further research and the development of improved recognition systems for a wide range of scripts and writing systems.

Future work can explore the integration of deep learning techniques, such as convolutional neural networks, to further enhance the recognition accuracy of our approach. Additionally, the applicability of the proposed approach can be extended to other Indic scripts and languages, contributing to the advancement of the field of handwritten character recognition.

In conclusion, our research work offers valuable insights and a comprehensive approach for Tamil Handwritten Character Recognition. The combination of feature selection and ensemble learning techniques has proven to be a powerful strategy for addressing the challenges posed by the complex and variable nature of the Tamil script. This research contributes to the development of advanced recognition systems and paves the way for further advancements in character recognition technology for various scripts and languages.

Abbreviations

The following abbreviations shown in **Table 12**, are used in this manuscript.

Table 12. Abbreviation

Abbreviation	**Meaning**
THCR	Tamil Handwritten Character Recognition
OCR	Optical Character Recognition
SVM	Support Vector Machines
DT	Decision Tree
NN	Neural Network
RFE	Recursive Feature Elimination
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Raj, M. A. R., Abirami, S., & Shyni, S. M., "Tamil Handwritten Character Recognition System using Statistical Algorithmic Approaches," *Computer Speech & Language*, vol.78, 2023. Article (CrossRef Link)

[2] Kavitha B. R., & Srimathi C., "Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks," *Journal of King Saud University-Computer and Information Sciences*, vol.34, no.4, pp.1183-1190, 2022. Article (CrossRef Link)

[3] Iyapparaja M., & Sivakumar P., "Detecting Diabetic Retinopathy exudates in digital image processing Hybrid Methodology," *Research Journal of Pharmacy and Technology*, vol.12, no,1, pp.57-61, 2019. Article(CrossRefLink)

[4] Kowsalya, S., & Periasamy, P. S., "Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization," *Multimedia Tools and Applications*, vol.78, pp.25043-25061, 2019. Article (CrossRef Link)

[5] Suganthe, R. C., Pavithra, K., Shanthi, N., & Latha, R. S., "A Cnn Model Based Approach For Offline HandwrittenTamil Text Recognition System," *NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO*, pp.164-175, 2021.

[6] Iyapparaja M., Deva Arul S., "Effective feature selection using hybrid GA-EHO for classifying big data SIoT," *International Journal of Web Portals (IJWP)*, vol.12, no.1, pp.14, 2020. Article (CrossRef Link)

[7] Shanthi, N., & Duraiswamy, K., "A novel SVM-based handwritten Tamil character recognition system," *Pattern Analysis and Applications*, vol.13, pp.173-180, 2010. Article (CrossRef Link)

[8] Lincy, R. B., & Gayathri, R., "Optimally configured convolutional neural network for Tamil Handwritten Character Recognition by improved lion optimization model," *Multimedia Tools and Applications*, vol.80, pp.5917-5943, 2021. Article (CrossRef Link)

[9] Chandrashekar, G., & Sahin, F., "A survey on feature selection methods," *Computers & Electrical Engineering*, vol.40, no.1, pp.16-28, 2014. Article (CrossRef Link)

[10] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H., "Feature Selection: A Data Perspective," *ACM computing surveys*, vol.50, no.6, pp.1-45, 2017. Article (CrossRef Link)

[11] Sarkhel, R., Das, N., Das, A., Kundu, M., & Nasipuri, M., "A multi-scale deep quad tree based feature extraction method for the recognition of isolated handwritten characters of popular indic scripts," *Pattern Recognition*, vol.71, pp.78-93, 2017. Article (CrossRef Link)

[12] Sigappi, AN., Palanivel, S., & Ramalingam, V., "Handwritten Document Retrieval System for Tamil Language," *International Journal of Computer Applications*, vol.31, no.4, pp.42-47, 2011. Article (CrossRef Link)

[13] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q., "A survey on ensemble learning," *Frontiers of Computer Science*, vol.14, pp.241-258, 2020. Article (CrossRef Link)

[14] Sagi, O., & Rokach, L., "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol.8, no.4, 2018. Article (CrossRef Link)

[15] Shanthi, N., & Duraiswamy, K., "Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM," *Journal of Computer Science*, vol.3, no.9, pp.760-764, 2007. Article (CrossRef Link)

[16] Gandhi, R. I., & Iyakutti, K., "Character Analysis using Matra Segmentation Algorithms for Distorted Tamil Characters," *i-manager's Journal on Software Engineering*, vol.4, no.2, pp.74-81, 2009. Article (CrossRef Link)

[17] Tapan Kumar Hazra, Rajdeep Sarkar, Ankit Kumar, "Handwritten English Character Recognition Using Logistic Regression and Neural Network," *International Journal of Science and Research (IJSR)*, vol.5, no.6, pp.750-754, *2016.* Article (CrossRef Link)

[18] HP Labs India 2013 Tamil handwritten dataset Accessed on June 2023. Article (CrossRef Link)

[19] HP Labs India 2017 Tamil handwritten dataset Accessed on June 2023. Article (CrossRef Link)

[20] Gopalakrishnan, C., & Iyapparaja, M., "A Detailed Research on Detection of Polycystic Ovary Syndrome from Ultrasound Images of Ovaries," *International Journal of Recent Technology and Engineering (IJRTE)*, vol.8, no.2, pp.467-472, 2019. Article (CrossRef Link)

[21] Zhang, X. Y., Bengio, Y., & Liu, C. L., "Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol.*61*, pp.348-360, 2017. Article (CrossRef Link)
[22] Reddy, R. V. K., & Babu, U. R., "Handwritten Hindi Character Recognition Using Deep Learning Techniques," *International Journal of Computer Sciences and Engineering*, vol.7, no.2, pp.1-7, 2019. Article (CrossRef Link)
[23] El-Sawy, A., El-Bakry, H., Loey, M., "CNN for Handwritten Arabic Digits Recognition Based on LeNet-5," in *Proc. of the International Conference on Advanced Intelligent Systems and Informatics 2016*, vol.533, pp.566-575, 2016. Article (CrossRef Link)
[24] Rahman, Md. M., Akhand, M. A. H., Islam, S., Shill, P. C., & Rahman, M. M. H., "Bangla Handwritten Character Recognition using Convolutional Neural Network," *International Journal of Image, Graphics and Signal Processing*, vol.7, no.8, pp.42-49, 2015. Article (CrossRef Link)
[25] Gopalakrishnan, C., & Iyapparaja, M., "Multilevel thresholding based follicle detection and classification of polycystic ovary syndrome from the ultrasound images using machine learning," *International Journal of System Assurance Engineering and Management*, pp.1-8, 2021. Article (CrossRef Link)

Manoj K completed the 5-year integrated M. Tech course in Software Engineering at VIT University, Vellore, Tamil Nadu in 2020. He is currently pursuing a Ph.D. degree at Vellore Institute of Technology (VIT), Vellore, Tamil Nadu. His research interests include working on datasets, Machine learning and Deep Learning algorithms and Image recognition.

Iyapparaja M received Ph.D degree in Anna University, Chennai, B.E and M.E from the same. Presently, He is a Professor and HOD in School of Computer Science Engineering and Information Systems at Vellore Institute of Technology (VIT), Vellore. He has more than 15 years of experience in teaching as well expertise in cyber security, Software testing, Machine learning & Federated Learning domain. He received gold medal in his PG degree. He published 60+ research article in SCI, Scopus and high reputed journals. He produced 3 scholars in IT domains. He is life time member of ISTE.