SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.04 NO. 3 2025
DOI: 10.51984/SUCP.V4I3.4228

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

Developing a Benchmark Data Set for Optical Character Recognition of Drug Names in Medical Prescriptions

*Maryam Aldaeb, Mohamed fadeel

Department of Computer Science, Faculty of Science, Sabha University, Libya

Keywords:
Data Set
Handwriting
Medical Errors
Optical Character Recognition
Medical Prescription

A B S T R A C T

Interpreting handwritten medication names remains a global challenge that often causes medical errors and delays in drug dispensing. This issue threatens patient safety and demands effective solutions. To address it, a data set was developed of handwritten drug names to support the training of optical character recognition (OCR) systems. The goal is to improve prescription readability, enable digital archiving, and support electronic health records. The data set can also be used in smart health applications to help patients manage their medications. The data set was built using contributions from 250 physicians and medical students in southern region of Libya. Each participant wrote 20 predefined drug names in standardized forms, producing 9,225 handwriting samples. Image processing techniques were used, such as converting color images to grayscale, to reduce memory requirements by reducing the size to one-third of the original size and increasing processing speed by ~66%. In addition, input normalization was applied by converting pixel. These techniques accelerate model training, enhance accuracy, and address artifacts caused by different handwriting sizes. This resource aims to reduce prescription-related errors, enhance data accessibility through searchable archives, and support faster and safer medication processing. The paper also proposes a practical approach to integrating intelligent systems in healthcare and highlights the importance of collaboration between clinicians and researchers to improve patient safety and care quality.

تطوير مجموعة بيانات مرجعية للتعرف البصري على اسماء الأدوية الخطية في الوصفات الطبية

*مريم الدائب و محمد فضيل

قسم الحاسوب، كلية العلوم، جامعة سبها، ليبيا

الكلمات المفتاحية:
الأخطاء الطبية
التعرف البصري على الحروف
الكتابة اليدوية
الوصفة الطبية
مجموعة البيانات

الملخص

لا يزال تفسير أسماء الأدوية المكتوبة بخط اليد يُمثل تحديًا عالميًا، وغالبًا ما يُسبب أخطاءً طبيةً وتأخيرًا في صرف الأدوية. تُهدد هذه المشكلة سلامة المرضى وتتطلب حلولًا فعّالة. ولمعالجة هذه المشكلة، طُوّرت مجموعة بيانات لأسماء الأدوية المكتوبة بخط اليد لدعم تدريب أنظمة التعرف الضوئي على الحروف .(OCR) الهدف هو تحسين قابلية قراءة الوصفات الطبية، وتمكين الأرشفة الرقمية، ودعم السجلات الصحية الإلكترونية. كما يُمكن استخدام مجموعة البيانات في تطبيقات الصحة الذكية لمساعدة المرضى على إدارة أدويتهم. بُنيت مجموعة البيانات باستخدام مساهمات من 250 طبيبًا وطالب طب في المنطقة الجنوبية من ليبيا. كتب كل مشارك 20 اسمًا دوائيًا مُحددًا مسبقًا بصيغ موحدة، مما أدى إلى إنتاج 9225 عينة خطية. استُخدمت تقنيات معالجة الصور مثل تحويل الصور الملونة إلى تدرج الرمادي، لتقليل متطلبات الذاكرة عن طريق تقليل الحجم إلى ثلث الحجم الأصلي وزيادة سرعة المعالجة بنسبة 66% تقريبًا. بالإضافة إلى ذلك، طُبّقت معايرة الإدخال عن طريق تحويل قيم البيكسل، تعمل هذه التقنيات على تسريع تدريب النماذج وتعزيز الدقة ومعالجة العيوب الناتجة عن أحجام الكتابة اليدوية المختلفة. يهدف هذا المورد إلى تقليل الأخطاء المتعلقة بالوصفات الطبية، وتعزيز إمكانية الوصول إلى البيانات من خلال الأرشيفات القابلة للبحث، ودعم معالجة الأدوية بشكل أسرع وأكثر أمانًا. وتقترح الورقة أيضًا نهجًا عمليًا لدمج الأنظمة الذكية في الرعاية الصحية، وتُسلط الضوء على أهمية التعاون بين الأطباء والباحثين لتحسين سلامة المرضى وجودة الرعاية المقدمة.

*Corresponding author:
E-mail addresses: Mar.adaeb@ sebhau.edu.ly ,(M. fadeel) fadeel1@sebhau.edu.ly
Article History : Received 20 February 2025 - Received in revised form 01 September 2025 - Accepted 07 October 2025

1. Introduction

Medical Prescriptions can have serious and sometimes fatal consequences[1] [2][3]. According to the Institute of Medicine (IoM), medical errors cause at least 44,000 preventable deaths annually in the United States, with approximately 7,000 linked to illegible handwriting [4][5]. This is a global issue, as evidenced by research from Egypt's National Academy of Sciences, which revealed over 1.5 million people are harmed each year due to misinterpreted prescriptions [2][6][7]. The problem of misreading medication names is also widespread among healthcare workers, with Brits et al. reporting a rate of 20%, potentially leading to incorrect drug dispensing, wrong dosages, and fatal outcomes [8][4]. This challenge is particularly pronounced in developing countries like Bangladesh, where physicians often work under intense pressure and 97% of handwritten prescriptions were found to be illegible [9][10].

Driven by these alarming statistics, researchers are increasingly turning to technological solutions like handwriting recognition (OCR) systems.

Using machine learning and image processing, OCR offers a promising approach to automate interpretation and reduce handwriting errors, thus improving patient safety [1][3][7]. However, developing OCR models requires large-scale, high-quality, domain-specific datasets of English medical terminology. To address this gap, this paper introduces a comprehensive dataset of handwritten drug names, meticulously curated to reflect real-world writing variations. This resource supports training and validation of state-of-the-art OCR and deep learning models, aiming to significantly enhance recognition accuracy, facilitate prescription digitization, and ultimately advance healthcare quality and safety.

2. Literature Review

Over recent decades, many methods have been proposed to recognize handwriting in medical prescriptions, aiming to reduce errors in medication dispensing. Several data sets of handwritten drug names have been developed to support these efforts. Below is a summary of key studies that have addressed the development and use of such data sets:

One of the earliest studies [11] introduced a handwritten data set developed by the University of Bern for writer recognition. Initially, it contained 1,066 samples from 400 writers, which was later expanded to 1,539 samples from 657 writers. This data set includes metadata such as writer identity, original text, and segmentation at the line, sentence, and word levels.

In addition, another study [12] analyzed common spelling errors in 30 drug names from hospital medical prescriptions in Birmingham, UK. A test set of 325,979 entries revealed 3,872 misspellings (1.17%) due to hidden reference variables, while a control set of 470,064 entries showed 766 such cases (0.16%). The most frequent errors involved letter substitution (e.g., 'i' to 'y') and deletions.

Moreover, study [7] proposed a method for recognizing handwritten drug names using English prescriptions. Text lines were segmented using horizontal projection, while word segmentation relied on vertical histograms. Features were extracted with the convex hull method, and classification was performed using an SVM model. On a data set of 50 prescriptions, the system achieved 95% accuracy in line segmentation, 92% in word segmentation, and 85% in drug name recognition. However, challenges included connected characters and variations in handwriting styles.

Similarly, study [4] addressed the challenge of recognizing handwritten medical prescriptions by proposing a word detection approach. This method was tested on 500 medical prescriptions using a string-matching algorithm with distance correction and a Hidden Markov Model (HMM) trained on the IAM English Sentences data set. The system achieved 97.5% accuracy in word recognition and improved keyword retrieval accuracy by 6.8%. The authors recommended using such systems for archiving handwritten medical prescriptions and enabling further analysis.

In this context, Farjado et al. [13] developed a data set containing 1,800 images of 12 handwritten drug names, collected from 50 physicians in clinics and hospitals across Metro Manila, Quezon City, and Taytay, Rizal (Philippines). The recognition system trained on this data set achieved a validation accuracy of 72%.

Furthermore, study [2] used a Convolutional Recurrent Neural Network (CRNN) to recognize handwritten English prescriptions. Given the difficulty of interpreting abbreviations and medical terms in illegible handwriting, the system aimed to reduce related errors. It was trained on 50 medical prescriptions (45 valid, 10 invalid) consisting of short texts, typically two to three words. The model achieved an accuracy of 98% in recognizing prescription content.

Similarly, Hassan et al. [3] developed a convolutional neural network (CNN) model to recognize handwritten drug names using a data set collected from Egyptian prescriptions. The model achieved a training accuracy of 73% and a test accuracy of 50%, highlighting the complexity of the task.

Additionally, study [14] introduced the HP_DocPres data set, which includes 11,340 samples of handwritten and printed words from various prescriptions. This data set was designed to classify and distinguish between handwritten and printed medical text.

Furthermore, study [15] employed a medical terminology corpus of 480 handwritten words (360 in English and 120 in Bengali). A graphical user interface was used to visualize the output, and the model was trained using over 200 Medical prescriptions from the IAM data set. The architecture combined five CNNs, two RNNs, and a CTC layer. The system achieved 50% accuracy on the test set, which included 17,431 handwritten samples.

Moreover, study [16] collected over 2,700 medical documents—including medical prescriptions and reports—from pharmacies and clinics in Sidi Bel Abbes, Algeria. These documents were segmented into word-level samples, manually classified, and pre-processed, resulting in 5,700 examples of handwritten medical terms.

In a related effort, study [9] addressed the issue of prescription accuracy using a deep learning approach based on the VGG16 model. This study is among the first to apply this model to handwritten medical prescriptions. The data set, collected from hospitals in Bangladesh and published on Kaggle, included 3,120 training samples and 779 test samples. The data featured a variety of handwriting styles and medical terms.

Finally, study [10] introduced an advanced method for extracting drug names using a combination of Mask R-CNN and Transformer-based OCR (TrOCR) with Multi-Head Attention and Positional Embeddings. The data set included 1,000 handwritten medical prescriptions from 50 physicians in Pakistan. Data augmentation increased the sample size to 9,920 images. The model achieved a Character Error Rate (CER) of just 1.4%, demonstrating high accuracy and generalizability across handwriting styles.

3. Materials and Methods

The data set of handwritten drug names was built through a structured process, as illustrated in Figure 1. The methodology includes form design, data collection, image acquisition, processing, sorting, and normalization.

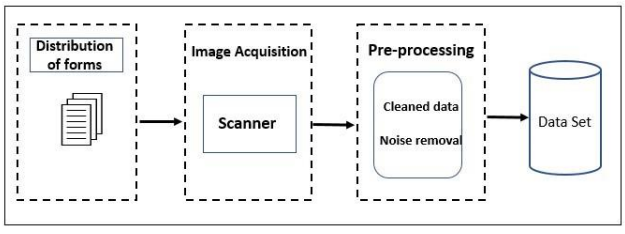

Fig.1: The Methodology used to identify handwritten drug names.

1- **Data Preparation**: Four standardized forms were designed, each containing 20 drug names commonly used in handwritten prescriptions. These forms were distributed to participants for data entry. Figure 2 presents the classification of drug names used in the forms.

Names of medicines	Model No
Panadol - Osteocare - Augmenting - Fucidin - Buscopan	1
Aspirin75 - Flagyl - Ceftriaxoron - Folic Acid - Apisal	2
Omprazl - Uricol eff - Suprax - Feffole - Congestat	3
Voltaren- Azithromycin - Ketofan - Adol Syrp - Fucicort	4

Fig. 2: The classification of drug names used in the forms.

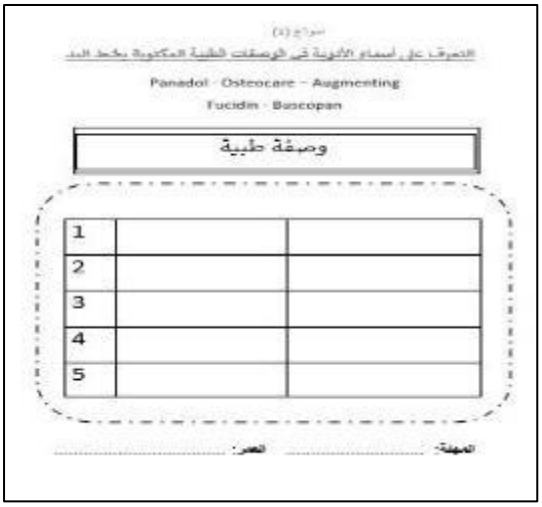

Fig.3: Models for writing drug names in manual prescriptions.

2- **Data collection:** Over 250 forms were distributed in various regions of southern Libya. Participants included male and female doctors and medical students aged between 25 and 66. After data collection, the results were analyzed using SPSS.

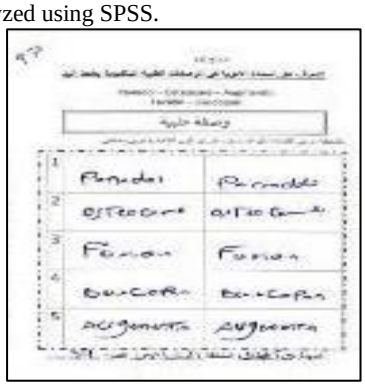

Fig. 4: Illustrates a sample of the completed forms.

3- **Image Acquisition:** Scanned images were captured using an Epson scanner at 400 dpi and saved in JPEG format. The data was organized into 10 main folders, each containing 50 completed forms. Subfolders were created for each drug name.

Image processing was performed using FastStone Image Viewer, which allowed for cropping, resizing, rotating, and enhancing image quality. Adjustments included contrast, brightness, and color correction. Figure 5 shows the process of cropping and preparing drug name image.

Fig.5: The drug name cropping and processing**.**

4- **The sorting:** After cropping and cleaning, noise was removed from the images to enhance data quality. The images were then sorted into folders based on drug names. A total of 9,225 usable handwriting samples were obtained.

5- **Image processing:**

- **Standardizing the image dimensions**: Rresizing the images to a fixed size, such as 300x120 pixels, ensures consistency and uniformity of inputs, Using the mathematical equation:

$$x' = x \cdot \frac{Wsrc}{W\,target}$$

$$y' = y \cdot \frac{Hsrc}{H\,target}$$

- **Processing acceleration:**

$$\text{Processing Time} = \text{Number of pixels} \times \text{Number of channels}$$

Reducing channels from 3 to 1 reduces time by ~66%.

6- **Grayscale Conversion:** Grayscale conversion involves converting color images to a single-channel image by standardizing the colors. This transformation aims to simplify the processing process, allowing focus on the textual structure of the image through:

• Converting color images to grayscale using the equation:
***I*gray** = 0.299 * R + 0.587 * G + 0.114 * B

• Reduces computational complexity by reducing the number of data channels from three to one Leads to reducing memory size.

$$Grayscale\ image\ size = \frac{color\ image\ size}{3}$$

7- **Normalization:** The normalization process aims to convert images of handwritten drug names into a standard format by normalize the input by converting pixel values from [0, 255] to [0, 1] or [-1, 1].

1. Normalize the values to [0, 1], The following equation can be used:

$$\mathbf{Inorm} = \frac{I}{255}$$

Where:

I: is the original pixel value.

Inorm: is the pixel value after normalization.

2.Normalize the values to [-1, 1] To convert pixel values to the range [-1, 1], the following equation can be used:

$$\mathbf{Inorm} = \frac{I}{127.5} - 1$$

The previous equations are used to speed up the training process and improve recognition accuracy. They also address distortions resulting from varying handwriting sizes.

It also adjusts the pixel value range from [0 , 255] to [0 , 1] or [-1,1] to speed up the training process and improve recognition accuracy. It also addresses distortions resulting from varying handwriting sizes.

4. Data Analysis Results:

Total Distributed Forms: More than 250 forms were distributed across various regions in southern Libya.

• **Geographic Coverage**:The survey included participants from Wadi Atba, Murzuq, Taraghin, Al-Mahalla Al-Sharqiya, Al-Qatrun, Ubari, Ghat, Sabha, Al-Buwanis, and Al-Shati (see Figure 6).

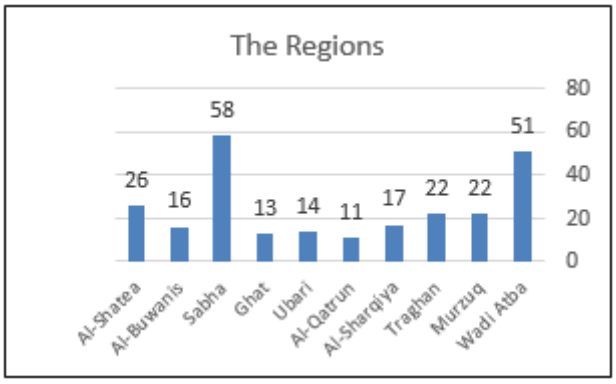

Fig.6: The regions distribution for filling out the forms.

• **Target Groups**: Participants were either medical students or practicing physicians (see Figure 7).

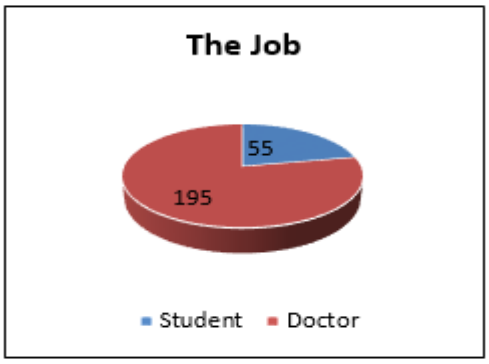

Fig.7: The target groups for filling out the form.

- **Age Distribution:** Participants ranged from 25 to 66 years of age, covering a broad age spectrum (see Figure 8).

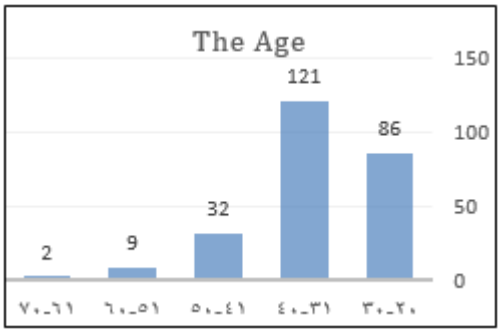

Fig.8: The Age distribution for filling out the forms.

- **Missing Data:** A total of 775 images were discarded due to incomplete entries or incorrect drug names that did not match the predefined list.

- **Final Data set:** Figure 9 shows examples of the valid medicine name images obtained from the forms.

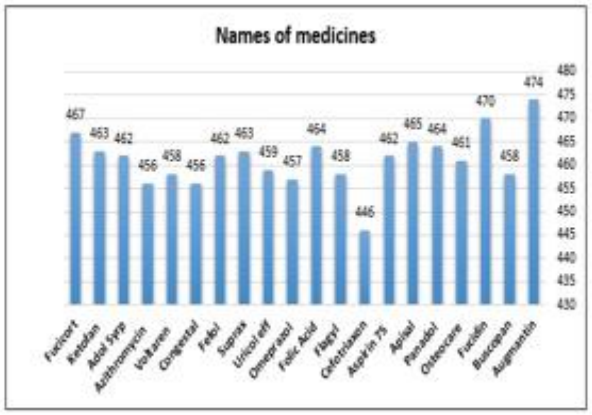

Fig.9: Images of medicines obtained from the distributed forms**.**

5. Discussion

The collected data set represents a valuable step toward addressing the widespread issue of illegible handwritten prescriptions, particularly in low-resource and high-pressure medical environments like those in southern Libya. Compared to previous data sets —many of which included limited drug names or small-scale contributions—this data set offers broader coverage with 9,225 samples from 250 participants across diverse regions and age groups.

One major contribution of this study is its regional specificity. While earlier works focused on data sets from South Asia or urban medical centers in Europe, this data set captures the handwriting variability of practitioners in North Africa. This aspect fills a notable gap in current research, making it especially useful for building models that generalize across cultural and linguistic contexts.

Another strength is the inclusion of both students and experienced doctors, ensuring a wide range of handwriting styles—from structured to highly cursive. This diversity adds complexity but also realism, which is essential for training robust OCR and deep learning models.

However, the study also faced challenges. About 775 samples were excluded due to incomplete or incorrect data, reflecting difficulties in collecting standardized handwriting samples in real-world settings. Furthermore, the study relied on traditional image enhancement and organization tools; future work could benefit from automated preprocessing pipelines and more advanced annotation systems.

In light of existing literature, the data set's size is comparable or superior to many similar works, but accuracy rates of recognition models still depend heavily on the preprocessing quality and the models used. It is therefore recommended that future research explore how this data set performs with various machine-learning models such as CNNs, CRNNs, or Transformers, as well as evaluate cross-data set transferability.

Overall, this work demonstrates the feasibility of building localized, high-quality handwriting data sets for medical use, and opens the door for the development of practical AI systems to support prescription digitization and error reduction.

6. Conclusion

This study underscores the importance of collecting samples of handwritten medical prescriptions as a fundamental step towards improving prescription accuracy and enhancing patient safety. A total of 9,225 handwriting samples were collected, constituting a valuable resource for training automated recognition models. By leveraging machine learning and computer vision techniques, these samples can be utilized to build systems capable of converting handwritten prescriptions into reliable digital records. This will contribute to reducing medication dispensing errors and increasing the efficiency of healthcare delivery. The use of image processing techniques—such as converting color images to grayscale—has contributed to reducing memory requirements to approximately one-third of the original size and increasing processing speed by up to nearly 66%. Furthermore, the process of standardizing inputs by converting pixel values from the range [0–255] to [0–1] or [-1–1] accelerates model training, enhances accuracy, and addresses distortions resulting from variations in handwriting size.

The researchers recommend fostering increased collaboration between the medical sector and the research community to build upon these findings. It is hoped that this work will inspire further research and practical initiatives to address this pressing challenge in the field of healthcare.

7. Recommendations

Based on the findings of this study, the following recommendations are proposed:

- Develop OCR Models Tailored to Regional Handwriting:

Future research should use this data set to train and evaluate machine-learning models that specifically address the characteristics of Arabic handwriting in medical prescriptions.

- Expand the Data set Scope:

Collect additional data from more geographic areas, involving pharmacists and nurses, to capture a wider range of writing styles and terminology.

- Integrate the Data set in Mobile Health Applications:

Collaborate with developers to build smartphone applications that use the data set for real-time drug name recognition and verification at the point of care.

- **Standardize Prescription Writing in Healthcare Institutions:**

Encourage hospitals and clinics to gradually shift toward digital prescription systems, or at least standardized paper templates that facilitate scanning and recognition.

- **Conduct Comparative Studies:**

Compare the performance of different deep learning architectures (e.g., CNN, CRNN, Transformers) using this data set to determine the most effective model for Arabic medical handwriting.

- **Public Data set Release:**

Make the data set publicly available for academic use, with appropriate ethical guidelines, to support reproducibility and cross-study validation.

8. References

[1]- A. Kumar, I. Dangi, S. Chowdary, and K. K. Pandey, "Ideal drug prescription writting," vol. 8, no. 3, pp. 634–654, 2019, doi: 10.20959/wjpps20193-12989.

[2]- R. Achkar, K. Ghayad, R. Haidar, S. Saleh, and R. Al Hajj, "Medical handwritten prescription recognition using CRNN," CITS 2019 - Proceeding 2019 Int. Conf. Comput. Inf. Telecommun. Syst., pp. 1–5, 2019, doi: 10.1109/CITS.2019.8862004.

[3]- E. Hassan, H. Tarek, M. Hazem, and S. Bahnacy, "Medical Prescription Recognition using Machine Learning," IEEE 11th Annu. Comput. Commun. Work. Conf. (CCWC, pp. 973–979, 2021.

[4]- S. Kumar Sarkar, "A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions Arghya Mukhejee 1a , Arunit Halder 1b , Subhrapratim Nath 1c," Adv. Ind. Eng. Manag., vol. 6, no. 2, pp. 90–96, 2017, doi: 10.7508/aiem.2017.02.006.

[5]- Y. Amol Rathod et al., "Handwritten Character Recognition Using CNN, KNN and SVM," Int. J. Technol. Eng. Arts Math. Sci., vol. 1, no. 2, pp. 2583–1224, 2022, doi: 10.11591/eei.v9i3.xxxx.

[6]- [6] E. Kamalanaban, M. Gopinath, and S. Premkumar, "Medicine box: Doctor's prescription recognition using deep machine learning," Int. J. Eng. Technol., vol. 7, no. 3.34 Special Issue 34, pp. 114–117, 2018, doi: 10.14419/ijet.v7i3.34.18785.

[7]- P. S. Dhande and R. Kharat, "Character Recognition for Cursive English Handwriting to Recognize Medicine Name from Doctor's Prescription," 2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017, pp. 1–5, 2017, doi: 10.1109/ICCUBEA.2017.8463842.

[8]- H. Brits et al., "Illegible handwriting and other prescription errors on prescriptions at National District Hospital , Bloemfontein Illegible handwriting and other prescription errors on prescriptions at National District Hospital , Bloemfontein," South African Fam. Pract., vol. 6190, pp. 1–4, 2017, doi: 10.1080/20786190.2016.1254932.

[9]- M. A.-A.-S. C. A. A. M. A. M. R. N. T. Mia, Abdur Rahim, "A Deep Neural Network Approach with Pioneering Local Dataset to Recognize Doctor's Handwritten Prescription in Bangladesh," IEEE Int. Conf. Adv. Comput. Commun. Electr. Smart Syst.), Dhaka, Bangladesh, pp. 1–6, 2024, doi: 10.1109/iCACCESS61735.2024.10499631.

[10]- U. Ali, M. Nadeem, H. Ishfaq, and W. Ali, "Leveraging Deep Learning with Multi-Head Attention for Accurate Extraction of Medicine from Handwritten Prescriptions," vol. 978-1–7281, 2024.

[11]- U. V. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, no. November 1999, pp. 709–712, 1999, doi: 10.1109/ICDAR.1999.791885.

[12]- R. E. Ferner and J. K. Aronson, "Nominal ISOMERs (Incorrect Spellings Of Medicines Eluding Researchers)— variants in the spellings of drug names in PubMed : a database review," 2016, doi: 10.1136/bmj.i4854.

[13]- L. J. Fajardo et al., "Doctor's Cursive Handwriting Recognition System Using Deep Learning," 2019 IEEE 11th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2019, pp. 6–11, 2019, doi: 10.1109/HNICEM48295.2019.9073521.

[14]- S. R. Dhar D, Garain A, Singh P, "HP_DocPres: a method for classifying printed and handwritten texts in doctor's prescription," Multimed. Tools Appl., 2020.

[15]- N. P. and N. Sampath., "Detecting and extracting information of medicines from a medical prescription using deep learning and computer vision," Int. Conf. Knowl. Eng. Commun. Syst., pp. 1–6, 2022.

[16]- K. Sehimi, "A Transfer Learning approach for handwritten drug names recognition," 2023 IEEE Int. Conf. Networking, Sens. Control, vol. 1, pp. 1–6, 2023, doi: 10.1109/ICNSC58704.2023.10318974.