IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2022, accepted May 10, 2022, date of publication May 16, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175197

CArDIS: A Swedish Historical Handwritten Character and Word Dataset

AMIR YAVARIABDI[1], HUSEYIN KUSETOGULLARI[2], (Member, IEEE), TURGAY CELIK[3,4,5], SHIVANI THUMMANAPALLY[2], SAKIB RIJWAN[2], AND JOHAN HALL[6]

[1]Department of Mechatronics Engineering, KTO Karatay University, 42020 Konya, Turkey
[2]Department of Computer Science, Blekinge Institute of Technology, 37141 Karlskrona, Sweden
[3]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa
[4]Wits Institute of Data Science, University of the Witwatersrand, Johannesburg 2000, South Africa
[5]Faculty of Engineering and Science, University of Agder, 4630 Kristiansand, Norway
[6]Arkiv Digital, 54873 Stockholm, Sweden

Corresponding author: Huseyin Kusetogullari (huseyinkusetogullari@gmail.com)

This work was supported by the Research Project "Scalable Resource Efficient Systems for Big Data Analytics" through by the Knowledge Foundation, Sweden, under Grant 20140032.

ABSTRACT This paper introduces a new publicly available image-based Swedish historical handwritten character and word dataset named **C**haracter **Ar**kiv **D**igital **S**weden (CArDIS) (https://cardisdataset.github.io/CARDIS/). The samples in CArDIS are collected from 64, 084 Swedish historical documents written by several anonymous priests between 1800 and 1900. The dataset contains 116, 000 Swedish alphabet images in RGB color space with 29 classes, whereas the word dataset contains 30, 000 image samples of ten popular Swedish names as well as 1, 000 region names in Sweden. To examine the performance of different machine learning classifiers on CArDIS dataset, three different experiments are conducted. In the first experiment, classifiers such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbor (k-NN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Random Forest (RF) are trained on existing character datasets which are Extended Modified National Institute of Standards and Technology (EMNIST), IAM and CVL and tested on CArDIS dataset. In the second and third experiments, the same classifiers as well as two pre-trained VGG-16 and VGG-19 classifiers are trained and tested on CArDIS character and word datasets. The experiments show that the machine learning methods trained on existing handwritten character datasets struggle to recognize characters efficiently on the CArDIS dataset, proving that characters in the CArDIS contain unique features and characteristics. Moreover, in the last two experiments, the deep learning-based classifiers provide the best recognition rates.

INDEX TERMS Character and word recognition, machine learning methods, optical character recognition (OCR), old handwritten style, Swedish handwritten character dataset, Swedish handwritten word dataset.

I. INTRODUCTION

Digitization of historical handwritten documents has become essential to preserve valuable source and cultural heritage information such as birth, death, marriage, military, municipal census, and court records as well as property registers. Generally, a manual transcription approach is used to extract information and access the content of these documents which is a tedious task and requires a lot of time and effort. Therefore, it is crucial to develop efficient automatic transcription algorithms based on Optical Character Recognition (OCR) systems. However, efficient and reliable automatic transcription of ancient or historical handwritten document images is challenging for OCR systems due to the lack of data availability and the existence of high complexities and degradation presence in these documents. It is worth noting that the complexities and degradation are due to the age of documents, bleed-trough, existence of variety of noise types, stains, intensity variations, fading, merged, overlapped and broken characters, and different handwriting styles. To overcome some of those problems faced by OCR systems, multiple image-based handwritten datasets, which are collected from historical and modern document images written in different languages such as English, Spanish, Chinese, Arabic, Persian, Russian and many others, have been introduced [1]–[9].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

The existing handwritten character and word recognition datasets have several limitations: 1) the lack of availability of samples in Latin and Swedish languages; 2) the samples are generally from relatively recent document images with mild degradations; and 3) the samples are primarily written with ballpoint pens and in modern handwriting styles. Therefore, to alleviate these limitations, we introduce a new publicly available image-based historical handwritten character and word dataset, named CArDIS. The samples in CArDIS are from 64, 084 Swedish historical handwritten birth record documents (e.g. Fig. 1) written by several anonymous priests with various inks, nibs, and dip pens. The CArDIS consists of character and word datasets comprising 116, 000 single-character images in Latin and Swedish alphabets with 29 classes and 30, 000 Swedish names with 10 classes as well as 1, 000 region names. The experiments demonstrate that machine learning methods trained on existing character datasets and tested on the CArDIS character dataset, provide low recognition accuracy. Thus, it is necessary to create a new handwritten character and word dataset for historical handwritten text recognition. As a summary, the main contributions of this work are as follows:

- Introducing a new handwritten historical character and word image dataset named Character Arkiv Digital Sweden (CArDIS) (publicly available from: (https://cardisdataset.github.io/CARDIS/).
- The CArDIS is the first publicly available Swedish handwritten character and word image dataset.
- The CArDIS consists of 116, 000 letters with 29 classes, 30, 000 Swedish female and male names with 10 classes and 1, 000 region names.
- An extensive analysis of machine learning methods on created dataset and existing Handwritten Character Recognition (HCR) datasets is carried out.
- Examining the similarities and differences between created CArDIS character dataset and existing character datasets which are Extended Modified National Institute of Standards and Technology (EMNIST), IAM and CVL.

II. RELATED WORK

OCR is one of the leading research topics in pattern recognition, and it has been widely used to recognize handwritten or machine-printed characters in document images collected from heterogeneous sources [10]. Generally, the existing OCR systems include four main steps comprising pre-processing, segmentation, feature extraction, and recognition [11]. The first step aims to eliminate undesired artifacts or characteristics in a document image using binarization, skew correction, and denoising techniques. The second step aims at isolating text elements in a document image from the image's background. Usually, this process starts with line segmentation, followed by word segmentation, and, if necessary, ends with character segmentation. In the third step, various features from character/word images are extracted. In the last step, the extracted features are fed into a classifier to identify characters/words. Here, instead of reviewing the vast body of literature on OCR, we discuss commonly used machine learning methods in alphabetic character and word recognition. A comprehensive survey of OCR methods can be found in [12]–[17].

A. ALPHABETIC CHARACTER RECOGNITION

In the optical alphabetic character recognition, statistical classifiers such as Hidden Markov Model (HMM), Decision Trees (DT), k-Nearest Neighbor (kNN) have been widely used. For instance, [18] propose an HMM-based alphabetic character recognition system for Greek Polytonic on historical documents. Firstly, in this approach, geometric and Principal Component Analysis (PCA) features are extracted from character images. Next, a Gaussian Mixture Model (GMM) is used to model the feature vector. Lastly, alphabetic character images are classified using HMM through the probability calculated by the GMM. This method obtains a character error rate of 8.61%. [19] develop a method for recognizing Telugu handwritten alphabet characters. In this approach, characters are segmented from palm leaves, and then a DT algorithm is used to recognize the character images with an overall accuracy of 93.10%. Amongst all the statistical classifiers, kNN is one of the most used machine learning approaches in OCR systems [20], [21]. For example, [22] proposes a kNN-based machine learning method for Lanna Dharma handwritten alphabet character recognition on palm leaves manuscript images. To achieve this, firstly, two different wavelet transforms, and region properties are used to extract 3 different features from input alphabet character images. Then, the kNN classifier is adopted for recognition and achieves an accuracy of 95.48%. Many other kNN-based OCR methods have also been proposed [23], [24].

In OCR, another recognition approach is the use of Support Vector Machine (SVM) technique. For instance, [25] introduce an SVM-based OCR system for English handwritten character recognition. The proposed model starts with using a thinning pre-processing algorithm to produce unique skeletons representing the original handwriting characters. Then, to extract features, Freeman chain code is used. Finally, the English character images are classified using an SVM with a radial basis kernel function. The machine learning methods are trained and tested on NIST dataset [1] and the average classification accuracy of 86% is achieved. [26] develop a handwritten character recognition system for Gurmukhi alphabets. Firstly, the method extracts horizontal and vertical projection features. Then, the feature vector is fed into SVM classifier with linear and polynomial kernel functions. This approach obtains an average accuracy of 97.4%. Moreover, many other SVM-based OCR systems have been proposed to recognize alphabetic characters in different languages [27], [28].

Artificial Neural Networks (ANNs) is another widely used classifier in OCR. For example, [29] presents a character recognition method for broken Kannada characters using ANN. This method consists of three steps. Firstly, an end

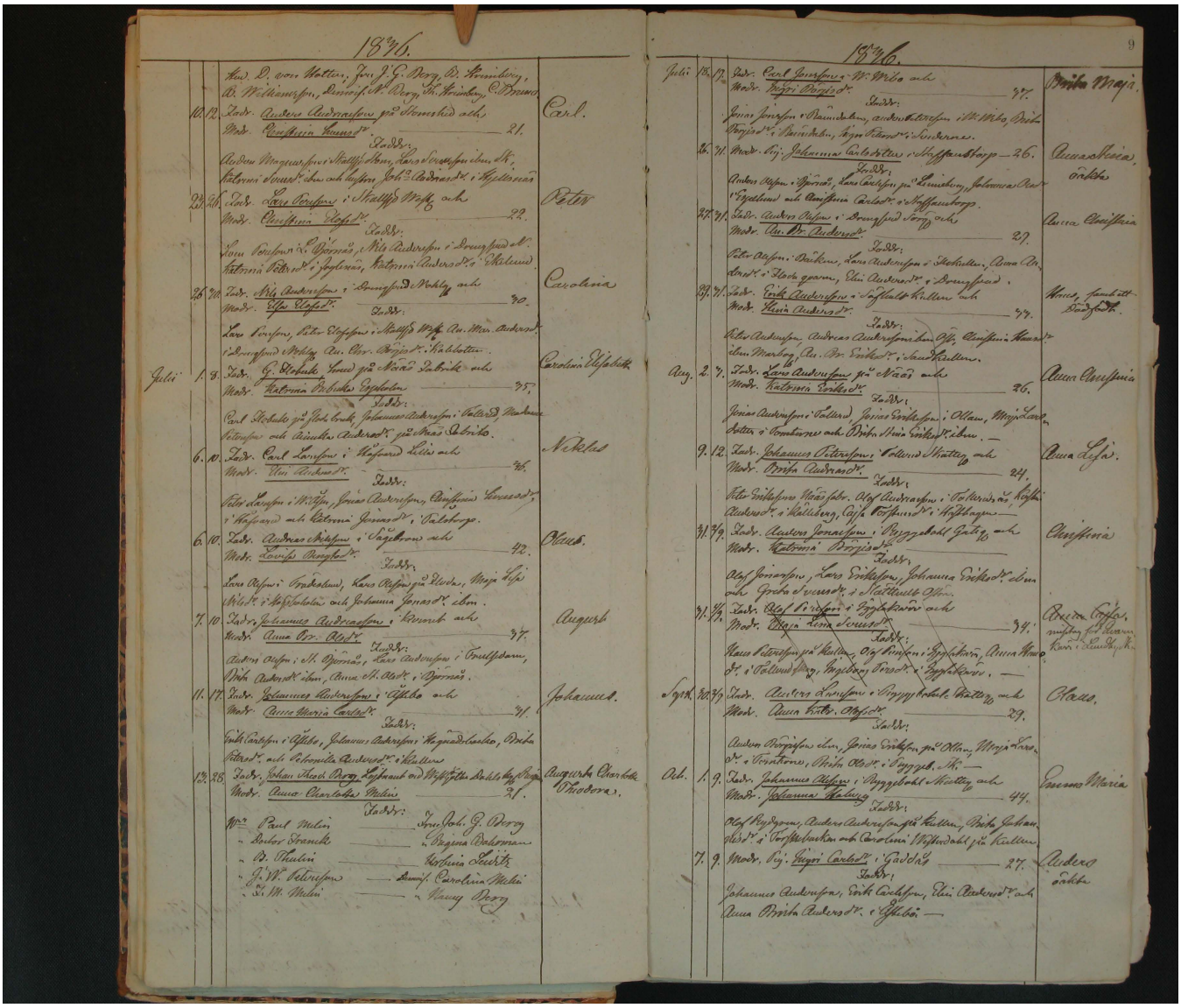

FIGURE 1. An illustration of historical document recorded in 1836.

point technique is applied to reconstruct broken characters. Secondly, zonal features are extracted from character images. Finally, ANN is used to recognize character images and obtains recognition accuracy of 98.9%. [30] introduce a handwritten character recognition system using ANN to classify English handwritten letters. This method has three phases. Firstly, the English handwritten character images are converted into binary images. Secondly, the binary image is segmented into individual characters, and then each character image is resized into 30 × 20 pixels. Finally, the character images are classified and recognized using an ANN classifier with an overall accuracy of 94.15%. [31] propose an ANN-based OCR system for recognition of handwritten characters of the English language. In this approach, binary characters images are recognized using a multi-layered ANN classifier and deliver average classification accuracy of 85.62%.

In the last decade, Convolutional Neural Network (CNN) has achieved outstanding performance in character recognition. For instance, [32] propose a two-stage CNN method to detect and classify tight Chinese characters in historical documents. To achieve this, two simultaneously CNNs are used. The first CNN aims to localize characters with bounding boxes, whereas the second CNN, based on VGG-16, aims to recognize characters in each bounding box. [33] design a CNN-based OCR system for handwritten Arabic character recognition. The CNN consists of three convolutional layers and a fully connected layer. The CNN model is tested on two different Arabic handwritten character datasets and

achieves average recognition accuracy of 94.7% and 94.8%, respectively. [34] develops a model focused on integrating CNN and SVM for Arabic handwriting recognition. A CNN is used for extracting image features in this approach, and SVM is utilized as a recognizer. The CNN architecture is tested on IFN/ENIT [35] database and achieves an error rate of 7.05%. [36] propose a CNN-based OCR method to recognize Arabic handwriting characters automatically. The CNN architecture consists of three convolutional and two fully connected layers. This method achieves an accuracy of 97%. Many other deep-learning-based OCR frameworks have been designed to achieve a high accuracy rate for different handwritten character datasets [37]–[39].

B. WORD RECOGNITION

Generally speaking, in word recognition, two main types of strategies have been applied: 1) analytical approach and 2) holistic approach. A word must first be segmented into units such as letters, graphemes, strokes, or pseudo-letters in the former one. Then, the word units are recognized using a machine learning algorithm. Finally, the likelihood for each word in the lexicon can be estimated to recognize the word. For instance, [40] propose a hybrid handwritten word recognition approach based on ANN and HMM. First, a slicing technique is used to build a graph which shows all possibilities to segment a word into letters. Then, ANN is utilized to compute probabilities for each letter in the graph. Finally, HMM is used to classify the words. For evaluation, a French word database has been used, namely IRONOFF [41] handwriting database, and the system achieves an accuracy of 99.1%. [42] develop a CNN-based method to recognize words in RIMES dataset. A CNN architecture is used first to measure and then re-sample an input word image to a canonical representation in this approach. Then, a fully connected CNN architecture is designed to predict the characters. Finally, the words are recognized by a vocabulary-matching method. Another CNN-based method is proposed in [43]. In [43] an attention-based sequence-to-sequence model is used for handwritten word recognition in IAM dataset. To form encoder stage, the ResNet feature extraction is combined with bidirectional LSTM. Then, to predict words, a decoder is integrated with a content-based attention mechanism. [44] proposes another attention-based encoder-decoder model to recognize handwritten text using sequences of characters, extracted from IAM dataset. In another work [45], an attention-based method combines CNN Recurrent Neural Networks (RNN) encoder with an RNN-decoder to recognize line or word. In this method, the encoder extracts features from the handwritten texts and sequentially encodes temporal context. Then, the decoder recognizes a character one by one, using an attention mechanism. In [46], a Generative Adversarial Network (GAN) architecture is proposed to recognize handwritten words qt character levels using IAM dataset. The method consists of two main steps. The first step is a discriminator which consists of a path signature features extractor and a CNN-LSTM binary classifier to distinguish realistic and forgery handwritten data. The second step is a generator used to produce random handwritten characters.

In the holistic approach, word recognition is performed on the whole representation of words, without segmenting them into units (e.g. letters). In this manner, [47] propose a holistic-based lexicon reduction method to recognize 200 region names written in Farsi/Arabic language. First, the words holistic features such as single, double, and triple dots, their order from left to right and their up or down position in a word are extracted. Then, the extracted features are fed into an HMM model to recognize words. [48] present a handwritten Arabic word recognition system. The system uses Pseudo Zernike Moments as a feature extraction technique. Then, the HMM is used as a classifier. The OCR framework is tested on 100 Arabic names and provides 88% of the word recognition accuracy. [49] propose a word recognition method based on HMM. In this method, three feature sets based on black–and–white transition, image gradient, and contour chain code are employed. Then, each of those is modeled with an individual HMM. In the recognition step, the outputs of the HMMs are combined using a multi-layer perceptron. The method is tested on the Iranshahr 3 dataset and provides 89% recognition accuracy. [50] propose a word recognition system based on a 12-layer CNN and canonical correlation. The method is tested on 3 different datasets such as IAM [51], RIMES [52] and IFN/ENIT [35] and achieves error rate of 6.45%, 3.9%, and 3.24%, respectively.

Different handwritten character and word datasets with different languages have been created from handwritten document images. The generated datasets are used for developing machine learning based OCR models. Extensively used and the generated CArDIS handwritten character and word datasets are tabulated in Table 1 and explained below.

The EMNIST letters dataset [1] consists of 145, 600 isolated handwritten letter images with 26 balanced classes. The letters in the dataset was collected from handwritten documents written in English language by 3, 600 writers. This dataset contains gray-scale letter images which are size-normalized and denoised. It is publicly available to the research community. The QUWI dataset contains a handwritten dataset in Arabic and English languages written by 1, 017 volunteers of different ages, nationalities, genders, and education levels [53]. This dataset is mostly used for developing writer identification systems. The dataset has 4, 068 document images and consists of 60, 000 words in Arabic and 100, 000 words in English languages. The dataset is available upon request. The IAM dataset contains 5, 685, 13, 353 and 115, 320 isolated and annotated handwritten sentences, text lines, and handwritten word images, respectively [51]. In this dataset, the words were automatically collected from the handwritten document images using hidden Markov model (HMM) based automatic segmentation model [51] and were verified manually. All the images in this dataset are scanned with a resolution of 300 DPI and stored in greyscale color space. This dataset was collected

TABLE 1. Handwritten character and word datasets in different languages.

Dataset Name	Language	Color	Letter Samples	Word Samples	Availability	Historical/Modern
EMNIST [1]	English	Gray-scale	145.600	-	Yes	Modern
IAM [51]	English	Gray-scale	-	115.320	Yes	Modern
QUWI [53]	Arabic, English	RGB	-	160.000	No	Modern
CVL [54]	English, German	RGB	-	101,069	Yes	Modern
CEDAR [55]	English	Gray-scale, Binary	21,308	10,570	No	Modern
IRONOFF [41]	French	Gray-scale	32,000	50,000	No	Modern
IFN/ENIT [35]	Arabic	Binary	212,211	26,459	Yes	Modern
KHATT [56]	Arabic	Binary	-	-	Yes	Modern
Chars74K [58]	English, Hindu	RGB	74,000	-	Yes	Modern
uTHCD [57]	Tamil	Binary	90,000	-	No	Modern
GRPOLY-DB [59]	Greek	Gray-scale	102,596	171,511	Yes	Historical
CArDIS	**Swedish, English**	**RGB**	**116,000**	**30,000**	**Yes**	**Historical**

from 1, 539 scanned handwritten documents written in the English language by 657 writers. The IAM dataset is publicly available [51]. The CVL is a dataset that contains only handwritten words in the English and German languages [54]. This dataset was collected from different document images from 310 different writers. It consists of 101, 069 words with 292 different word classes. Handwritten word samples are stored in RGB color space with their labels [54]. This dataset is publicly available and it is used for writer identification and word spotting. The CEDAR dataset [55] comprises of 10, 570 handwritten words in gray-scale. This dataset contains 12, 821 isolated handwritten uppercase letters and 8, 487 isolated lowercase letters in binary. The database is imbalanced, which was collected manually from the scanned USA mails. This dataset is not publicly available. The IRONOFF online/offline handwritten dataset contains isolated French characters, digits, and cursive words [41]. This dataset was extracted from approximately 1, 000 digitized forms written by French writers.

III. EXISTING HANDWRITING DATASETS

A. HANDWRITTEN CHARACTER AND WORD DATASETS

In addition to the Latin handwritten character and word datasets, other handwritten character and word datasets have been generated in different languages. Several of these datasets are explained and described below.

The IFN/ENIT [35] is an Arabic handwritten character dataset that consists of 212, 211 handwritten Arabic characters and 26, 459 words. It was created to develop Arabic OCR systems. This dataset was created from 2, 200 binary handwritten document forms written by 411 different writers, including names of towns and villages in Tunisia. The dataset is publicly available. The KHATT dataset [56] is another Arabic handwritten dataset that contains 1, 000 handwritten forms written by 1, 000 different writers. All the handwritten forms are scanned at 200, 300, and 600 dpi resolution, and they are pre-processed using OTSU's method to convert the handwritten images into binary images. The dataset is publicly available. The Chars74K dataset [58] composes of 7, 705 handwritten, 3, 410 hand-drawn, and 62, 992 synthesized characters which were collected from different natural images of street scenes in Bangalore, India. This dataset has 74, 000 handwritten characters with 64 classes, and they were written in Latin, Hindu, and Arabic languages. This dataset is publicly available. The GRPOLY-DB [59] is a character and word dataset collected from printed and handwritten polytonic Greek document images. The scanned printed and handwritten documents were written between 1838 and 1912. This dataset was extracted from 399 document images and consists of 102, 596 words and 171, 511 characters. This dataset is publicly available.

In addition to the datasets as mentioned above, there are other character and word datasets that are generated in other languages such as Urdu [8], Chinese [60], Persian [5] etc. A comprehensive review of the OCR systems and datasets is discussed in [12].

B. HANDWRITTEN DOCUMENT IMAGES

In recent years, various handwriting document image databases in Latin have been introduced to solve different problems in document analysis applications. For instance, George Washington [61], [62] database is one of the well-known databases which contains 20 different historical handwritten document images. The documents were written in English in the eighteenth century. The handwritten document images are labelled with 4894-word instances, 1471 different word classes, 82 letters and 656 text lines. Another database is Esposalles database [2], [63] which includes 173 Spanish handwriting document images. The Spanish documents were written between fifteen and twentieth centuries. In the handwriting document images, the text blocks and lines, as well as the transcriptions are annotated. Germana database [64] is another Spanish database, and the handwritten documents are from 1891. The Germana contains 764-page annotated Spanish document images. In addition to Esposalles and Germana, the Rodrigo database [65] was created from an older document named ''Historia de Espana∼del arçobispo Don Rodrigo'' which is from the sixteen centuries. The database has nearly 20.000 annotated text lines and 231.000 annotated words. The Parzival database [66] has 47 handwritten document images in Medieval German language in the

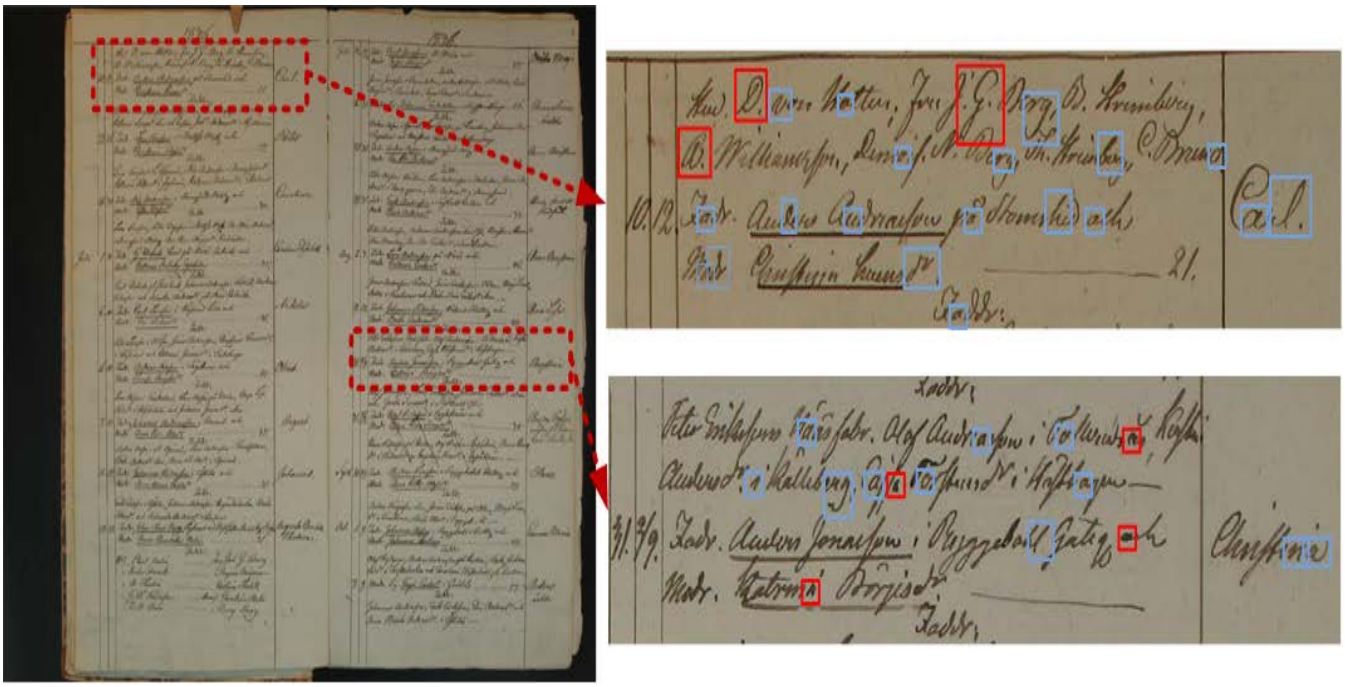

FIGURE 2. A birth record historical document image with examples of annotated and non-annotated characters with blue and red boxes, respectively.

thirteen centuries. The document images in the databases are labelled with 4, 477 text lines, 23,478-word instances, 4, 934-word classes and 93 letters. Saint Gall database [67] consists of 60 handwritten document images and 1, 410 text lines and 11 597 words annotations. The manuscripts were written in Latin in the ninth century. In [68], the largest Swedish database is introduced which contains 15, 000 high-resolution handwritten document images. The Swedish documents were written in nineteenth centuries.

IV. CARDIS DATA COLLECTION

The CArDIS dataset consists of sample Swedish historical handwritten character and word images collected from 64, 084 Swedish birth record handwritten document images acquired by Arkiv Digital. In the handwritten document images, each Swedish birth record contains a newly born child's name, born date, baptized date, born place, father's name, and mother's name. Various anonymous priests recorded the handwritten documents between 1800 and 1900 in Swedish churches located in different counties such as Gotland, Gävleborg, Norrbotten, Västerbotten, Västernorrland, Västmanland, Älvsborg, and Örebro. The scanned document images are with the resolution of 6000 × 4000 in RGB color space, including various complexities such as handwriting styles, background color, and variety of degradations. The collections of CArDIS dataset (publicly available from: ((https://cardisdataset.github.io/CARDIS/) are clearly explained below.

CArDIS Dataset I is generated from 64, 084 historical Swedish birth record handwritten document images and contains only isolated lowercase handwritten Latin letters from 'a' to 'z' as well as special Swedish letters (e.g., å, ä, ö). Each letter is manually segmented and cropped from handwritten document images as illustrated in Fig. 2. Note that only lowercase letters, as well as the characters which can be read and perceived, are selected (e.g. blue boxes in Fig. 2). In contrast, the uppercase letters and degraded lowercase letters are ignored as depicted in red boxes in Fig. 2. To the best of our knowledge, the CArDIS dataset is the first historical handwritten Swedish lowercase letter that provides image samples in RGB color space with original sizes as shown in Fig. 3. Moreover, in this dataset, the lowercase letter images may consist of extra parts from neighboring characters and include various artifacts such as degradation, noise, line dashes, and underlines. This dataset contains 116, 000 lowercase letter images with 29 classes, where 26 classes ('a'–'z') belong to Latin alphabets and 3 classes (å, ä, ö) belong to Swedish alphabets (see Fig. 4), with 4,000 images per class. This dataset is generated to further improve lowercase letter recognition and segmentation for OCR systems in historical document images with different degradation and complex backgrounds. This dataset will be publicly available.

CArDIS Dataset II is a Swedish word dataset that is manually obtained from 64, 084 birth record handwritten document images. To generate this dataset, the ten most popular Swedish female and male names as well as Swedish region names are collected from these birth record documents. The female names are Anna, Brita, Maria, Johanna, Christina, whereas, the male names are Anders, Olof, Lars, Carl, and Pehr. Moreover, there are various Swedish region names in

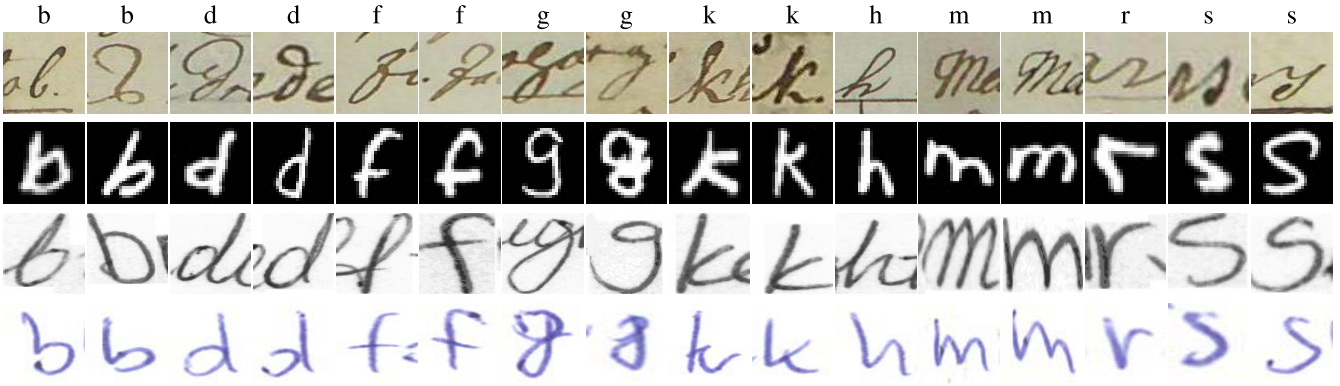

FIGURE 3. Sample images of the Latin alphabetic characters from CArDIS (first row), EMNIST (second row), IAM (third row), and CVL (fourth row) datasets.

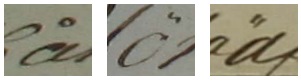

FIGURE 4. Illustration of special Swedish characters.

the CArDIS dataset II. This dataset includes 30, 000 Swedish female and male names' images with 10 classes, and each class contains 3, 000 images. In addition, the Swedish region names includes 1, 000 images. Fig. 5 shows several female, male, and region names' images in the CArDIS Dataset II. The Swedish male, female, and region names are manually segmented and cropped from handwritten document images and stored in original sizes as well as RGB color space. Moreover, the collected images may contain artifacts such as noise, dash lines, underline, bleed-through, faint, and many others, as shown in Fig. 5. This dataset can be used in name indexing and word recognition applications and will be publicly available.

A. SWEDISH CHARACTER AND WORD DATA CHARACTERISTICS

The CArDIS dataset is generated based on the Swedish historical document records written by different priests in the 19th century. Thus the dataset has multiple unique characteristics, as explained below.

- Degradation: The low quality of the used ink and papers in the 19^{th} century, age of documents, and distortions affect the characteristics of the words and letters in the CArDIS dataset. These issues result in multiple degradation and artifacts. For instance, the age of documents causes deterioration of texts and characters (i.e. faint). Moreover, other artifacts in the document images are background variation, show-through, bleed-through, and smear. Consequently, the CArDIS dataset is exhibited with many different inter- and intra-class variations.
- Handwriting styles: History indicates that each person has its own unique writing style [12]. In the birth record documents, the texts were written in Gothic, cursive, and copperplate styles by various priests using different inks, nibs, and dip pens, which result in distinct appearances. For instance, applying different pressures on a nip can result in flowing different amounts of ink, generating different character appearances. Moreover, in the documents, the same word and character were written in many different sizes. Thus, the shapes can be diverse. Hence, in the CArDIS dataset, the words and characters are scripted in various writing styles, sizes, directions, widths, and arrangements. These variations in handwriting patterns due to individual writing styles and materials used to write the texts generate endless inter-class variations.
- Special characters: The birth record documents were written in the Swedish language. Thus the documents do not follow the standard Latin alphabets. Although the overall writing of the documents are quite similar to the Latin, 3 extra letters such as å, ä, ö are included (see Fig. 4).

The characteristics mentioned above generate many distortions in the appearance of words and characters and lead to a unique dataset where the words and characters appear with many inter- and intra-writing variations. Thus, the CArDIS dataset overcomes multiple limitations over the existing datasets. For instance, most datasets such as EMNIST are based on characters in Latin language and written in modern handwriting styles with ballpoint and rollerball pens. Besides this, they are collected from non-degraded documents, and they are size normalized. These characteristics of the existing datasets restrict the application of existing methods for handwritten historical character and word recognition where the variability and complexity become more dominant. Therefore, to support the research in the Latin and Swedish handwritten character and word recognition, a new dataset based on historical handwritten documents is generated to resolve the problem of the existing ones.

FIGURE 5. Illustration of Swedish female and male names, as well as region names.

V. RESULTS AND DISCUSSION

A. LEARNING ALGORITHMS AND HYPERPARAMETERS

For quantitative evaluations, various learning classifiers have been used. In this work, k-Nearest Neighbour (k-NN), random forest, one-versus-all SVM classifier with RBF kernel, recurrent neural network (RNN), convolutional neural networks (CNNs) and two different pre-trained deep learning methods have been selected to recognize handwritten characters and words. In the k-NN classifier, the distance is first calculated using the Euclidian distance and then handwritten characters and words are identified by the majority class of k-nearest neighbors. It is important to note that, the raw pixel values of image samples are used in the k-NN classifier, and the k value is empirically selected as 5 for classification of handwritten characters and words.

Random Forest is another classifier used to evaluate the quantitative results. In this classifier, the raw pixels of image samples are first normalized between 0 and 1. After that, the random forest classifier is applied to the normalized pixel values. The classifier consists of two different parameters which are: (1) the number L of trees and, (2) the number K of random features preselected in the splitting process. In the Random Forest classifier, we set the parameters as $L = 100$ and $k = 12$. The complete assessment regarding to these parameters is analyzed and discussed in [16].

The third classifier is RBF kernel SVM. In order to get the results, two different input types are used for the SVM classifier which are the raw pixels of image samples and the features of image samples extracted by using histogram of oriented gradients (HOGs). As a result, these create two experimental structures named SVM and SVM-HOG in the rest of the paper. In the SVM classifier, we set two parameter values as $\gamma = 0.001$ and $C = 1$.

Another handwritten character and word classifier is developed based on Recurrent Neural Network (RNN). In RNN classifier, four-layer neural network model is designed and used to obtain the results. Firstly, the pixel values of the image sample are normalized and then normalized pixel values are used as inputs for the RNN classifier. The batch size and iteration size are selected as 64 and 10, respectively. Moreover, Rectifier Linear Unit (ReLU) is employed as an activation function in the hidden layers and in the output layer, Softmax function is used to estimate probabilities of output classes of handwritten characters and words. Artificial Neural Network (ANN) based classifier includes 5 hidden layers and output layer. This classifier is employed with the same strategy of RNN classifier.

The CNN-based handwritten character and word classifier consists of following layers; 1) Input layer, 2) three convolutional layers, 3) three fully connected layers, and 4) one output layer. The first two convolutional layer use 64 filters with the filter or kernel size of 5×5, and the last convolutional layer uses 128 filters or kernels with the same kernel size. The convolutional layers are followed by fully connected layers which each one contains 128 nodes. In addition, the ReLU is employed as an activation function in all connected layers except final layer. Softmax is used in the final layer to estimate the probabilities of output classes of handwritten characters and words. In the CNN model, the batch size and iteration size is set to 200 and 10, respectively. In VGG-16 and VGG-19, the number of neurons in the last fully connected layer has been changed to the number of classes. In both methods, the learning rate, epoch and batch size are set to 0.001, 200, and 32, respectively.

B. EXPERIMENTAL SETUP

Since the EMNIST dataset is the only available Latin character dataset, we manually collected character samples from IAM and CVL datasets to demonstrate characters in the CArDIS dataset effectively. Each of these collected datasets consists of 104,000 character images with 26 classes. Moreover, 80% of the character datasets are randomly selected and used for training, and the rest is used for testing. To obtain the results, six different classifiers, which are RNN, k-NN ($k = 5$), RF ($L = 100$ and $k = 12$), SVM with RBF kernel function, ANN with 5 hidden layers, and CNN with 3 convolutional layers and 3 fully connected layers are used. In addition to these algorithms, Histogram Oriented Gradient (HOG) feature extraction technique is used with two conventional classifiers which are SVM and ANN. In addition to the eight

TABLE 2. Handwritten character recognition accuracy rates using different classifiers trained on EMNIST, IAM and CVL, and tested on CArDIS.

Method	EMNIST	IAM	CVL
RNN	23.45	36.65	27.47
kNN	14.56	28.38	17.89
RF	12.25	23.76	16.34
SVM	28.65	43.89	32.51
SVM-HOG	31.76	46.32	40.67
ANN	26.54	39.89	31.23
ANN-HOG	30.12	41.48	37.78
CNN	34.16	53.18	45.79

aforementioned methods, two different deep learning-based pre-trained models which are VGG-16 and VGG-19 have been applied to further analyze the characteristics of CArDIS character and word dataset. The classifiers are trained and tested on a computer with a single NVIDIA GTX 1060TI GPU, an Intel i7-7700HQ CPU, and 16 GB RAM.

C. COMPARING CLASSIFIERS ON VARIOUS CHARACTER DATASETS

The first experiment focuses on understanding and analyzing performance of different classifiers trained on existing Latin character datasets (i.e. EMNIST, IAM, and CVL) and tested on the Latin characters in the CArDIS. Besides of this, diversities and similarities between different character datasets are evaluated. The obtained accuracy rates are shown in Table 2. The results prove that the classifiers trained on the existing datasets cannot perform well when the character samples in the CArDIS are used for testing. Based on the results, the lowest recognition accuracy rate is obtained using RF classifier with EMNIST, which is 12.25%, whereas the best accuracy rate is achieved using CNN classifier with IAM, which is 53.18%. The poor performance of the classifiers trained with the existing datasets shows that the CArDIS has different characteristics due to several diversities such as (1) the characters are written by different priests with the 19^{th} century Swedish handwritten styles using various types of dip-pens; (2) the handwritten characters are with various orientations and thickness, and (3) variations in handwritten character patterns and appearances as they are collected from old Swedish handwritten document images. Furthermore, the quantitative results show that the HOG- and CNN-based classifiers outperform as compared with the classifiers trained on the normalized pixels. According to the results, all the classifiers perform poorly, indicating that handwritten characters in the CArDIS contain different features than the ones in the existing datasets.

D. COMPARING CLASSIFIERS ON CARDIS CHARACTER DATASET

In contrast to the experiment one, experiment two contains Swedish characters which includes 29 classes. To conduct this experiment, 87, 000 handwritten samples in the CArDIS are used to train the classifiers and 29, 000 samples are used to evaluate the performance of them. Fig. 6 depicts the accuracy rates of ten classifiers on CArDIS dataset. Overall, results show that the classifiers give high recognition rates. The highest recognition rate is obtained using deep learning models which are VGG-16, CNN and VGG-19 with 96.12%, 95.28% and 93.92%, respectively. The fourth- and fifth-best results are achieved by SVM and ANN classifiers with HOG features with the accuracy rate of 93.82% and 93.53%, respectively. SVM and ANN on the normalized pixels gives the error rates of 8.75% and 10.32%, respectively. RNN performs slightly worse than ANN and achieves 87.83% accuracy rate. The worse recognition performances are obtained using RF, and kNN methods with error rates of 14.42% and 15.40%, respectively. As a result, the CNN classifier overcomes the challenges arised from handwritten character images in the CArDIS.

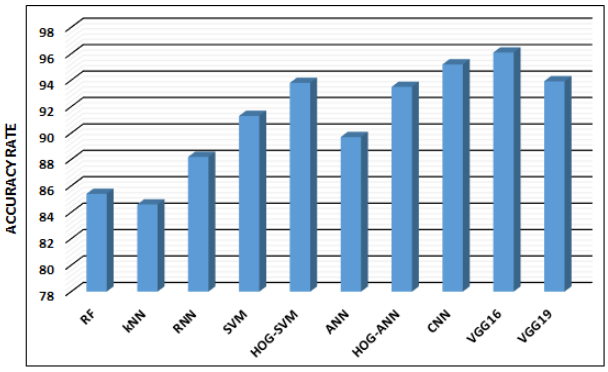

FIGURE 6. Recognition performance of classifiers on CArDIS dataset.

TABLE 3. Handwritten word recognition accuracy rates using different classifiers trained on word dataset in CArDIS.

Method	CArDIS Word Dataset
RNN	48.18
kNN	32.12
RF	35.57
SVM	57.88
SVM-HOG	68.06
ANN	42.26
ANN-HOG	65.18
CNN	72.19
VGG-16	75.23
VGG-19	76.89

E. PERFORMANCE OF CLASSIFIERS ON CARDIS WORD DATASET

The third experiment aims at understanding and evaluating the performance of the machine learning classifiers using CArDIS word dataset which contains 30, 000 Swedish names with 10 classes. To achieve the results, the word dataset is first divided into 80% training and 20% testing, thus 24, 000 image samples are used to train the classifiers and 6, 000 image samples are used to test the performance of the classifiers. Table 3 tabulates the recognition accuracy rates using ten different classifiers trained and tested on CArDIS

word dataset. According to the results, all classifiers provide accuracy rates between 32.12% and 76.89%. The best and second-best performances are obtained using VGG-19 and VGG-16 deep learning methods with 76.89% and 75.23%, respectively. In addition, the designed CNN classifier obtains 72.19% and SVM with HOG features achieves 68.06% recognition rates. ANN method with HOG features provides slightly worse than SVM with HOG features with 65.18%. SVM, RNN and ANN achieve recognition accuracy rates of 57.88%, 48.18% and 42.26%, respectively. The worst recognition performances are determined by using RF and kNN classifiers with the error rates of 64.43% and 67.88%, respectively. Moreover, the results in Table 3 indicate that the Swedish names in the CArDIS dataset are complex and difficult to recognize since the classifiers provide not very high recognition performance.

VI. CONCLUSION

In this paper, a new historical handwritten character and word dataset, named CArDIS, is introduced and publicly available for the research community ((https://cardisdataset.github.io/CARDIS/). The CArDIS is manually collected from Swedish birth record handwritten document images written in the 19^{th} century. This handwritten dataset consists of (1) alphabetic character images in Latin and Swedish languages with original appearances in RGB; and (2) 10 popular female and male Swedish names' image samples and Swedish region names' image samples with original appearances. In this paper, various classifiers have been trained on three different handwritten Latin character datasets and tested on the CArDIS character dataset. The results verify that classifiers perform poorly and give less recognition accuracy, indicating that the characters in the CArDIS have different features and characteristics than the existing Latin character datasets. Moreover, deep learning-based methods provide the best recognition performance to recognize the Swedish characters and names comparing to the other comparing methods. The CArDIS will be publicly available to improve the performance of OCR systems further.

REFERENCES

[1] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.

[2] V. Romero, A. Fornes, N. Serrano, J. A. Sanchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Llados, "The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognit.*, vol. 46, pp. 1658–1669, Jun. 2013.

[3] L. Xu, Y. Wang, X. Li, and M. Pan, "Recognition of handwritten Chinese characters based on concept learning," *IEEE Access*, vol. 7, pp. 102039–102053, 2019.

[4] H. M. Balaha, H. A. Ali, M. Saraya, and M. Badawy, "A new Arabic handwritten character recognition deep learning system (AHCR-DLS)," *Neural Comput. Appl.*, vol. 33, pp. 6325–6367, Oct. 2020.

[5] S. Mozaffari and H. Soltanizadeh, "ICDAR 2009 handwritten Farsi/Arabic character recognition competition," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, 2009, pp. 1413–1417.

[6] A. Abdallah, M. Hamada, and D. Nurseitov, "Attention-based fully gated CNN-BGRU for Russian handwritten text," *J. Imag.*, vol. 12, no. 12, pp. 1–21, 2020.

[7] N. Sandhya, R. Krishnan, and D. R. R. Babu, "A novel local enhancement technique for rebuilding broken characters in a degraded Kannada script," in *Proc. IEEE Int. Advance Comput. Conf. (IACC)*, Jun. 2015, pp. 176–179.

[8] M. N. Asim, M. U. Ghani, M. A. Ibrahim, W. Mahmood, A. Dengel, and S. Ahmed, "Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification," *Neural Comput. Appl.*, vol. 33, pp. 5437–5469, Sep. 2020.

[9] H. T. Nguyen, C. T. Nguyen, P. T. Bao, and M. Nakagawa, "A database of unconstrained Vietnamese online handwriting and recognition experiments by recurrent neural networks," *Pattern Recognit.*, vol. 78, pp. 291–306, Jun. 2018.

[10] A. T. Sahlol, M. A. Elaziz, M. A. A. Al-Qaness, and S. Kim, "Handwritten Arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set," *IEEE Access*, vol. 8, pp. 23011–23021, 2020.

[11] A. Rehman, S. Naz, and M. I. Razzak, "Writer identification using machine learning approaches: A comprehensive review," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 10889–10931, Apr. 2019.

[12] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020.

[13] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 712–724, May 2006.

[14] A. Baldominos, Y. Saez, and P. Isasi, "A survey of handwritten character recognition with MNIST and EMNIST," *Appl. Sci.*, vol. 9, pp. 3169–3180, Jan. 2019.

[15] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions," *IEEE Access*, vol. 9, pp. 72894–72936, 2021.

[16] N. Babu and A. Soumya, "Character recognition in historical handwritten documents—A survey," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2019, pp. 299–304.

[17] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of OCR applications," *Int. J. Mach. Learn. Comput.*, vol. 2, pp. 314–318, Jun. 2012.

[18] V. Katsouros, V. Papavassiliou, F. Simistira, and B. Gatos, "Recognition of Greek polytonic on historical degraded texts using HMMs," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Apr. 2016, pp. 346–351.

[19] P. N. Sastry, R. Krishnan, B. Venkata, and S. Ram, "Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach," *J. Eng. Appl. Sci.*, vol. 5, no. 3, pp. 22–32, 2010.

[20] H. Kusetogullari, A. Yavariabdi, A. Cheddad, H. Grahn, and J. Hall, "ARDIS: A Swedish historical handwritten digit dataset," *Neural Comput. Appl.*, vol. 32, no. 21, pp. 16505–16518, 2020.

[21] H. Kusetogullari, A. Yavariabdi, J. Hall, and N. Lavesson, "DIGITNET: A deep handwritten digit detection and recognition method using a new historical handwritten digit dataset," *Big Data Res.*, vol. 23, Feb. 2021, Art. no. 100182, doi: 10.1016/j.bdr.2020.100182.

[22] P. Inkeaw, C. Chueaphun, J. Chaijaruwanich, A. Klomsae, and S. Marukatat, "Lanna dharma handwritten character recognition on palm leaves manuscript based on wavelet transform," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Oct. 2015, pp. 253–258.

[23] P. Romulus, Y. Maraden, P. D. Purnamasari, and A. A. P. Ratna, "An analysis of optical character recognition implementation for ancient batak characters using K-nearest neighbors principle," in *Proc. Int. Conf. Quality Res. (QiR)*, Aug. 2015, pp. 47–50.

[24] S. Alirezaee, H. Aghaeinia, M. Ahmadi, and K. Faez, "Recognition of middle age Persian characters using a set of invariant moments," in *Proc. 33rd Appl. Imag. Pattern Recognit. Workshop*, 2004, pp. 196–201.

[25] D. Nasien, H. Haron, and S. S. Yuhaniz, "Support vector machine (SVM) for English handwritten character recognition," in *Proc. 2nd Int. Conf. Comput. Eng. Appl.*, 2010, pp. 249–252.

[26] M. K. Mahto, K. Bhatia, and R. K. Sharma, "Combined horizontal and vertical projection feature extraction technique for Gurmukhi handwritten character recognition," in *Proc. Int. Conf. Adv. Comput. Eng. Appl.*, Mar. 2015, pp. 59–65.

[27] S. RajaKumar and V. S. Bharathi, "Eighth century Tamil consonants recognition from stone inscriptions," in *Proc. Int. Conf. Recent Trends Inf. Technol.*, Apr. 2012, pp. 40–43.

[28] J. John, K. V. Pramod, K. Balakrishnan, and B. B. Chaudhuri, "A two stage approach for handwritten Malayalam character recognition," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Sep. 2014, pp. 199–204.

[29] N. Sandhya and R. Krishnan, "Broken Kannada character recognition—A neural network based approach," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 2047–2050.
[30] J. Pradeep, "Neural network based recognition system integrating feature extraction and classification for English handwritten," *Int. J. Eng.*, vol. 25, no. 2, pp. 99–106, May 2012.
[31] A. Choudhary, R. Rishi, and S. Ahlawat, "Off-line handwritten character recognition using features extracted from binarization technique," *AASRI Proc.*, vol. 4, pp. 306–312, Jan. 2013.
[32] H. Yang, L. Jin, W. Huang, Z. Yang, S. Lai, and J. Sun, "Dense and tight detection of Chinese characters in historical documents: Datasets and a recognition guided detector," *IEEE Access*, vol. 6, pp. 30174–30183, 2018.
[33] K. Younis and A. Khateeb, "Arabic hand-written character recognition based on deep convolutional neural networks," *Jordanian J. Comput. Inf. Technol.*, vol. 3, no. 3, p. 186, 2017.
[34] M. Elleuch, R. Maalej, and M. Kherallah, "A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition," *Proc. Comput. Sci.*, vol. 80, pp. 1712–1723, Jan. 2016.
[35] H. El Abed and V. Margner, "The IFN/ENIT-database—A tool to develop Arabic handwriting recognition systems," in *Proc. 9th Int. Symp. Signal Process. Appl.*, Feb. 2007, pp. 1–4.
[36] N. Altwaijry and I. Al-Turaiki, "Arabic handwriting recognition system using convolutional neural network," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2249–2261, Jun. 2020.
[37] M. Amrouch, M. Rabi, and Y. Es-Saady, "Convolutional feature learning and CNN based HMM for Arabic handwriting recognition," in *Proc. Int. Conf. Image Signal Process.*, 2018, pp. 265–274.
[38] M. Rajalakshmi, P. Saranya, and P. Shanmugavadivu, "Pattern recognition-recognition of handwritten document using convolutional neural networks," in *Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS)*, Apr. 2019, pp. 1–7.
[39] D. S. Maitra, U. Bhattacharya, and S. K. Parui, "CNN based common approach to handwritten character recognition of multiple scripts," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1021–1025.
[40] Y. Haur Tay, P. M. Lallican, M. Khalid, C. Viard-Gaudin, and S. Kneer, "An offline cursive handwritten word recognition system," in *Proc. IEEE Region 10 Int. Conf. Electr. Electron. Technol. (TENCON)*, Aug. 2001, pp. 519–524.
[41] C. Viard-Gaudin, P. M. Lallican, S. Knerr, and P. Binter, "The IRESTE on/off (IRONOFF) dual handwriting database," in *Proc. 5th Int. Conf. Document Anal. Recognit.*, 1999, pp. 455–458.
[42] R. Ptucha, F. P. Such, S. Pillai, F. Brockler, V. Singh, and P. Hutkowski, "Intelligent character recognition using fully convolutional neural networks," *Pattern Recognit.*, vol. 88, pp. 604–613, Apr. 2019.
[43] D. Kass and E. Vats, "AttentionHTR: Handwritten text recognition based on attention encoder-decoder networks," 2022, *arXiv:2201.09390*.
[44] J. Poulos and R. Valle, "Character-based handwritten text transcription with attention networks," *Neural Comput. Appl.*, vol. 33, pp. 10563–10573, Feb. 2021.
[45] J. Michael, R. Labahn, T. Gruning, and J. Zollner, "Evaluating sequence-to-sequence models for handwritten text recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1286–1293.
[46] B. Ji and T. Chen, "Generative adversarial network for handwritten text," 2019, *arXiv:1907.11845*.
[47] S. Mozaffari, K. Faez, V. Märgner, and H. El-Abed, "Lexicon reduction using dots for off-line Farsi/Arabic handwritten word recognition," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 724–734, Apr. 2008.
[48] I. El-Feghi, F. Elmahjoub, B. Alswady, and A. Baiou, "Offline handwritten Arabic words recognition using Zernike moments and hidden Markov models," in *Proc. Int. Conf. Comput. Appl. Ind. Electron.*, Dec. 2010, pp. 165–168.
[49] S. A. A. A. Arani, E. Kabir, and R. Ebrahimpour, "Handwritten Farsi word recognition using NN-based fusion of HMM classifiers with different types of features," *Int. J. Image Graph.*, vol. 19, no. 1, Jan. 2019, Art. no. 1950001.
[50] A. Poznanski and L. Wolf, "CNN-N-gram for handwriting word recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2305–2314.
[51] M. Zimmermann and H. Bunke, "Automatic segmentation of the IAM off-line database for handwritten English text," in *Proc. Int. Conf. Pattern Recognit.*, vol. 4, 2002, pp. 35–39.
[52] W. Swaileh and T. Paquet, "A syllable based model for handwriting recognition," 2018, *arXiv:1808.07277*.
[53] S. A. Maadeed, W. Ayouby, A. Hassaine, and J. M. Aljaam, "QUWI: An Arabic and English handwriting dataset for offline writer identification," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, Sep. 2012, pp. 746–751.
[54] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, "CVL-DataBase: An off-line database for writer retrieval, writer identification and word spotting," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 560–564.
[55] S. Singh and M. Hewitt, "Cursive digit and character recognition in CEDAR database," in *Proc. 15th Int. Conf. Pattern Recognit. (ICPR)*, 2000, pp. 569–572.
[56] S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, M. Alshayeb, M. T. Parvez, V. Märgner, and G. A. Fink, "KHATT: An open Arabic offline handwritten text database," *Pattern Recognit.*, vol. 47, no. 3, pp. 1096–1112, 2014.
[57] N. Shaffi and F. Hajamohideen, "UTHCD: A new benchmarking for Tamil handwritten OCR," *IEEE Access*, vol. 9, pp. 101469–101493, 2021, doi: 10.1109/ACCESS.2021.3096823.
[58] O. Akbani, A. Gokrani, M. Quresh, F. M. Khan, S. I. Behlim, and T. Q. Syed, "Character recognition in natural scene images," in *Proc. Int. Conf. Inf. Commun. Technol. (ICICT)*, Dec. 2015, pp. 1–6.
[59] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, V. Papavassiliou, F. Sunistira, and V. Katsouros, "GRPOLY-DB: An old Greek polytonic document image database," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 646–650.
[60] H. Zhang, J. Guo, G. Chen, and C. Li, "HCL2000—A large-scale handwritten Chinese character database for handwritten character recognition," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, 2009, pp. 286–290.
[61] (Jan. 30, 2022). *The Washington Database*. [Online]. Available: http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database
[62] George Washington. (Dec. 25, 1755). *George Washington Papers, Series 2, Letterbooks 1754 to 1799: Letterbook 1*. [Online]. Available: https://www.loc.gov/item/mgw2.001/
[63] (Jan. 30, 2022). *The ESPOSALLES Database*. [Online]. Available: http://dag.cvc.uab.es/the-esposalles-databas/
[64] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. R. Terrades, and A. Juan, "The GERMANA database," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, 2009, pp. 301–305.
[65] N. Serrano, F. Castro, and A. Juan, "The RODRIGO database," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2010, pp. 2709–2712.
[66] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 934–942, May 2012.
[67] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proc. Workshop Historical Document Imag. Process.*, 2011, pp. 29–36.
[68] A. Cheddad, H. Kusetogullari, A. Hilmkil, L. Sundin, A. Yavariabdi, M. Aouache, and J. Hall, "SHIBR—The Swedish historical birth records: A semi-annotated dataset," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15863–15875, Nov. 2021.

AMIR YAVARIABDI received the Ph.D. degree in medical image analysis from the University of Auvergne, Clermont-Ferrand, France, in 2014. In 2015, he joined the Department of Mechatronics Engineering, KTO Karatay University, as an Assistant Professor. His research interests include deep learning, image registration, 3-D computer vision, and remote sensing.

HUSEYIN KUSETOGULLARI (Member, IEEE) received the Ph.D. degree from the University of Warwick, U.K., in 2012. After completing his Ph.D. degree, he worked as a Postdoctoral Researcher with the Image Processing and Expert Systems Laboratory, School of Engineering, University of Warwick. After that, he worked on multi-objectivization problems with the Department of Computer Science, Aberystwyth University, as a Research Associate. He is currently working as a Senior Lecturer with the Department of Computer Science, Blekinge Institute of Technology, and the School of Informatics, University of Skövde. His research interests include image and video processing, artificial intelligence, evolutionary methods, remote sensing, and optimization. He is an Associate Editor of the *Signal, Image and Video Processing* (Springer).

TURGAY CELIK received the second Ph.D. degree from the University of Warwick, Coventry, U.K., in 2011. He is currently a Professor of digital transformation and the Director of the Wits Institute of Data Science, University of Witwatersrand, Johannesburg, South Africa. His research interests include computer vision, (explainable) artificial intelligence, (health) data science, data-driven optimal control, and remote sensing. He is an Associate Editor of *BMC Medical Informatics and Decision Making*, *IET Electronics Letters*, IEEE Access, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), and *Signal, Image and Video Processing* (Springer).

SHIVANI THUMMANAPALLY is currently pursuing the master's degree with the Department of Computer Science, Blekinge Institute of Technology. Her research interests include image processing, handwritten document analysis, computer vision, and machine learning.

SAKIB RIJWAN is currently pursuing the master's degree with the Department of Computer Science, Blekinge Institute of Technology. His research interests include image processing, handwritten document analysis, computer vision, and machine learning.

JOHAN HALL received the Ph.D. degree from Växjö University, Växjö, Sweden, in 2008. He did his research in natural language processing (NLP) with special interest in natural language parsing. He also worked as a Researcher at Uppsala University. He is currently working as the Chief Technology Officer (CTO) at Arkiv Digital AB, Sweden. His research interests include machine learning, image processing, and natural language processing.