http://doi.org/10.35784/iapgos.6521 received: 27.08.2024 | revised: 14.01.2025 | accepted: 13.03.2025 | available online: 31.03.2025

CKSD: COMPREHENSIVE KURDISH-SORANI DATABASE

Jihad Anwar Qadir[1], Samer Kais Jameel[2], Wshyar Omar Khudhur[3], Kamaran H. Manguri[1]
[1]University of Raparin, Software Engineering Department, Ranya, Iraq, [2]University of Raparin, Department of Computer Science, Ranya, Iraq, [3]Erbil Polytechnic University, Koya Technical Institute, Department of Information Technology, Koya, Erbil, Iraq

Abstract. *Every individual has a specific language with which he/she communicates. Each language has special letters and features distinguishing it from other languages. Ideas, cultures, and sciences are exchanged through some notions of languages, including retrieval, translation, and classification of texts from journals, books, journals, research, and the internet. It is accomplished through database availability. Unfortunately, due to some reasons, Kurdish language databases may be rare or non-existent. In the present study, a Comprehensive Kurdish-Sorani Database (CKSD) is generated, which contains datasets of dates, letters, and common words in the Kurdish language, as well as the documents employed for the extraction of these datasets. Elements of these collections were extracted from the written documents in 27 different fonts. It bestows a comprehensiveness feature to the CKSD database that can be utilized by researchers. In order to determine the extent to which classifiers can categorize such data, these data were utilized in this study. Indeed, this study demonstrated the reliability of this data and its suitability for use in the fields of machine learning and other artificial intelligence applications.*

Keywords: CKSD, OCR, font recognition, character recognition, font style

CKSD: KOMPLEKSOWA BAZA DANYCH KURDYJSKO-SORANI

Streszczenie. *Każda osoba ma określony język, którym się komunikuje. Każdy język ma specjalne litery i cechy odróżniające go od innych języków. Idee, kultury i nauki są wymieniane za pośrednictwem niektórych pojęć języków, w tym wyszukiwania, tłumaczenia i klasyfikacji tekstów z czasopism, książek, badań i Internetu. Jest to możliwe dzięki dostępności baz danych. Niestety, z pewnych powodów bazy danych w języku kurdyjskim mogą być ograniczone lub nie istnieć. W niniejszym badaniu wygenerowano kompleksową bazę danych kurdyjsko-sorani (CKSD), która zawiera zbiory danych dat, liter i popularnych słów w języku kurdyjskim, a także dokumenty wykorzystane do ekstrakcji tych zbiorów danych. Elementy tych zbiorów zostały wyodrębnione z dokumentów pisanych 27 różnymi czcionkami. Nadaje to bazie danych CKSD cechę kompleksowości, która może być wykorzystywana przez badaczy. W celu określenia zakresu, w jakim klasyfikatory mogą kategoryzować takie dane, dane te zostały wykorzystane w tym badaniu. Badanie to wykazało wiarygodność tych danych i ich przydatność do wykorzystania w dziedzinie uczenia maszynowego i innych zastosowań sztucznej inteligencji.*

Słowa kluczowe: CKSD, OCR, rozpoznawanie czcionek, rozpoznawanie znaków, styl czcionki

Introduction

The number of documents in the Kurdish language uploaded to the Internet, including emails, articles, and different news types, such as political, cultural, sports, and artistic [14, 15]. Thus, it is required to arrange, archive, categorize, and retrieve these data [22] using various approaches, such as text classification, as well as the techniques used in languages other than Kurdish. Moreover, as electronic data grows, the necessity for such techniques as text mining is increased for manipulating and understanding the linguistic and character patterns used in a large dataset [10, 16].

The interaction among people of the world is done through language [12]. People need communication with others worldwide for understanding others' traditions and cultures and for social and cultural exchange, which is accomplished via the translation of articles, scientific research, and books [13]. These require techniques for quick translation and understanding of texts in different languages, and given a large number of databases, the difficulty of understanding and analysis of texts can be overcome [4].

There are a larger number of studies carried out on texts in Arabic [20], Urdu [3], and English [7], including text classification, handwriting detection, translations, text mining, etc. These studies can be hardly compared with the studies conducted in those languages because of easy access to databases, the availability, comparison of results, etc. [8].

Although various techniques such as artificial intelligence, deep learning, and machine learning have been developed for text recognition, text classification, and text exploration [5] and despite the increased number of Kurdish-speaking individuals, still there are rare numbers of studies dealing with the classification [23] and retrieval [9] of the Kurdish language [6], which is because of the absence of a common database containing Kurdish symbols and letters to be used by researchers.

The present study proposes a database known as the Comprehensive Kurdish-Sorani Database (CKSD) that would be accessible by researchers to facilitate of their work and motivation for studying the Kurdish language. This database is applicable in various areas since it includes datasets of Kurdish language letters, with the letter written in 27 fonts. This database was composed of different resources on the archives and Internet. These datasets are important because allow application of the text classification problems on Kurdish-Sorani text documents.

1. Literature review

In the realm of OCR and font classification, a multitude of research endeavors have been undertaken, reflecting the dynamic and evolving nature of this field. This abundance of scholarly exploration underscores the significance and multifaceted dimensions inherent in OCR and font classification. Some of the related works have been reviewed and listed below.

Yaseen and Hassani [24] stands out with its focused attention on the complexities of Kurdish scripts. Their proposed method utilizing contour labeling-based segmentation effectively addresses cursive and diacritic features, resulting in an impressive average recognition rate of 90.82%. This research fills a crucial gap in the OCR domain, especially for languages with unique scripts like Kurdish.

Wang et al. [21] is remarkable, as it not only presents a large-scale dataset for font recognition but also leverages Convolutional Neural Networks (CNNs) for improved performance. The achieved accuracy of 80% (top-5).

Another research study [17] highlighted the significance of single font OCR for enhanced accuracy. Their proposed CNN-based framework and dataset creation demonstrate state-of-the-art performance in font recognition across different languages with 98.8%-line level accuracy. on the challenging dataset of 40 Arabic fonts showcases the model's effectiveness, providing valuable font similarity measures for font selection and suggestion.

Ahmed et al. [2] researched on Kurdish Handwritten Character Recognition Using Deep Learning Techniques is groundbreaking, addressing the dearth of OCR tools for Kurdish. The deep learning model achieves a remarkable accuracy of 96%, making significant strides in offline Kurdish handwriting recognition.

A novel combination of Hidden Markov Models and Harmony Search algorithms for online Kurdish character recognition is presented in [25]. Achieving a 93.52% recognition rate, the proposed method surpasses similar systems using HMM as their main recognizer.

Tofiq and Hussein [18] worked on Kurdish Character Segmentation Algorithm for Optical Character Recognition provided an efficient algorithm for segmenting Kurdish characters. The projection-based approach showcases an impressive accuracy of 98.6% in segmenting Kurdish words, even in the face of over-segmentation challenges.

Overall, these research papers significantly contribute to the advancement of OCR technology for various languages and scripts. They address critical challenges and offer promising solutions that are poised to make a profound impact on the field.

2. An overview of the Kurdish language

The Kurdish-Sorani language is spoken by individuals living in the Kurdistan region of Iraq [1], some regions of Iran and northern Syria, and some people in southern areas of Turkey. The Kurdish script is the closest to Arabic since it is written from right to left and in terms of drawing letters [11]. The difficulty in dealing with the Kurdish language is that every word includes a set of letters that are related to each other in a specific way since one letter has many shapes based on its occurrence in the word, that is, the middle of the word, beginning of a word, or end of a word (Table 1).

Table 1. The Kurdish letters and its sound

letter sound	Kurdish letter	letter sound	Kurdish letter	letter sound	Kurdish letter
G	گ	Ř	ڕ	Î	ئ
L	ل	Z	ز	A	ا
Ļ	ڵ	J	ژ	B	ب
M	م	S	س	P	پ
N	ن	Š	ش	T	ت
H	هـ	E / Ɛ	ع	C	ج
E	ه	Ğ	غ	Ç	چ
W	و	F	ف	Ḥ	ح
O	ۆ	V	ڤ	X	خ
Û	وو	Q	ق	D	د
Y	ی	K	ک	R	ر
Ê	ێ				

3. An overview of database

The Kurdish language includes 34 letters [19]. The data collection process is a labor-intensive and difficult process, which requires separating the letters from every single word. It should be noted that Arabic names are written and used as they are in the Kurdish language. In the proposed database, it was taken into account that some letters were added that are originally Arabic letters, like (ع, ص, ض, ظ, ط). To this end, there is a total number of 39 letters.

In the second part of the database, there are the Kurdish language words. Moreover, a section was added for the date, which contained the dates written in the Latin and Kurdish styles. The last part of the database includes some original files from which these letters were collected. The general structure of CKSD database is shown in Fig. 1.

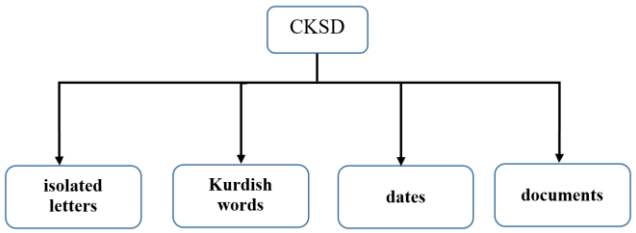

Fig. 1. Structure of the CKSD of database

3.1. Data collection

Official documents, daily newspapers, research articles, and other resources were used for collecting the data presented in the present work. We considered each letter as an individual case known as an isolated letter. As previously noted, each letter has a different shape based on its appearance location in the beginning, middle, or end of the word. Therefore, the dataset includes all cases. Furthermore, the letters were gathered from documents written in different fonts. Only 27 types of fonts are adopted in this dataset. Thus, each letter is presented in different shapes based on where it appears in the word, making the proposed dataset a comprehensive dataset. The Kurdish letters and the related sounds are illustrated in Table 1.

3.2. Isolated letters dataset

The dataset includes a total of 39 letters. Each letter was written with 27 types of fonts. Additionally, there are three different shapes for each letter. There exist no lowercase and uppercase letters in the Kurdish language as is the case in English. Letters are related to each other and form words that have different meanings in the language. The letter could change by changing its appearance location. For instance, the shape of the letter (ح) at the middle and beginning of the word is (حـ) as in the words (محمد, حمد), while its shape is (ح) at the end of the word as in the word (سامح). The shapes of the letters in different places of incidence are shown in Table 2.

There exist letters with no direct relationship to the letter following it, and they only may a have relationship with the letter before it, like (و) in the word (چونی) and letter (ه) as in the word (مەریم). The letter (أ) does not connect if it appears at the beginning of the word. All the characters with this property are presented in Table 2.

Table 2. The shapes of letters according to their occurrence in words (beginning of the word, the middle word, and the last word)

end of the word	center of the word	start word	end of the word	center of the word	start word	end of the word	center of the word	start word
ـڵ	ـڵـ	ڵـ	ـز	ـز	ز	ـئ	ـئـ	ئـ
ـم	ـمـ	مـ	ـژ	ـژ	ژ	ـا	ـا	ا
ـن	ـنـ	نـ	ـس	ـسـ	سـ	ـب	ـبـ	بـ
ـە	ـھـ	ھـ	ـش	ـشـ	شـ	ـپ	ـپـ	پـ
ه	ه	ه	ـع	ـعـ	عـ	ـت	ـتـ	تـ
ـو	ـو	و	ـغ	ـغـ	غـ	ـج	ـجـ	جـ
ـۆ	ـۆ	ۆ	ـف	ـفـ	فـ	ـچ	ـچـ	چـ
ـوو	ـووـ	ووـ	ـڤ	ـڤـ	ڤـ	ـح	ـحـ	حـ
ـی	ـیـ	یـ	ـق	ـقـ	قـ	ـخ	ـخـ	خـ
ـێ	ـێـ	ێـ	ـک	ـکـ	کـ	ـد	ـدـ	دـ
			ـگ	ـگـ	گـ	ـر	ـرـ	رـ
			ـل	ـلـ	لـ	ـڕ	ـڕـ	ڕـ

Some letters contain language movements such as (ۆ) in the word (مامۆستا). The letters with marks below or above them are presented in Table 3.

Table 3. All Kurdish letters have language movements

ڕ	ڵ	ێ	ۆ	

3.3. Kurdish word dataset

There is a file in the database containing a group of Kurdish words that are commonly used, like the names of the months in the Kurdish language, the Kurdish season's name, the numbers from 0 to 9, and the names of the weekdays. Each word was written 27 times with different fonts. This file could be interesting for researchers since it serves as a facilitator for the classification, retrieval, translation, and, distinction processes. The numbers, name of seasons, and months names in the Kurdish language are illustrated in Table 4, 5, and 6 respectively.

Table 4. Name of the numbers in Kurdish and Arabic languages

Kurdish number names	Arabic number	Kurdish numbers
سفر	0	٠
یەک	1	١
دوو	2	٢
سێ	3	٣
چوار	4	٤
پێنج	5	٥
شەش	6	٦
حەوت	7	٧
هەشت	8	٨
نۆ	9	٩

Table 5. Names of the seasons in Kurdish language

season	Kurdish season
Spring	بەهار
Summer	هاوین
Fall	پایز
Winter	زستان

Table 6. Names of the months in the Kurdish language

English months name	Kurdish months name
March	نەورۆز (خاکەلێوە)
April	بانەمەڕ
May	جۆزەردان
June	پووشپەڕ
July	خەرمانان
August	گەلاوێژ
September	ڕەزبەر
October	گەلاڕێزان(خەزەڵوەر)
November	سەرماوەز
December	بەفرانبار
January	ڕێبەندان
February	ڕەشەمێ

3.4. Date dataset

The dates file includes a collection of dates. Each group is written in a different font and in a different style. Two calendars, the English calendar, and the Kurdish calendar, were used (Table 7).

Table 7. Two types of calendars Kurdish and English

English style	Kurdish style
01/01/2020	٠١/٠١/٢٧١٩
26/07/2011	٢٦/٠٧/٢٧١٠
15/12/1998	١٥/١٢/٢٦٩٧

Various studies and literature have explored the recognition of letters and fonts in different languages. We were unable to find any prior research related to font recognition specific to the Kurdish language in the available literature. This study aims to identifying the font used and extracting text from the images.

4. Proposed methods

In our proposed approach, we employ deep learning methods to recognize both fonts and letters within images. Our system involves input an image and applying a deep convolutional neural network to extract letters.

This method is dedicated to recognizing Kurdish, Arabic, and Persian letters and implementing a font recognition system. The proposed algorithm is developed specifically for the identification of Kurdish letters. The process involves many steps which are shown in Fig. 2, and listed as follows:

1. Collecting data: Collecting the printed documents to find the most used fonts.
2. Writing words: Writing words to have all letters in the different style and places (connected and disconnected words).
3. Dataset Preparation: Segment all written words to letters for the selected fonts.
4. Applying preprocessing: perform preprocessing algorithms, including image resizing, sharpening, smoothing, and contrast enhancement.
5. Data augmentation: to increase number of letters of the dataset based on different transformation.
6. Dataset splitting: Datasets are portioning to cross-validation of 80% (training data) and 20% (test and validation).
7. Train the proposed deep and transfer models: with tuning hyper-parameters of each model.
8. Performance metrics' evaluation: precision, recall, F1-score, and average accuracy.
9. Font recognition and classification: The multi-classification method identifies font style.

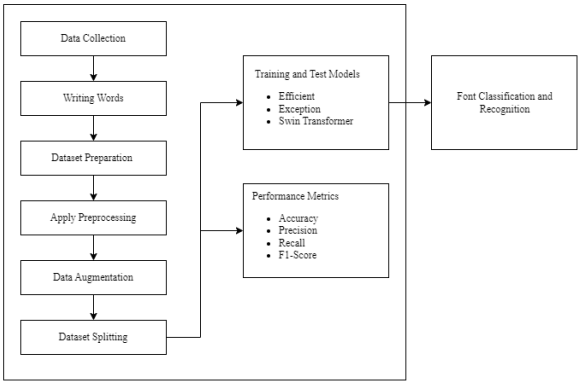

Fig. 2. General diagram of the study

5. Experimental results

The proposed dataset was the subject of a series of experiments, which demonstrated the effectiveness of the well-known classifiers in accurately classifying the dataset. The testing was done with efficient, exception, and swin transformer. The results are presented in Table 8, which indicates that our dataset can be a trustworthy group in many applications that involve words and letters in the same way as Latin letters are used in other languages or other scripts.

Table 8. Performance of the classifiers to detect letters from CKSD using one type of font

measurements\ classifiers	efficient	exception	swin transformer
F1-score	0.9836	0.9649	0.79
Recall	0.984	0.9652	0.79
Precision	0.9868	0.9654	0.98
Accuracy	0.984	0.9654	0.98

The data set consists of 27 groups, each of which represents a different font, as we already explained. from Table 1, each font type was employed independently, which means that each font was used in the process of the model individually. The ratio of the train set and test set, is 20% and 80% respectively. The results are quite good, with the employed measures having values of not less than 96%, which is a good indication of the validity and dependability of the CKSD dataset for use in research projects.

For all fonts type in the data set, experiment results are shown in Table 9. The model is tested with a different sort of font after being trained on one. This experiment is repeated numerous times using every type of currently used font, the confusion matrix is calculated, indicating that the classifiers employed are not very effectively able to distinguish between letters of various types of fonts. This is thought to be an obstacle for this dataset, which researchers can use in order to develop models capable of meeting this challenge.

Table 9. Performance of the classifiers to detect letters from CKSD using different types of font

measurements\ classifiers	efficient	exception	swin transformer
F1-score	0.68	0.6807	0.63
Recall	0.7029	0.6991	0.65
Precision	0.6771	0.6904	0.63
Accuracy	0.7032	0.6993	0.65

Despite significant efforts to improve font classification outcomes, the task remains challenging for several reasons. The similarity between fonts, particularly those used in official writing, creates difficulty in recognize them due to subtle shape and size differences that lead to confusion during classification. Additionally, dealing with a large number of font classes (27 types) poses recognition challenges for the model. Using single letters for classification lacks contextual information needed to understand each font's unique characteristics. Fonts are commonly used in real-world scenarios within words and sentences, where character combinations provide additional cues for recognition. Often similar fonts require the model to learn fine-grained details, demanding a vast amount of data and a sophisticated architecture. Training on individual letters alone might fail to capture the overall structure and stylistic nuances of each font, as fonts follow consistent design principles across various characters. Font types exhibit considerable variability in style due to factors like weight, slant, and thickness. A dataset lacking comprehensive coverage of these variations can hinder the model's ability to generalize effectively for specific font types. Furthermore, the unique characteristics of the Kurdish language, with closely resembling letters such as "ب" and "پ," "ق" and "ڤ," or "ن" and "ت," contribute to the difficulty of the font classification task.

6. Conclusion

With the growing use of the language, the necessity for distinguishing and separating the line has also been increased. Hence, building a comprehensive database with the groups of letters and different units seems necessary. The CKSD provides a comprehensive database containing all the letters that are used in the Kurdish language. Moreover, this database includes a set of dates written in the style used by these language speakers. A file is included in this database that presents the most common words in the Kurdish language, including name of months and seasons, and numbers in the Kurdish language. The CKSD is the first database that contains documents from which words, dates, and letters were obtained.

References

[1] Abdulrahman R. O. et al.: Developing a Fine-Grained Corpus for a Less-Resourced Language: The Case of Kurdish. arXiv 11467, 2019.
[2] Ahmed R. M. et al.: Kurdish Handwritten Character Recognition Using Deep Learning Techniques 46, 2022, 119278.
[3] Akhter M. P. et al.: Exploring Deep Learning Approaches for Urdu Text Classification in Product Manufacturing 16(2), 2022, 223–248.
[4] Allahyari M. et al.: A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques. arXiv 1707.02919v2, 2017.
[5] Alwehaibi A., Roy K.: Comparison of Pre-Trained Word Vectors for Arabic Text Classification Using Deep Learning Approach. 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, 1471–1474.
[6] Celik S.: Collaborative English Language Learning in Primary School: A Sequential Explanatory Study in Kurdistan Region of Iraq. Id No. 2520, 2019.
[7] Chen K. et al.: Defect Texts Mining of Secondary Device in Smart Substation with GloVe and Attention-Based Bidirectional LSTM. Energies 13(17), 2020, 4522.
[8] Choudhary P. et al.: A Four-Tier Annotated Urdu Handwritten Text Image Dataset for Multidisciplinary Research on Urdu Script. Information Processing. 15(4), 2016, 1–23.
[9] Gómez L. A. et al.: Single Shot Scene Text Retrieval. European Conference on Computer Vision (ECCV), 2018, 700–715.
[10] Hakim L. et al.: Text Mining of UU-ITE Implementation in Indonesia. Journal of Physics: Conference Series 1, 2018.
[11] Hashimi A. O.: Ajami Tradition in Non-Islamic Society: The Roles of Ajami-Arabic Scripts in Keeping Records and Documentation. KIU Journal of Humanities 5(2), 2020, 373–379.
[12] Jana H. P.: The Tools of Language and Literature in Sustainable Development of the Globizen: An Enquiry with Special Reference to English Language and Literature. International Journal of Yogic, Human Movement and Sports Sciences 3(2), 2018, 318–324.
[13] Mallery G.: Sign Language among North American Indians Compared with That among Other Peoples and Deaf-Mutes. Vol. 14, Walter de Gruyter GmbH & Co KG, 2019.
[14] Rashid T. A. et al.: A Robust Categorization System for Kurdish Sorani Text Documents. Information Technology Journal 16(1), 2017, 27–34.
[15] Sheyholislami J.: Identity, Language, and New Media: The Kurdish Case. Language Policy 9, 2010, 289–312.
[16] Sun W. et al.: Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. Journal of Healthcare Engineering 2018, 4302425 [https://doi.org/10.1155/2018/4302425].
[17] Tensmeyer C. et al.: Convolutional Neural Networks for Font Classification. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 1, 2017, 985–990.
[18] Tofiq T. A., Hussein J. A.: Kurdish Text Segmentation Using Projection-Based Approaches. UHD Journal of Science and Technology 5(1), 2021, 56–65.
[19] Veisi H. et al.: Toward Kurdish Language Processing: Experiments in Collecting and Processing the Asosoft Text Corpus. Digital Scholarship in the Humanities 35(1), 2020, 176–193.
[20] Wahdan A. et al.: A Systematic Review of Text Classification Research Based on Deep Learning Models in Arabic Language. International Journal of Electrical and Computer Engineering (IJECE) 10(6), 2020, 6629–6643.
[21] Wang Z. et al.: DeepFont: Identify Your Font from an Image. 23rd ACM International Conference on Multimedia, 2015.
[22] Wiedemann G., Wiedemann: Text Mining for Qualitative Data Analysis in the Social Sciences. Vol. 1, Springer, 2016.
[23] Yao L. et al.: Graph Convolutional Networks for Text Classification. AAAI Conference on Artificial Intelligence 3(1), 2019, 7370–7377.
[24] Yaseen R., Hassani H.: Kurdish Optical Character Recognition. UKH Journal of Science and Engineering 2(1), 2018, 18–27.
[25] Zarro R. D. et al.: Recognition-based online Kurdish character recognition using hidden Markov model and harmony search. I. J. Technology 20(2), 2017, 783–794.

M.Sc. Jihad Anwar Qadir
e-mail: jihad.qadir@uor.edu.krd

He received his B.S. degree in Computer Science from the University of Sulaymaniyah, Sulaymaniyah, Iraq, in 2011, and his M.Sc. degree in Electronics and Computer Engineering from the Institute of Natural and Applied Sciences at Hasan Kalyoncu University, Gaziantep, Turkey, in 2016. Currently, he is a member of Department of Software and Informatics Engineering, University of Raparin.

https://orcid.org/0000-0003-3958-814X

Ph.D. Samer Kais Jameel
e-mail: samer.kais@uor.edu.krd

He received his B.Sc. in Information Systems and Computer Science from Al-Mansour University, Baghdad, Iraq, his M.Sc. in Computer Science from B.A.M.U University, India, and his Ph.D. from Aksaray University, Turkey. He served as a lecturer in the Department of Computer Science at Dijla University, Baghdad, Iraq, from 2011 to 2012. Since 2012, he has been a lecturer at Raparin University, Sulaymaniyah, Iraq. His research interests include Machine Learning, Computer Vision, and Deep Learning.

https://orcid.org/0000-0003-2236-9303

M.Sc. Wshyar Omar Khudhur
e-mail: wshyar.khudhur@epu.edu.iq

Wshyar Omar Khudhur received his M.Sc. in Computer Engineering from Cyprus International University in the North Cyprus Region, Turkey, in 2015. He is currently an Assistant Lecturer at Erbil Polytechnic University, Koya Technical Institute, Information Technology Department, where he also served as the Head of the IT Department from 2019 to 2023.

https://orcid.org/0009-0009-8762-2884

Ph.D. Kamaran Manguri
e-mail: kamaran@uor.edu.krd

Lecturer in the Department of Software Engineering at the College of Engineering, University of Raparin. He holds a Bachelor's degree in Computer Systems Engineering from the Faculty of Engineering at Koya University in Iraq, a Master's degree in Electronics and Computer Engineering from Hasan Kalyoncu University in Turkey, and a Ph.D. in Computer Vision and Machine Learning from Erbil Technical Engineering College at Erbil Polytechnic University in Iraq. Kamaran's research interests include computer vision, machine learning, and related fields.

https://orcid.org/0000-0001-8567-3367