Received: July 27, 2021. Revised: August 25, 2021. 301

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Handwritten Character Recognition Using Unsupervised Feature Selection and Multi Support Vector Machine Classifier

Asha Kathigi[1]* Krishnappa Honnamachanahalli Kariputtaiah[2]

[1]*GM Institute of Technology, Davangere, Karnataka, India*
[2]*RV College of Engineering, Bangalore, India*
* Corresponding author's Email: ashak@gmit.ac.in

Abstract: In recent times, identifying Kannada, Arabic and English handwritten characters is a challenging task in pattern recognition application. The low resolution, complex backgrounds, text orientation, text size, and the variations in the writing styles makes character recognition as a challenging task. To address the above stated issues, a new automated character recognition model is introduced in this paper. Firstly, skewed line segmentation technique is applied to the handwritten character for dissecting the document images into line elements and then split into phrases and single characters. Next, AlexNet model is used for extracting the deep features from the individual characters, where the extracted feature vectors are multi-dimensional in nature that increases the system complexity. So, an unsupervised feature selection algorithm is proposed to select the active feature vectors that are fed to multi support vector machine classifier for individual character classification such as 64 classes in English language, 10 classes in Arabic language, and 657 classes in Kannada language. Experimental analysis showed that the proposed model obtained 85.80%, and 95.55% of accuracy on the chars74K dataset for Kannada, and English characters, respectively. In addition, the proposed model obtained 71.79%, and 99.97% of recognition accuracy in Kannada and Arabic handwritten character recognition on a real time dataset and MADbase digits dataset. The obtained results are better compared to the existing deep learning models; DIGI-Net, feed forward neural network, and context aware model.

Keywords: AlexNet, Handwritten character recognition, Multi support vector machine, Skewed line segmentation technique, Unsupervised feature selection.

1. Introduction

In recent decades, the automated handwritten character recognition frameworks include many applications such as forensic research, zip code identification, writer recognition, etc. [1]. The aim of such frameworks is to automatically extract, and recognize the characters, which are embedded in the manuscript, where the quality of the manuscript is also an essential factor for enhancing the recognition accuracy [2, 3]. In addition to this, it is necessary to deal with the other concerns, which occur in the manuscripts like background noise appeared during the scanning process, and distortions in a character image [4, 5]. In recent periods, uttermost research works are carried out on the offline handwritten character recognition for the scripts; Tamil, Kannada, Bangla, etc. [6, 7]. Though, the Kannada handwritten character recognition is vital in the applications like banking, post office etc. [8]. The recognition of Kannada and Arabic characters in the handwritten and printed documents is a difficult process, since the numerals have more angles and curves [9, 10]. In addition, external factors like slided numerals, number of holes and strokes, and different writing styles affects the detection performance [11, 12]. The main aim of this research is to propose a novel model, which obtain higher recognition accuracy in the scripts; Kannada, Arabic and English.

Firstly, handwritten Kannada, Arabic and English characters are acquired from chars74K, MADbase digits dataset and a real time dataset. The intended operations like segmentation, feature extraction, feature selection and classification are performed well by collecting the data from proper datasets. Then,

segmentation is performed on the characters by using skewed line segmentation technique for distinguishing the overlapped characters into distinct characters. Next, AlexNet model is used to extract the feature vectors from the segmented characters. The AlexNet model extracts more accurate object features that results in high classification accuracy, even in lower lighting conditions. The extracted deep feature vectors are multi-dimensional in nature that leads to "curse of dimensionality" issue and increases the system complexity. So, UFS algorithm is proposed in this paper for selecting the discriminative feature vectors for better character recognition. The selected feature vectors are fed to MSVM to classify the Kannada, Arabic, and English characters; 657 classes in Kannada language, 10 classes in Arabic language, and 64 classes in English language. In the resulting phase, the proposed UFS-MSVM model performance is validated in light of accuracy, f-score, sensitivity, specificity and Matthews's correlation coefficient (MCC).

This paper is arranged as follows. Some recent research papers on the topic "handwritten character recognition" are surveyed in the Section 2. The UFS-MSVM model is briefly detailed in the Section 3. The quantitative, and comparative analysis of UFS-MSVM model is indicated in the Section 4, and the conclusion of this research work is stated in the Section 5.

2. Related works

Chandio [13] implemented a Multi-Level Feature Fusion (MLFF) and a Multi Scale Feature Aggregation (MSFA) networks for recognizing Urdu characters. At first, up sampling, and addition operations were used to aggregate the multi scale feature vectors of the convolution layers, and then the respective feature vectors were combined with the high level feature vectors. Finally, the outputs of the MLFF and MSFA networks are merged together for generating a more powerful and discriminative feature vectors. In this study, the developed network performance was validated on ICDAR03 and Chars74K datasets in terms of precision, f-score and recall. Hence, the developed network has a major issue of image degradation, which occurs due to multi orientated text, and un-even lighting conditions. Akhtar [14] presented an Optical Character Recognition (OCR) system on the basis of multi-characteristics feature fusion, and selection techniques. The extracted feature vectors were fused using serial formulation, and then the selection was accomplished using partial least square selection approach which works on the basis of entropy fitness function. Lastly, the selected feature vectors were fed to the ensemble classifier for final classification. The simulation results confirmed that the presented model obtained significant performance in OCR on Chars74k dataset. Still, the presented model faced difficulties in segmenting the characters, which have multi oriented structure.

Sampath and Gomathi [15] implemented a hybrid neural network model for English handwritten character recognition. Firstly, image denoising was performed using median filter and then structural, positional and feature set were extracted from the denoised images. After extracting the features, Firefly and Levenberg Marquardt (FLM) algorithm was integrated with Feed Forward Neural Network (FFNN) for classification. The simulation result showed that the prescribed model obtained better performance in English handwritten character recognition compared to prior deep learning models. However, the hybrid neural network model showed comparable performance in real scene dataset (camera captured text image), because it was affected from environmental noise, background complexity, and deformations. Madakannu and Selvaraj [16] introduced DIGI Net model for learning feature vector and recognizing printed font, natural images and handwritten images. In this study, DIGI Net model attained effective performance on three benchmark datasets like Chars74K, MNIST and CVL single digit datasets. Hence, the DIGI Net model finds difficulties in recognizing un-constrained natural images, due to large appearance variability.

Hallur and Hegadi [17] introduced a new deep learning model for Kannada numeral recognition. After collecting handwritten Kannada data, pre-processing was carried out by utilizing thinning, normalization, binarization, skew amendment and noise removal. Further, feature vectors were extracted from the denoised images using curvelet transfiguration wrapping, discrete wavelet transform, drift length count and direction related progression code. Lastly, data classification was performed using deep Convolution Neural Network (CNN) classifier. Experimental results showed that the presented deep learning model obtained significant performance in handwritten Kannada numerals recognition by means of accuracy. However, the high declination in Kannada digits leads to more issues in the recognition process. Sampath and Gomathi [18] extracted feature vectors from the chars74k dataset by applying histogram of oriented gradient descriptor. The obtained features were given as the input to fuzzy based multi-kernel spherical support vector machine classifier for character classification. The simulation analysis showed that the prescribed model obtained

better performance in English character recognition on the chars74K dataset by means of false rejection rate, classification accuracy, and false acceptance rate. Still, English handwritten character recognition was complex in this study, because English characters differ from writing devices, shapes and styles.

Ahmed [19] developed a novel context aware model on the basis of deep neural networks for addressing the issues of detecting offline handwritten Arabic characters. Alkhawaldeh [20] presented a new model; adapted deep hybrid transfer model using CNN and long short term memory network for Arabic character recognition. In this literature, the CNN model learns the relevant feature vectors of Arabic characters, and the long short term memory network extract the long term dependence features. The developed model obtained comparable performance in real time Arabic handwritten character detection, due to dots, cursive writing, loops, overlapping, ligatures and diacritics. In order to address the above mentioned issues, a new model named UFS-MSVM is proposed to enhance the performance of Kannada, Arabic and English handwritten character recognition.

3. Methodology

The proposed handwritten character recognition model consists of five major phases like image collection: chars74K, MADbase digits dataset, and a real time dataset, segmentation: skewed line segmentation technique, feature extraction: AlexNet features, feature selection: UFS based on subspace randomization and collaboration and classification: MSVM classifier. The flow diagram of the UFS-MSVM model is graphically specified in Fig. 1.

3.1 Image collection

In this work, the proposed UFS-MSVM model performance is validated on chars74K dataset, where it comprises of both the Kannada and English characters [21]. In this work, Kannada and English handwritten characters are utilized for performance analysis, and experimentation. English characters comprise of 64 classes that composed of a lower case, numbers and upper case (a-z, 0-9 and A-Z). English characters include 7705 natural images, 62992 synthesized images, and 3410 handwritten images. Additionally, the Kannada characters cannot be differentiated between lower and upper case characters. Kannada language has 49 basic characters in alpha syllabary, but vowels and consonants combine to give 657 visually distinct classes. Sample images of chars74K dataset is denoted in Fig. 2. In addition to this, real time data are collected for Kannada language, which includes 657 handwritten Kannada characters. Sample image is graphically shown in Fig. 3.

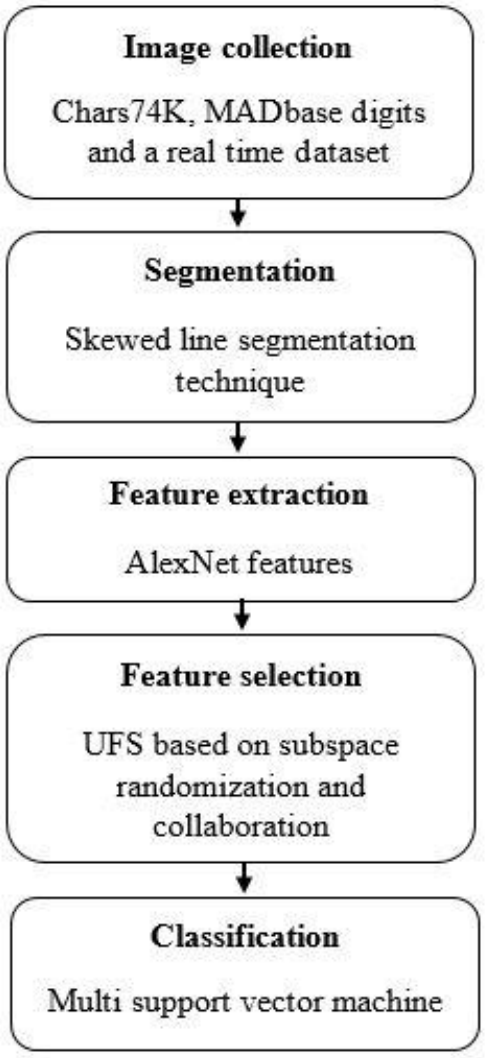
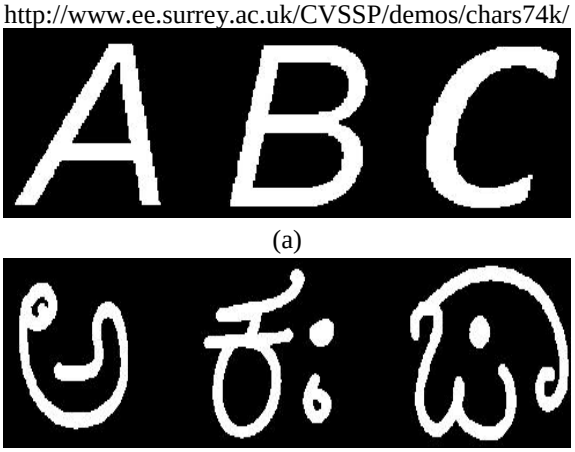


Figure. 1 Flow diagram of proposed UFS-MSVM model

Dataset link of chars74K dataset:
http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/



(a)

(b)

Figure. 2 Sample images of chars74K dataset: (a) English characters and (b) Kannada characters

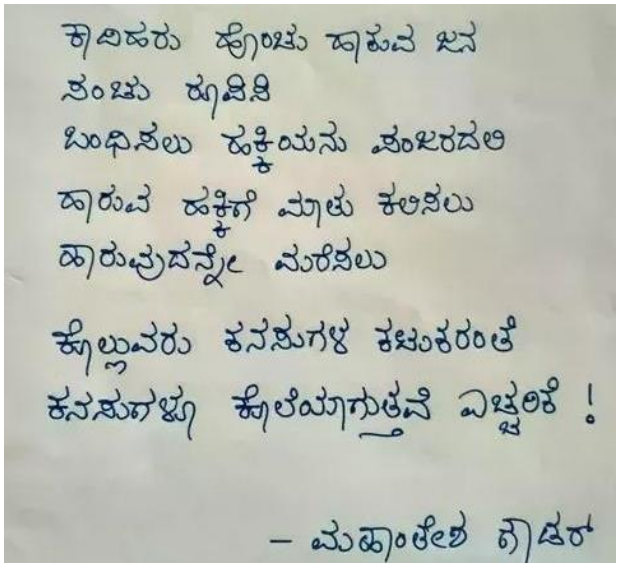


Figure. 3 Sample image of real time handwritten Kannada characters

Dataset link of MADbase digits dataset:
http://datacenter.aucegypt.edu/shazeem/

Figure 4. Sample handwritten characters of MADbase digits dataset

In addition, Arabic handwritten character recognition is accomplished on MADbase digits dataset that consists of 10,000 testing, and 60,000 training handwritten images, which are written by 700 writers [22]. In MADbase digits dataset, the images are acquired from dissimilar institution to ensure dissimilar writing styles such as college of engineering and law, Open University, high school, and government institutions. The sample handwritten characters of MADbase digits dataset is graphically denoted in Fig. 4.

3.2 Segmentation

After collecting the characters, skewed line segmentation technique is applied to the real time handwritten characters for distinguishing the individual characters [23]. At first, the skew correction technique is applied to the collected handwritten images, which works based on image pixel intensity information. The aim of skew correction technique is to distinguish the area between text sentences or lines. Skew correction technique generally works on page level, and it identifies the titled line angles and rotate the images around center pixel. Next, line segmentation is carried out on the skew less images to distinguish black text and white background regions. Next, the pixel strength is computed for black text in the handwritten documents that finds the threshold value of text image. Hence, the standard deviation value of the document is less than the rows having dark pixels. The text line (combination of successive rows) are extracted as pixels, which lies between footer and header.

The two lines in a document are separated, while adaptive threshold in rows of the document is higher than the black pixels in a row. The text line is distinguished using two parameters; header (first row of the line), and baseline (ending point of the line), which are determined by the number of white and black pixels in rows. In this scenario, adaptive threshold value is calculated based on the standard deviation of the text pixels. Lastly, text segmentation is carried out using projection profile technique. For ligature segmentation, the text lines are converted into words using the vertical profile technique. In this research, the collected Kannada and Arabic characters are considered as input. Firstly, the handwritten document is segmented into many text lines and fed to ligature/word segmentation algorithm that distinguishes the text lines into smaller ligatures. Hence, the ligatures are automatically arranged in sequence, because the algorithm segment words in a sequential order. The output of line segmentation and ligatures segmentation are shown in Fig. 5 and 6.

Figure. 5 Output of line segmentation

Figure. 6 Output of ligatures segmentation

3.3 Feature extraction and selection

After segmenting the individual ligatures, AlexNet (pre-trained CNN) model is applied for feature extraction. The AlexNet model is an eight-layer neural network, which includes three fully connected layers, and five convolutional layers. Every convolutional layer performs max pooling operations by Rectifier Linear Unit (ReLU) [24-25]. In addition, the fully connected layer delivers output using softmax classifier. The convolutional layer's extracts feature vectors from segmented images by performing convolutional operations, and it utilizes filters to extract the feature vectors. Generally, the size of the convolutional layer depends on the filter size, and number of filters. The ReLU activation function replaces all the negative pixels of feature map by zero, and the pooling layer is utilized for reducing the dimensions of feature map.

Different pooling operations are performed in feature extraction such as sum pooling, average pooling, and max pooling. In this research study, max pooling is utilized in AlexNet model, where the largest element is considered for the defined neighbourhoods. The AlexNet model extracts the feature vectors in a hierarchical manner, where the sub-sequent layers extract higher level feature vectors, and the initial layers extracts lower level feature vectors. The total extracted 4096 feature vectors are fed to UFS algorithm to select the active feature vectors for better classification. In the UFS algorithm, a multi-subspace randomization and collaboration technique is applied for creating several feature partitions with same sub-space size [26]. The proposed UFS algorithm utilizes two parameters q and e to adjust the number of created sub-spaces. Where, q is specified as number of random sub-spaces, and e is indicated as number of feature partitions. Based on the created random sub-spaces $q \times e$, K nearest neighbor and Laplacian graphs are constructed in order to preserve the locality power. Every feature partition's full vector score is calculated by integrating partial vector score of multiple sub-spaces. By computing the mean feature score of the feature partitions, overall feature vector score is determined. Final outcome is achieved by arranging the overall feature score in an ascending order. Feature partition G^i, and random sub-space G^{ij} are mathematically depicted in the Eqs. (1) and (2).

$$G^i = \{G^{i,1}, G^{i,2}, \ldots . G^{i,q}\} \tag{1}$$

$$G^{ij} = (g_1^{(i,j)}, g_2^{(i,j)}, \ldots . g_{D_{ij}}^{(i,j)}) \tag{2}$$

The multiple basic features, and the constructed KNN graph are denoted in the Eqs. (3) and (4).

$$G = \{G^1, G^2, \ldots G^e\} \tag{3}$$

$$Lap^{i,j} = \{X, Adj^{i,j}\} \tag{4}$$

Where, $Adj^{i,j}$ is denoted as adjacent matrix and X is stated as node set that consists of data samples. Next, calculate the Laplacian score of the features $g_{D_{ij}}^{(i,j)}$ with sub-space G^{ij}, and then compute the Laplacian score of D_{ij} with subspace $G^{(i,j)}$. By averaging the Laplacian score vectors of $e = D_{ij} + g_{D_{ij}}^{(i,j)}$, the final score vector G_{final} is obtained. The total selected relevant feature vectors G_{final} are 1000, which are fed to MSVM classifier for character recognition.

3.4 Classification

The SVM classification technique is developed for binary classification problems in the image processing application. It is essential to develop a multi-SVM classifier with hierarchical structure in order to deal with the multi class classification problems [27]. The MSVM classifier is used for decomposing the M class problems into a set of binary classification problem and then integrate M binary classification techniques. Most commonly used methodologies in multi-class classification problems are One against One (OAO) and One against All (OAA).

The OAO method generates $M_{OAO} = M_2(M_2 - 1)/2$ binary SVM classifiers that discriminates between class A and B. Similarly, OAA method generates $M_{OAA} = M_1$ binary SVM classifiers that distinguish one class from other classes. The i^{th} SVM classifier is trained with all the training sets of i^{th} class with positive labels, and the remaining classes are trained with negative labels. Lastly, the OAO, and OAA methodologies are integrated for constructing $M(M_1 + M_2)$ class problems in order to classify the individual Kannada, Arabic, and English characters. Here, polynomial kernel is used in MSVM classifier to deliver a better performance for classification problems. Polynomial kernel is mathematically depicted in Eq. (5). The trade-off between the training error rate, and the complexity of decision making rule is controlled by utilizing the degree of polynomial kernel d. The experimental analysis showed that the best result is obtained by polynomial kernel in MSVM, classifier, which is depicted in the Section 4.

$$K(x_1, x_2) = (\langle x_1 \times x_2 \rangle + 1)^d \quad (5)$$

4. Simulation results

The proposed handwritten character recognition model is simulated using MATLAB (2018a) environment. In this research, the effectiveness of the proposed UFS-MSVM model performance is validated by comparing with benchmark models; FLM-FFNN [15], DIGI Net model [16] and context aware model [19] and adapted deep hybrid transfer model [20] on chars74K and MADbase digits datasets. In addition, the performance of the proposed UFS-MSVM model is evaluated in terms of accuracy, MCC, f-score, sensitivity and specificity. Mathematical equations of undertaken performance measures are represented in the Eqs. (6) to (10). Whereas, TN indicates true negative, FN represents false negative, TP denotes true positive, and FN states false negative.

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \times 100 \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \quad (7)$$

$$F - score = \frac{2TP}{FP+2TP+FN} \times 100 \quad (8)$$

$$Sensitivity = \frac{TP}{FN+TP} \times 100 \quad (9)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (10)$$

4.1 Analysis of handwritten Kannada characters on chars74K dataset

In this segment, chars74K dataset (Kannada characters) is used for validating the performance of UFS-MSVM model. In chars74K dataset, 80% of the text images are utilized for model training, and the residual 20% of the text images are utilized for model testing. In this scenario, performance analysis is carried out by several classification techniques; K-Nearest Neighbor (KNN), random forest, decision tree and MSVM and feature selection algorithms; Person Correlation Coefficient (PCC), UFS and reliefF. As seen in Tables 1 and 2, the combination: UFS-MSVM showed better performance in handwritten character recognition related to the comparative classifiers and feature selection algorithms. In chars74K database, the combination: UFS-MSVM achieved a maximum classification accuracy of 85.8%, sensitivity of 88.65%, specificity of 85.68%, MCC of 86.5% and f-score of 85.22%, which are higher related to the comparative algorithms in Kannada handwritten character recognition.

4.2 Analysis of handwritten English characters on chars74K dataset

In this segment, chars74K dataset (3410 handwritten English images) is used for evaluating the performance of the UFS-MSVM model. In this circumstance, 80% (2728 images) of the English text images are used for model training, and 20% (682 images) of the English text images are used for model testing. By inspecting Tables 3 and 4, the combination: UFS-MSVM obtained maximum

Table 1. Performance evaluation of different classifiers in Kannada handwritten character recognition on chars74K dataset

UFS algorithm					
Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
KNN	78.44	85.19	77.71	81.66	78.96
Random forest	45.38	44.49	45.75	51.16	45.34
Decision tree	38.77	46.45	38.28	53.12	37.59
MSVM	85.80	88.65	85.68	86.50	85.22

Table 2. Performance evaluation of different feature selection algorithms in Kannada handwritten character recognition on chars74K dataset

MSVM classifier					
Feature selection algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
Without feature selection	82.68	85.14	81.22	85.45	81.22
ReliefF algorithm	83.20	85.23	83.44	85.77	81.46
PCC	82.83	86.10	83.86	86.05	83.45
UFS	85.80	88.65	85.68	86.50	85.22

Table 3. Performance evaluation of different classifiers in English handwritten character recognition on chars74K dataset

UFS algorithm					
Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
KNN	91.35	91.92	91.38	88.53	91.57
Random forest	85.83	85.29	85.57	82.78	85.47
Decision tree	82.59	82.76	82.08	84.93	82.01
MSVM	95.55	94.75	96.05	93.13	91.34

Table 4. Performance evaluation of different feature selection algorithms in English handwritten character recognition on chars74K dataset

MSVM classifier					
Feature selection algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
Without feature selection	90.37	89.71	88.45	87.56	88.12
ReliefF algorithm	94.32	93.13	92.17	91.26	92.67
PCC	93.04	92.20	91.64	90.49	92.07
UFS	95.55	94.75	96.05	93.13	93.34

classification accuracy of 95.55%, sensitivity of 94.75%, specificity of 96.05%, f-score of 91.34%, and MCC of 93.13%. The obtained results are better in English handwritten character recognition related to the comparative classifiers and feature selection algorithms on chars74K dataset. Hence, the MSVM classifier easily determines the location of every sub-classifier that helps in achieving better classification result and also it has limited un-classifiable regions compared to other machine learning classifiers.

4.3 Analysis of handwritten Kannada characters on a real time dataset

In this segment, real time dataset (Kannada characters) is used for validating the performance of proposed UFS-MSVM model by means of accuracy, MCC, f-score, sensitivity, and specificity. Here, Kannada characters belong to chars74K dataset are utilized for model training, and the real time Kannada characters are utilized for model testing. By evaluating Tables 5 and 6, the proposed UFS-MSVM model achieved significant classification accuracy of 71.79%, sensitivity of 72.21%, specificity of 83.3%, MCC value of 70.8%, and f-score of 74.5% in Kannada handwritten character recognition on the real time dataset, which are better related to the comparative classifiers, and feature selection algorithms. In this research, UFS algorithm effectively chooses the active feature vectors from the total extracted features that helps in improving the character recognition performance.

4.4 Analysis of handwritten Arabic characters on MADbase digits datasets

In this segment, the MADbase digits dataset is used for analysing the performance of UFS-MSVM model. In this scenario, 10,000 handwritten Arabic images are used for testing, and 60,000 images are used for training. By viewing tables 7 and 8, the combination: UFS-MSVM obtained maximum classification accuracy of 99.97%, sensitivity of 99.33%, specificity of 96.64%, f-score of 98.44%,

Table 5. Performance evaluation of different classifiers in Kannada handwritten character recognition on real time dataset

UFS algorithm					
Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
KNN	64.1	62.5	73.33	69.53	71.43
Random forest	65.66	54.55	66.7	71.24	70.59
Decision tree	63.2	66.57	63.1	66.57	71.34
MSVM	71.79	72.21	83.33	70.89	74.5

Table 6. Performance evaluation of different feature selection algorithms in Kannada handwritten character recognition on real time dataset

MSVM classifier					
Feature selection algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
Without feature selection	68.53	68.1	76.57	68.78	69.12
ReliefF algorithm	69.45	69.45	69.45	69.45	69.45
PCC	70.81	71.53	75.53	70.15	74.28
UFS	71.79	72.21	83.33	70.89	74.5

Table 7. Performance evaluation of different classifiers in Arabic handwritten character recognition on MADbase digits dataset

UFS algorithm					
Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
KNN	92.49	93.08	96.20	90.13	95.34
Random forest	91.39	89.50	90.81	89.17	92.44
Decision tree	84.77	84.60	85.23	86.00	82.56
MSVM	99.97	99.33	96.64	98.91	98.44

Table 8. Performance evaluation of different feature selection algorithms in Arabic handwritten character recognition on MADbase digits dataset

MSVM classifier					
Feature selection algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC (%)	F-score (%)
Without feature selection	91.05	92.46	91.56	89.57	88.74
ReliefF algorithm	98.67	93.85	93.92	91.64	93.59
PCC	95.94	96.47	94.21	91.69	93.27
UFS	99.97	99.33	96.64	98.91	98.44

and MCC of 98.91% on MADbase digits dataset, where the obtained results are better related to comparative classifiers and feature selection algorithms.

4.5 Analysis of handwritten Arabic characters on MADbase digits datasets

In this segment, the MADbase digits dataset is used for analysing the performance of UFS-MSVM model. In this scenario, 10,000 handwritten Arabic images are used for testing, and 60,000 images are used for training. By viewing Tables 7 and 8, the combination: UFS-MSVM obtained maximum classification accuracy of 99.97%, sensitivity of 99.33%, specificity of 96.64%, f-score of 98.44%, and MCC of 98.91% on MADbase digits dataset, where the obtained results are better related to comparative classifiers and feature selection algorithms.

4.6 Comparative analysis

The comparative investigation between the proposed UFS-MSVM model and the existing models is stated in the Tables 9 and 10. Sampath and Gomathi [15] developed a novel hybrid neural network for English handwritten character recognition. Initially, median filter was used for image denoising, where the images were collected from chars74K dataset. Further, feature extraction was performed using G-descriptor, and H-descriptor to extract the most discriminative feature vectors. The obtained optimal feature vectors were fed to FLM-FFNN model for character recognition. The simulation results showed that the FLM-FFNN model achieved 95% of accuracy in English hand-written character recognition. Additionally, Madakannu and Selvaraj [16] has implemented a novel DIGI-Net CNN model for automatic feature learning, and recognizing English printed font, natural images, and handwritten images. Experimental analysis showed that the developed DIGI-Net CNN model obtained 85% of classification accuracy in handwritten character recognition on chars74K dataset. Ahmed [19] developed a context aware model based on deep neural networks to highlight the issue of recognizing offline handwritten Arabic characters. Experimental outcome showed that the developed model obtained 99.91% of recognition accuracy and 99.13% of sensitivity on MADbase digits dataset. Alkhawaldeh [20] presented adapted deep hybrid transfer model by using CNN and long short term memory network for Arabic character recognition. Experimental results showed that the developed model attained 97.94% of accuracy and 99% of sensitivity on MADbase digits dataset. Compared to these research works, UFS-MSVM model showed better performance, and achieved better accuracy and sensitivity in English and Arabic character recognition.

In this paper, skewed line segmentation technique is used for segmenting the overlapped characters that results in effective performance in the circumstances; multi orientated text, and uneven lighting conditions. As seen in the experimental section, the skewed line segmentation technique is effective in environmental noise, deformations, and background complexity. In this research, global level features are extracted by AlexNet CNN model, which are high dimensional in nature that leads to system complexity. Though, UFS algorithm is proposed in this research paper to select the discriminative feature vectors for better character recognition and to reduce the system complexity. As seen in the quantitative analysis section, the proposed

Table 9. Comparative analysis on chars74K dataset

English handwritten character recognition on chars74K dataset	
Models	**Accuracy (%)**
FLM-FFNN [15]	95
DIGI-Net CNN [16]	85
UFS-MSVM	95.55

Table 10. Comparative analysis on MADbase digits dataset

Arabic handwritten character recognition on MADbase digits dataset		
Models	**Accuracy (%)**	**Sensitivity (%)**
Context aware model [19]	99.91	99.13
Adapted deep hybrid transfer model [20]	97.94	99
UFS-MSVM	99.97	99.33

UFS-MSVM model significantly address the problems mentioned in the literatures [13-20].

5. Conclusion

In this article, UFS-MSVM model is proposed to improve handwritten character recognition on Kannada, Arabic, and English languages. The proposed model comprises of two major phases; segmentation, and feature selection. In this research, Kannada, Arabic and English characters are acquired from real time dataset, MADbase digits dataset, and chars74K dataset. The skewed line segmentation technique is used in this article for effectively segmenting the overlapped characters. Next, AlexNet CNN model is applied for extracting features from the segmented text images. The extracted feature vectors are multi-dimensional in nature, so UFS algorithm is proposed to select the relevant features which reduced "curse of dimensionality" problem, and system complexity. Lastly, the selected features are fed to MSVM classifier to classify the handwritten text characters. The proposed UFS-MSVM model achieved 85.80%, 95.55% and 99.97% of recognition accuracy in Kannada, English, and Arabic languages. The proposed model showed minimum of 0.06% and maximum of 10.55% improvement in both English and Arabic handwritten character recognition compared to FLM-FFNN, context aware model, and DIGI-Net CNN. In future work, a new deep learning based model is developed for further improving the performance of handwritten character recognition, especially on Kannada, Arabic and English languages.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper background work, conceptualization, methodology, dataset collection, implementation, result analysis and comparison, preparing and editing draft, visualization have been done by first author. The supervision, review of work and project administration, have been done by second author.

References

[1] R. Anand, T. Shanthi, R. S. Sabeenian, and S. Veni, "Real time noisy dataset implementation of optical character identification using CNN", *International Journal of Intelligent Enterprise*, Vol. 7, No. 1-3, pp. 67-80, 2020.

[2] K. Asha and H. K. Krishnappa, "Kannada handwritten document recognition using convolutional neural network", In: *Proc. of 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions*, pp. 299-301, 2018.

[3] M. R. Phangtriastu, J. Harefa, and D. F. Tanoto, "Comparison between neural network and support vector machine in optical character recognition", *Procedia Computer Science*, Vol. 116, pp. 351-357, 2017.

[4] S. H. Lee, W. F. Yu, and C. S. Yang, "ILBPSDNet: Based on improved local binary pattern shallow deep convolutional neural network for character recognition", *IET Image Processing*, 2021.

[5] C. S. Yang and Y. H. Yang, "Improved local binary pattern for real scene optical character recognition", *Pattern Recognition Letters*, Vol. 100, pp. 14-21, 2017.

[6] N. Modhej, A. Bastanfard, M. Teshnehlab, and S. Raiesdana, "Pattern Separation Network Based on the Hippocampus Activity for Handwritten Recognition", *IEEE Access*, Vol. 8, pp. 212803-212817, 2020.

[7] A. K. Sampath and N. Gomathi, "Decision tree and deep learning based probabilistic model for character recognition", *Journal of Central South University*, Vol. 24, No. 12, pp. 2862-2876, 2017.

[8] N. D. Cilia, C. D. Stefano, F. Fontanella, and A. S. D. Freca, "A ranking-based feature selection approach for handwritten character recognition", *Pattern Recognition Letters*, Vol. 121, pp. 77-86, 2019.

[9] O. Surinta, M. F. Karaaba, L. R. B. Schomaker, and M. A. Wiering, "Recognition of handwritten

characters using local gradient feature descriptors", *Engineering Applications of Artificial Intelligence*, Vol. 45, pp. 405-414, 2015.
[10] A. T. Sahlol, M. A. Elaziz, M. A. A. Qaness, and S. Kim, "Handwritten Arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set", *IEEE Access*, Vol. 8, pp. 23011-23021, 2020.
[11] A. A. Chandio, M. Leghari, M. A. Memon, M. Leghari, and A. H. Jalbani, "A database for Urdu text detection and recognition in natural scene images", *Mehran University Research Journal of Engineering and Technology*, Vol. 39, No.1, pp. 47-54, 2020.
[12] Z. Liu, X. Pan, and Y. Peng, "Character Recognition Algorithm Based on Fusion Probability Model and Deep Learning", *The Computer Journal*, 2020.
[13] A. A. Chandio, M. Asikuzzaman, and M. R. Pickering, "Cursive Character Recognition in Natural Scene Images Using a Multilevel Convolutional Neural Network Fusion", *IEEE Access*, Vol. 8, pp. 109054-109070, 2020.
[14] Z. Akhtar, J. W. Lee, M. A. Khan, M. Sharif, S. A. Khan, and N. Riaz, "Optical character recognition (OCR) using partial least square (PLS) based feature reduction: An application to artificial intelligence for biometric identification", *Journal of Enterprise Information Management*, 2020.
[15] A. K. Sampath, and N. Gomathi, "Handwritten optical character recognition by hybrid neural network training algorithm", *The Imaging Science Journal*, Vol. 67, No. 7, pp. 359-373, 2019.
[16] A. Madakannu and A. Selvaraj, "DIGI-Net: a deep convolutional neural network for multi-format digit recognition", *Neural Computing and Applications*, Vol. 32, No. 15, pp. 11373-11383, 2020.
[17] V. C. Hallur and R. S. Hegadi, "Handwritten Kannada numerals recognition using deep learning convolution neural network (DCNN) classifier", *CSI Transactions on ICT*, 2020.
[18] A. K. Sampath and N. Gomathi, "Fuzzy-based multi-kernel spherical support vector machine for effective handwritten character recognition", *Sādhanā*, Vol. 42, No. 9, pp. 1513-1525, 2017.
[19] R. Ahmed, M. Gogate, A. Tahir, K. Dashtipour, B. A. Tamimi, A. Hawalah, M. A. E. Affendi, and A. Hussain, "Deep neural network-based contextual recognition of arabic handwritten scripts", *Entropy*, Vol. 23, No. 3, p. 340, 2021.
[20] R. S. Alkhawaldeh, "Arabic (Indian) digit handwritten recognition using recurrent transfer deep architecture", *Soft Computing*, Vol. 25, No. 4, pp. 3131-3141, 2021.
[21] T. E. D. Campos, B. R. Babu, and M. Varma, "Character recognition in natural images", In: *Proc. of the International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, 2009.
[22] A. E. Sawy, M. Loey, and H. E. Bakry, "Arabic handwritten characters recognition using convolutional neural network", *WSEAS Transactions on Computer Research*, Vol. 5, pp. 11-19, 2017.
[23] S. Malik, A. Sajid, A. Ahmad, A. Almogren, B. Hayat, M. Awais, and K. H. Kim, "An Efficient Skewed Line Segmentation Technique for Cursive Script OCR", *Scientific Programming*, 2020.
[24] R. B. Hegde, K. Prasad, H. Hebbar, and B. M. K. Singh, "Feature extraction using traditional image processing and convolutional neural network methods to classify white blood cells: a study", *Australasian Physical & Engineering Sciences in Medicine*, Vol. 42, No. 2, pp. 627-638, 2019.
[25] F. Zhou, Y. Ma, B. Wang, and G. Lin, "Dual-channel convolutional neural network for power edge image recognition", *Journal of Cloud Computing*, Vol. 10, No. 1, pp. 1-9, 2021.
[26] D. Huang, X. Cai, and C. D. Wang, "Unsupervised feature selection with multi-subspace randomization and collaboration", *Knowledge-Based Systems*, Vol. 182, p. 104856, 2019.
[27] S. Choi and Z. Jiang, "Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique", *Computers in Biology and Medicine*, Vol. 40, No. 1, pp. 8-20, 2010.