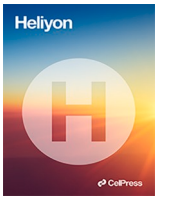
Heliyon 10 (2024) e36191

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Pixel embedding for grayscale medical image classification

Wensu Liu [a,b], Na Lv [a,b], Jing Wan [a,b], Lu Wang [c,d,**], Xiaobei Zhou [a,b,*]

[a] Key Laboratory of Obesity and Glucose/Lipid Associated Metabolic Diseases, China Medical University, Shenyang, Liaoning, 110122, China
[b] Institute of Health Sciences, China Medical University, Shenyang, Liaoning, 110122, China
[c] Library of Shengjing Hospital of China Medical University, Shenyang, Liaoning, 110122, China
[d] School of Health Management, China Medical University, Shenyang, Liaoning, 110122, China

ARTICLE INFO

Keywords:
Grayscale medical image
Classification
Pixel
Text embedding

ABSTRACT

In our paper, we present an extension of text embedding architectures for grayscale medical image classification. We introduce a mechanism that combines n-gram features with an efficient pixel flattening technique to preserve spatial information during feature representation generation. Our approach involves flattening all pixels in grayscale medical images using a combination of column-wise, row-wise, diagonal-wise, and anti-diagonal-wise orders. This ensures that spatial dependencies are captured effectively in the feature representations. To evaluate the effectiveness of our method, we conducted a benchmark using 5 grayscale medical image datasets of varying sizes and complexities. 10-fold cross-validation showed that our approach achieved test accuracy score of 99.92 % on the Medical MNIST dataset, 90.06 % on the Chest X-ray Pneumonia dataset, 96.94 % on the Curated Covid CT dataset, 79.11 % on the MIAS dataset and 93.17 % on the Ultrasound dataset. The framework and reproducible code can be found on GitHub at https://github.com/xizhou/pixel_embedding.

1. Introduction

Convolutional neural networks (CNNs) remain the standard paradigm in computer vision for image classification tasks. However, practical deployment of CNN models [1] relying on specialized hardware, such as Compute Unified Device Architecture (CUDA) devices, may not be feasible for all users, especially those who use cloud servers without independent Graphics Processing Unit (GPU) devices. Text embeddings, which can map entities (e.g., words) to a high-dimensional space, are widely used in various natural language processing (NLP) tasks [2]. The main advantage of embedding models is their high performance in capturing semantic and syntactic relationships between entities while maintaining a low computational cost. Based on the studies by Astolfi and López-Monroy in this field [3,4], we hypothesized that stablished text embedding architectures like FastText [5] or StarSpace [6] could be adapted to capture the spatial relationship features in grayscale images. This involves developing a method to efficiently flatten image pixels, extending these frameworks to embed pixels of grayscale medical images for classification tasks.

In this study, we propose a pixel embedding approach for classifying grayscale medical images as an extension of current text embedding architectures. Our approach involves converting each pixel of the image into a token by combining a unique ID and pixel value. These tokens are then flattened and arranged in specific sequences: column-wise (c), row-wise (r), diagonal-wise (d), and anti-diagonal-wise (v) to generate a document-like text representation. The flattened representation of tokens can be used to train a

* Corresponding author.
** Corresponding author.
E-mail addresses: lu.wang0526@gmail.com (L. Wang), xbzhou@cmu.edu.cn (X. Zhou).

https://doi.org/10.1016/j.heliyon.2024.e36191
Received 24 March 2024; Received in revised form 12 August 2024; Accepted 12 August 2024
Available online 13 August 2024

customized text embedding model that employs an n-gram mechanism for predictive purposes. Based on the extensive pre-experiments, we found that by treating an image as a sequence and arranging the pixels in different orders, text embedding models with a long n-gram setting (e.g., 10) can successfully capture the spatial relationship of image pixels. Results from evaluating different grayscale medical image datasets, varying in size and difficulty, consistently demonstrate the high performance of our proposed method. These findings highlight the potential of leveraging text embedding models in grayscale image classification tasks.

The remainder of the paper is structured as follows: The "Related Works" section provides an overview of pertinent literature. The "Aim and Contributions of This Paper" section summarizes the purpose and the main contribution of this study. In the "Method" section, we introduce the proposed methodology, the model training settings and the datasets used for benchmarking and evaluation. We present the datasets used in this study in detail in the "Datasets" section, followed by the "Statistical Analysis" section. The "Results" and "Discussion" sections present and analyze the findings of our study. Finally, in the "Conclusion" section, we conclude the proposed method.

1.1. Related works

One of the challenges faced by NLP models in grayscale image classification tasks is their difficulty in capturing spatial relationship features of the image. This limitation arises due to the models' reliance on one-dimensional text input, while grayscale image data is inherently represented as a two-dimensional matrix [7]. To address this challenge, researchers have proposed several solutions in the field. You et al. introduced bidirectional vectors to capture contextual relationships among pixels in remote sensing images [8]. Astolfi et al. constructed a bipartite graph network from the image and used a sequence of alphabet symbols to represent different parts of the image [3]. Kulkarni et al. focused on the extraction of n-gram features in a single direction and utilized a trained neural network for the classification of regions of interest (ROIs) into normal and abnormal categories [9]. López-Monroy et al. deployed visual n-gram features in an SVM classifier for histopathology image classification [4]. In their study, a visual n-gram feature was defined as an $N \times N$ pixel patch, which represented a local region within the input images.

Radiological grayscale medical images, which are easily accessible in clinical settings, have the potential to aid diagnosis as a non-invasive method. This is particularly valuable in areas with limited medical resources where pathology and genetic testing are challenging [10,11]. Specifically, during the early stages of urgent public health crises such as SARS-CoV-2, when quantitative real-time polymerase chain reaction (qRT-PCR) tests were not widely available, the automated and accurate classification of pneumonia using radiological images provided critical diagnostic support. This helped clinicians identify high-risk individuals and slow the spread of the pandemic [12,13]. In the field of medical image classification, the emerging deep learning has necessitated larger datasets and increased computational resources, such as GPU devices, which has introduced practical implementation challenges. For instance, using datasets like Chest X-ray Pneumonia for COVID-19 image classification, researchers employed the MobileNet architecture with 6.67M parameters, achieving 98.15 % accuracy in a three-class classification task [14]. Another study integrated CNN and Graph Neural Network (GNN) for ten-class classification using 17 datasets including, Chest X-ray Pneumonia, achieving 99.22 % accuracy but encountering prolonged training times (averaging 359 s per epoch) [15]. Additionally, a study leveraged state-of-the-art Tensor Processing Unit (TPU) hardware to train multiple neural networks (ResNet-152-v2, Densenet-201, Efficientnet-B7, and Xception), and achieved 99.88 % accuracy in a three-class classification task on the Curated Covid CT dataset with an average processing speed of 9 ms per image [16]. Based on the public INbreast and MIAS datasets, a previous study using variational autoencoders for classifying normal, benign, and malignant images achieved an AUC of 0.99 with only 228 trainable parameters (with 204.95 K FLOPs) [17]. Another study based on BiLSTM, using approximately 1000M parameters on the MIAS dataset for a three-class classification, achieved 97.60 % accuracy [18]. Furthermore, a study showed that the previously proposed ResNet achieved 99.6 % accuracy in a six-class task on the medical MNIST dataset [19]. This performance is comparable to recent networks like MobileNetV2 and Inception V3, which achieved approximately 99 % accuracy on the medical MNIST dataset but with an increased parameter count of 23.9M [20].

1.2. Aim and Contributions of This Paper

Previous studies indicate that current research on medical image classification is limited by the hardware requirements (e.g., GPU) for training deep learning networks. There is an urgent need to overcome this dependence on hardware resources for processing medial image datasets. This paper aims to propose and evaluate an approach for classifying grayscale medical images using a pixel embedding method inherited by text embedding architectures. The primary objective is to demonstrate the effectiveness of employing text embedding models and n-grams to capture spatial relationships within images and improve classification accuracy.

The contributions of this paper are summarized as follows:

- Development of a pixel embedding approach: We introduce a method that converts image pixels into tokens and arranges them in specific orders to generate a document-like representation of the image. This approach could leverage existing text embedding models for image classification tasks.
- Utilization of n-grams for spatial relationship features: By employing n-grams, we aim to capture the spatial relationships between pixels within the image enhancing the classification accuracy and performance of the model.
- Development of highly customized mechanisms integrated within the pixel embedding model: We have implemented 3 specialized mechanisms to enhance computational efficiency and mitigate the impact of blank or irrelevant pixels in an image. These include:

(i) filtering to eliminate noise and focus on significant pixel data; (ii) sampling strategies that selectively reduce the volume of data processed; and (iii) removal of 0-variance columns to discard redundant information.

- Evaluation on grayscale medical image datasets: We conduct evaluations on multiple grayscale medical image datasets of varying sizes and complexities to demonstrate the efficacy of our proposed method. The consistent high performance across different datasets highlights the robustness and effectiveness of our pixel embedding model.

2. Method

This section provides a detailed description of the stages involved in the embedding model, from image processing to training the embedding classifier. Fig. 1 visually represents the core steps of the pixel embedding model.

2.1. Converting pixel to token

To use existing text embedding frameworks, grayscale image data in the form of a two-dimensional matrix needs to be flattened into a one-dimensional structure. In our model setting, the grayscale image is firstly reshaped into a 28 × 28 dimension. Each pixel of an image, X_{28*28}, is converted into a token by combining a unique ID and normalized pixel expression value as following:

$$Image : X_{28*28} \rightarrow Text : id_1 : w_1 | id_2 : w_2 | \ldots | id_{784} : w_{784}, \tag{1}$$

where $w = \frac{x - min(X)}{max(X) - min(X)}$, and $min(X)$ and $max(X)$ represent the maximum and minimum grayscale values in the image X_{28*28}.

2.2. N-gram features in different directions

To capture the spatial relationship between pixels, a mechanism that combines n-gram features with an efficient token flattening technique is designed. The token flattening mechanism categorizes the orders or directions of flattening into four types: column-wise (c), row-wise (r), diagonal-wise (d) and anti-diagonal-wise (v). A visual illustration of these four flattening orders is provided in Fig. 2 (a–d). In the default setting, all input tokens are flattened using a combination of the column-wise, row-wise, diagonal-wise and anti-diagonal-wise orders together (crdv). This means that the pixels are arranged in a way that captures their spatial relationships in all four directions simultaneously. An example of 3 × 3-pixel image patch is used to illustrate this combined flattening process as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

In the c order, the tokens will be flattened in the following sequence: [1,4,7,2,5,8,3,6,9]. In the r order, the tokens will be flattened in the following sequence: [1,2,3,4,5,6,7,8,9]. In the d order, the tokens will be flattened in the following sequence: [1,5,9,2,6,7,3,4,8]. In the v order, the tokens will be flattened in the following sequence: [3,5,7,2,4,9,6,8,1]. When all four orders are combined (crdv), the tokens will be flattened in the sequence: [1,4,7,2,5,8,3,6,9,1,2,3,4,5,6,7,8,9,1,5,9,2,6,7,3,4,8,3,5,7,2,4,9,6,8,1]. By applying this combined flattening mechanism, the model can efficiently capture the spatial relationships between pixels in different directions, allowing for a comprehensive understanding of the structure and context of the image. This, in conjunction with the integration of n-gram features, enhances the model's ability to capture fine-grained patterns and dependencies within the image.

In addition, we propose 3 mechanisms to improve computational efficiency or dampen the effects by blank or irrelevant pixels in an image. The filtering mechanism involves applying a filter to remove pixel values in the border regions of the image while preserving the center pixels. The selection of filter regions depends on the focus of classification tasks, which can be customized by the user based on the status of the input image. A filter with regions (i, j, k, l) means keeping the region of the pixel matrix from the row i to j and column k to l. One common filter region, such as (3,26,3,26), indicates that emphasis is placed on the center regions while ignoring the border regions. As depicted in Fig. 3 (a and c), the border regions of 2 pixels were removed, and the center pixels were retained from 3 pixels from the top to 26 pixels from the bottom, and from 3 pixels to the left to 26 pixels to the right. Specific filters could be selected

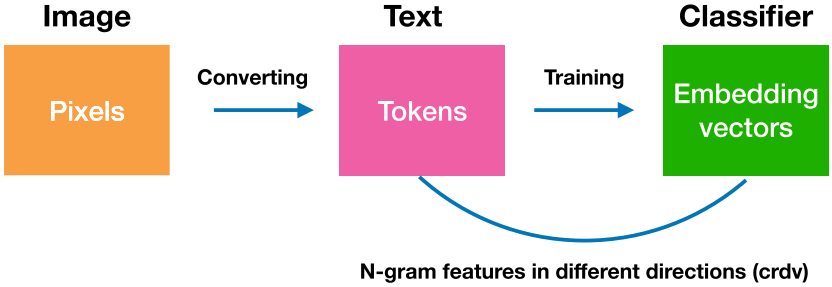


Fig. 1. The flowchart of the pixel embedding model.

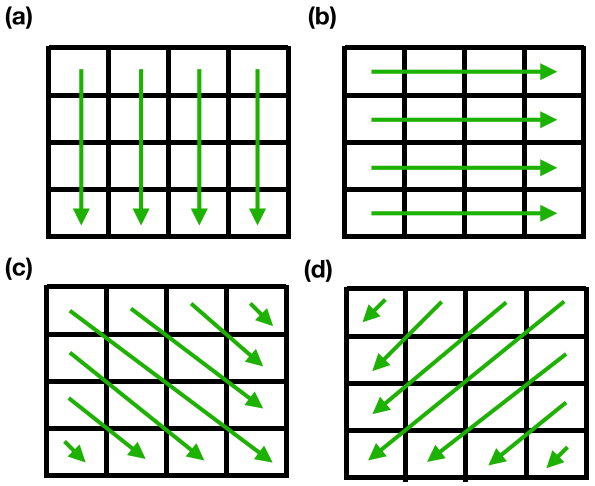


Fig. 2. Schematic illustration of four flatten orders. (a) Column-wise, (b) Row-wise, (c) Diagonal-wise and (d) Anti-diagonal-wise orders.

for specific purposes. For example, Fig. 3 (b) illustrated a filter (5,22,3,25) that focused only on the upper regions of the image. In this case (referring to the Chest X-ray Pneumonia dataset), the lower region of the image was found to be harmful to the task in the result section. The sampler mechanism involves randomly sampling a subset of pixels from the image. In the case of a 28 × 28 pixel image, 400 pixels are practically selected randomly. This sampling approach helps reduce the computational load by considering only a representative subset of pixels, rather than processing the entire image. While this may result in some loss of information, it can still capture the essential features of the image while significantly reducing computation time. Finally, the mechanism of removing 0-variance columns focuses on eliminating columns from the data matrix that contain pixel values with no variance. If a column has pixel values that are all the same, it is considered to have no meaningful information and is removed from further computations. This process helps eliminate redundant or uninformative pixel values, improving the efficiency and accuracy of subsequent analysis or processing tasks [21].

2.3. Text embedding model training

Our text embedding model for grayscale medical image classification is based on StarSpace [6], which utilizes a stochastic gradient descent (SGD) optimizer [22]. The model minimizes a loss function that measures the similarity between relevant entities (positive pairs of words and labels) and the dissimilarity between irrelevant entities (randomly selected irrelevant word and label pairs). The key parameters are as follows: size of embedding vectors [20 (default)], similarity measurement [dot (default)], loss function [hinge (default)], and length of word n-gram [10 (default)]. For more detailed information on training the model, including parameter settings, please refer to the project website for further details: https://github.com/xizhou/pixel_embedding.

3. Datasets

A total of 6 open-access medical image datasets were used for the model evaluation. **Medical MNIST dataset** [23] is a curated collection of grayscale medical images consisting of 58,954 medical images with dimensions of 64*64 pixels, categorized into 6 distinct classes. **Chest X-ray Pneumonia dataset** [24] includes chest X-ray images obtained from various healthcare facilities,

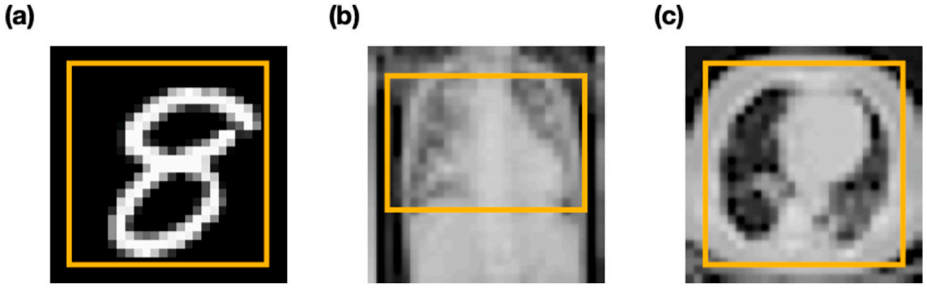


Fig. 3. Schematic illustration of three filters applied to grayscale images. (a) A filter is used to remove border regions of 2 pixels, while keeping the center from 3 pixels from the top to 26 pixels from the bottom, and from 3 pixels to the left to 26 pixels to the right, denoted as (3,26,3,26). (b) A filter with dimensions (5,22,3,25) is used, and (c) A filter with dimensions (3,26,3,26) is used.

including hospitals and medical institutions. The dataset comprises a total of 5863 images, categorized into two classes: pneumonia and normal. **Curated Covid CT dataset** [16] is a collection of lung CT scan images for COVID-19 from 7 public datasets containing 7593 COVID-19 cases, 2618 viral pneumonia cases, and 6893 normal cases. **MIAS dataset** contains images of 322 mammography scans [25]. **Ultrasound dataset** includes 360 ultrasound images of 6 abdominal organs, including the bladder, bowel, gallbladder, kidney, liver, and spleen [26]. **MNIST** dataset is a collection of 70,000 grayscale images of handwritten digits (0–9), with each image sized at 28 × 28 pixels [27].

4. Statistical analysis

The training and test datasets were divided with a ratio of 80:20 for the Medical MNIST, Curated Covid CT, and MIAS datasets. For the Chest X-ray Pneumonia dataset, we use the training set (3875 pneumonia and 1431 normal images) and test set (390 pneumonia and 234 normal images) that have been divided by previous researchers. In the ultrasound dataset, 300 images are used for training, consisting of 50 images per organ. The test set comprises 60 images, with 10 images per organ. MNIST dataset contains 60,000 images for training and 10,000 for testing. Ten-fold cross-validation was employed, and several metrics—including precision, recall, sensitivity, specificity, F1-score, and accuracy were used to evaluate the performance of the models across various datasets. Specifically, evaluations included 6-category classification on the Medical MNIST dataset, 3-category classification on the Curated Covid CT dataset, 6-category classification on the ultrasound dataset, 10-category classification on the MNIST dataset, and binary classification on 2 additional datasets. **The MNIST dataset is primarily used to evaluate the performance of different pixel embedding mechanisms rather than for benchmark comparisons.** The results reported in this study were the average of ten-fold cross-validation on the test set.

5. Result

The performance of the pixel embedding model was evaluated across various processing procedures, parameters, and datasets. Table 1 presents the accuracy performance of the proposed method. This method incorporates different settings for filtering (f), sampling (s), and the removal of 0-variance columns (0), alongside various n-gram (g) configurations, while maintaining the flattening methods (crdv) as fixed settings. Increasing the number of n-grams dramatically improved model performance, with configurations 10-g (10 g) and 12-g (12 g) achieving the highest accuracy rates. While filtering and sampling strategies tended to reduce model performance, they could be effectively when combined with n-grams. Notably, the combination of filtering with n-grams (f (5,22,3,25)× 10 g) achieved the best performance on the Chest X-ray Pneumonia dataset. Based on these findings, the "10-g (10 g)" has been set as the default in our text embedding model for downstream benchmark tests. The marginal benefits of using a larger number of n-grams, such as 12-g (12 g), are relatively small compared to the increased computational cost. However, a filtering strategy could be applied depending on the particular dataset, such as the Chest X-ray Pneumonia dataset, to optimize performance and resource usage. This strategy would involve adjusting the filter size based on the dataset's characteristics and the specific requirements of the task.

For the MNIST dataset, 10 g was used as the general setting. 4 different flattening methods (c, r, d, and v) and their combinations were evaluated, along with various filters (f), removal of 0-variance columns (0), and sampling of 400 pixels (k). The combinations tested included "crdv", "crdv0", "crdvf", "crdvk", "crdvkf", and "crdvk0". All combined results were shown in Fig. 4. It was evident that the combination of the 4 flattening methods (crdv) resulted in a significant improvement in accuracy compared to other individual flattening approaches (c, cr, and crd). Specifically, the inclusion of a filtering approach (crdvf) achieved an accuracy of 97.74 %, which

Table 1
Comprehensive benchmarks of the pixel embedding method incorporating various settings of mechanisms: filtering (f), sampling (k), and removal of 0-variance columns (0), alongside different n-gram (g) configurations.

Embedding methods	Medical MNIST	Curated Covid CT	Chest X-ray Pneumonia
none	98.76 (0.02)	86.51 (0.6)	86.15 (0.82)
0	95.47 (0.28)	86.33 (0.72)	85.95 (1.42)
f (2,27,2,27)	98.70 (0.04)	85.22 (1.18)	85.1 (1.85)
f (3,26,3,26)	98.22 (0.06)	83.59 (0.42)	84.44 (1.29)
f (5,22,3,25)	97.33 (0.06)	74.65 (0.46)	85.21 (1.01)
k (100)	98.08 (0.07)	77.33 (0.86)	85.38 (1.06)
k (200)	98.42 (0.07)	79.73 (0.57)	86.2 (0.63)
k (400)	98.71 (0.03)	83.49 (1.74)	85.45 (1.16)
2 g	99.91 (0.01)	96.61 (0.31)	85.45 (1.16)
5 g	99.93 (0.02)	97.28 (0.1)	89.38 (0.56)
10 g	99.92 (0.01)	96.94 (0.18)	90.06 (0.46)
12 g	99.90 (0.01)	96.79 (0.19)	90.38 (0.39)
0 × 10 g	99.88 (0.01)	96.95 (0.11)	90.02 (0.43)
f (2,27,2,27)×10 g	99.92 (0.01)	96.83 (0.09)	88.89 (0.53)
f (3,26,3,26)×10 g	99.93 (0.01)	96.45 (0.15)	88.21 (0.55)
f (5,22,3,25)×10 g	99.89 (0.01)	93.26 (0.12)	91.41 (0.24)
k (100)×10 g	99.89 (0.01)	96.36 (0.18)	88.78 (0.37)
k (200)×10 g	99.91 (0.01)	96.90 (0.09)	88.62 (0.54)
k (400)×10 g	99.90 (0.01)	96.80 (0.15)	89.10 (0.36)

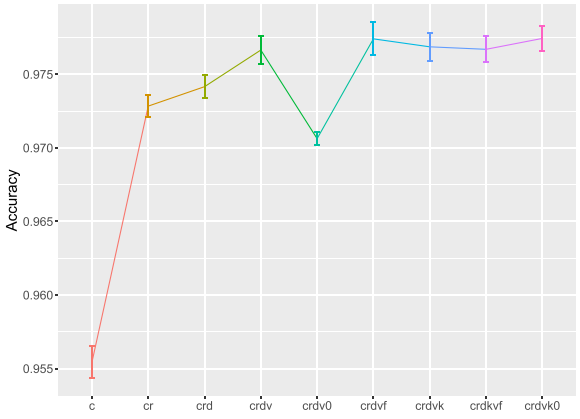


Fig. 4. Performance of various processing procedures on the MNIST dataset for image classification. These procedures included different flattening methods (c, r, d and v), filters (f), removal of 0-variance columns (0), and sampling processes (k). The accuracy scores of different combinations of these procedures were presented.

was the highest among the methods evaluated. Incorporating a sampler or combination of a filter and sampler (crdvk and crdvkf) led to slight decreases in accuracy by 0.05 % and 0.07 %, respectively. The removal of 0-variance columns (crdv0) resulted in a sharp decrease in accuracy of 0.68 %. However, this accuracy was restored to 97.74 % when the sampler (crdvk0) was applied in combination with the remaining features.

Table 2 presents the benchmark comparison results for the 5 datasets based on the general settings (crdv and 10 g). For the Medical MNIST dataset, text embedding model achieved an accuracy of 99.92 % with a standard deviation of 0.01. This is comparable to the mobileNetV2-based deep learning network (99.13 %) using GPU devices [20]. In the case of the Chest X-ray Pneumonia dataset, the embedding model achieved an accuracy of 90.06 % with a standard deviation of 0.46. This is higher than the accuracies of 87.28 %, 88.46 %, 77.56 % and 70.99 % of VGG16, VGG19, ResNet50 and Inception-v3 networks [28]. For the Curated Covid CT dataset, the embedding model achieved an accuracy of 96.94 % with a standard deviation of 0.18. This surpassed the accuracy of 95.31 % (median) reported in the research paper by Thomas et al. [16]. Our model achieved an accuracy of 79.11 % with a standard deviation of 1.21 on the MIAS dataset. The accuracy is inferior to the accuracy of 81.90 % reported in previous study for benign-malignant classifications on MIAS [17]. For the Ultrasound dataset, the model's accuracy of 93.17 % was lower than that achieved in the previous study [26]. Additionally, more detailed results including precision, re-call, sensitivity, specificity, and F1-score of the text embedding method are reported in Table 3, which demonstrate the effectiveness of the pixel embedding method in classifying medical images, with high accuracy and reliability.

Additionally, the detailed training information for various datasets across the proposed method and other reported methods were compared. To facilitate a clear comparison with previous studies, we have included information reported by the authors of previous studies using the four datasets mentioned in this study. Training speed, parameters, and hardware are summarized in Table 4. From the table, we observed that GPUs were commonly used across most methods, although TPU and CPU + GPU setups were also employed based on specific requirements and datasets. Parameter sizes varied significantly, from 228 in Karagoz et al. [17] to as high as 37.9 billion in Raiaan et al. [29] for the MIAS dataset. The pixel embedding method generally maintains a consistent parameter size of around 16.5 K but exhibits variability in processing time depending on the dataset on CPU usage. The parameter calculation formula for the pixel embedding method, using 10-g and 20 dimensions, is: $28 \times 28 \times 20 + 28 \times 28 + 1 + 20xN$, where N represents the number of labels. In this study, the image size was normalized to 28×28 pixels. The parameters of the pixel embedding method across different datasets depend on the number of labels, leading to minimal variation in outcomes. The computational time varies primarily based on the number of training datasets used.

Table 2
Benchmark comparison with the original results mentioned in the research corresponding to the medical image datasets. We report mean (%) and standard deviation of the accuracies, averaged over ten runs. Note that the parameter settings for the pixel embedding method are: crdv and 10 g.

Dataset	Pixel embedding	Previous studies
Medical MNIST	99.92 (0.01)	99.13 (Dash et al. [20])
Chest X-ray Pneumonia	90.06 (0.46)	87.28 (Jain et al. [28])
Curated Covid CT	96.94 (0.18)	95.31 (Thomas et al. [16])
MIAS	79.11 (1.21)	81.90 (Karagoz et al. [17])
Ultrasound	93.17 (1.83)	96.67 (Li et al. [26])

Table 3
The precision, re-call, sensitivity, specificity, and F1-score of the proposed embedding method in the datasets. We report mean (%) with standard deviation of the accuracies, averaged over ten runs. Note that the parameter settings for the pixel embedding method used in the proposed study are: crdv and 10-g.

Dataset	Precision	Re-call	Sensitivity	Specificity	F1-score
Medical MNIST	99.92 (0.01)	99.92 (0.01)	99.92 (0.01)	99.98 (0.00)	99.92 (0.01)
Chest X-ray Pneumonia	94.8 (0.4)	77.78 (1.47)	77.78 (1.47)	97.44 (0.24)	85.44 (0.8)
Curated Covid CT	97.47 (0.14)	97.4 (0.17)	97.4 (0.17)	98.26 (0.1)	97.43 (0.15)
MIAS	53.97 (2.16)	69.23 (0.00)	69.23 (0.00)	82.09 (1.57)	60.63 (1.37)
Ultrasound	94.17 (1.57)	93.17 (1.83)	93.17 (1.83)	98.63 (0.37)	93.02 (1.98)

Table 4
Training detailed information. Details across the proposed method and previous studies, including aspects such as training speed, number of parameters, and hardware used. Note: An entry marked as "Unknown" indicates that the information was not reported in the original papers.

Method	Dataset	Parameters	Time	Hardware
Asif et al. [14]	Chest X-ray Pneumonia	6.67M	29s for inference	GPU
Kaya et al. [30]	Chest X-ray Pneumonia	4.75M	Unknown	GPU
Rana et al. [15]	Chest X-ray Pneumonia	Unknown	359s/epoch	CPU + GPU
Pixel embedding	Chest X-ray Pneumonia	16.5 K	2.25s/epoch (10 cores)	CPU
Thomas et al. [16]	Curated Covid CT	1.15B–1.93B	6.71–11.34 ms/image	TPU
Pixel embedding	Curated Covid CT	16.5 K	8.01s/epoch (10 cores)	CPU
Aslan et al. [18]	MIAS	Unknown	Unknown	GPU
Raiaan et al. [29]	MIAS	37.9B	3–5s/epoch	GPU
Karagoz et al. [17]	MIAS	228	Unknown	GPU
Pixel embedding	MIAS	16.5 K	0.80s/epoch (10 cores)	CPU
Hassan et al. [29]	Medical MNIST	Unknown	Unknown	GPU
Dash et al. [20]	Medical MNIST	1.24M–23.9M	Unknown	GPU
Pixel embedding	Medical MNIST	16.5 K	24.83s/epoch (10 cores)	CPU

6. Discussion

In this study, we proposed a pixel embedding approach for classifying grayscale medical images by converting the image information to token-based document-like text representation. By employing the n-gram features in different directions, we aim to capture the spatial relationship features of the image and train a text embedding model for image classification. Compared to the traditional vision transformer models which divides the input image into patches and treats each patch as a token within the transformer framework, our approach is superior by employing a text embedding framework that represents each pixel of an image as a high-dimensional vector. Results from evaluating multiple grayscale medical image datasets, varying in size and difficulty, consistently demonstrate the high performance of the proposed method.

The image size of the dataset used in this study was normalized to 28×28. Resizing all the datasets to 28×28 is in line with the general research procedure of using these data sets, while ensuring the consistency of the data analysis in this study. In addition, our results show that we have achieved good accuracy on the distinct data sets after using the normalized image size. As stated in the manuscript, when using the smaller image size, it obtained a classification of 99.92 % on the Medical MNIST dataset, and regrading the Chest X-ray Pneumonia dataset and Curated Covid CT datasets, 90.06 % and 96.94 % were obtained for pneumonia classification. According to the findings of this study, we found that while the utilization of smaller-sized images like 28×28 may not effectively discern subtle disease disparities through visual observation, the pixel embedding approach proposed in this study can effectively detect imperceptible distinctions that elude the naked eye, leading to commendable predictive accuracy. Results showed that this method will improve the accuracy and efficiency of future medical data classification in the field of image recognition as a way to fill in the image information to text embedding.

The current deep learning approaches based on convolutional neural networks (CNN) involve direct pixel convolution on the image, and these methods has demonstrated promising results in image classification tasks using the image-based high-throughput data computation [31]. However, it inevitably relies on corresponding CUDA-based hardware resources, incurring substantial training costs and time consumption [32]. In contrast, the method proposed in this study captures the spatial relationship features of the image by treating it as a sequence and rearranging the pixels (i.e., tokens) in different orders. Subsequently, text embedding models with an extended n-gram configuration are utilized for feature extraction. Comparing with previously proposed deep learning analysis methods [33,34] on Medical MNIST dataset, Chest X-ray Pneumonia dataset, and Curated Covid CT dataset, this text embedding-based approach surpasses the traditional CNN-based image classification method. It no longer solely treats the image as the processing object but converts the image information into text, thereby reducing the computational workload and enhancing computational efficiency.

Our study demonstrated the importance of integrating multi-directional n-gram features in NLP-based image classification solutions. By considering pixel arrangements in column-wise, row-wise, diagonal-wise, and anti-diagonal-wise orders, we effectively capture different orientations and spatial dependencies present in the images. As a result, our approach demonstrates superior performance compared to Kulkarni's solution [9,35], which only considers n-gram features in a single direction. The notable increase in

classification accuracy from 95.62 % to 97.68 % further validates the effectiveness of our approach (as illustrated in Fig. 4 ("c" and "crdv")). Comparing the visual n-gram approach [4], our strategy is more efficient in terms of computation. For the example, as mentioned in the method section, when considering pixel 5 as the center, the visual 2-g method would extract all the pairs as [5-1,5-9, 5-3,5-7,5-2,5-8,5-4,5-6], resulting in 40 tokens in one image. However, our token flattening method only requires 36 tokens, represented as [1,4,7,2,5,8,3,6,9,1,2,3,4,5,6,7,8,9,1,5,9,2,6,7,3,4,8,3,5,7,2,4,9,6,8,1], to capture all spatial relations. Moreover, as the image size increases, the computational cost of incorporating multi-directional n-gram features is dramatically reduced compared to the visual n-gram approach. Furthermore, compared with the vision transformer architecture, which divides the input image into patches and treats each patch as a token within the transformer framework, our approach is essentially a text embedding framework that represents each pixel of an image as a high-dimensional vector and therefore it showed efficiency in image information recognition.

7. Limitations

This study has several limitations, and further efforts should be made in future research. First, the n-gram feature approach still struggles to effectively capture complete contextual information and dependencies between neighboring pixels, especially when compared to traditional CNN architectures. Future research should focus on finding methods that can better preserve the spatial relationships and local context in image data while leveraging the advantages of n-gram features. Second, it is important to consider the increased computational complexity associated with converting each pixel into a token and processing them as text embeddings, particularly for high-resolution image datasets or datasets with a large sample size. This conversion and subsequent training times may be longer, making real-time applications challenging or resource-intensive. Although the current solution of resizing all datasets to 28 $\times$ 28 aligns with the general research procedure, it may not be optimal for capturing fine-grained details in high-resolution images. Future studies should explore solutions to optimize the computational efficiency of n-gram feature approach to reduce computational demands and make the text embedding model more feasible for real-time applications with limited computational resources.

8. Conclusion

This study presented an innovative approach for classifying grayscale medical images using text embedding architectures. The findings of this research could provide benefits for traditional computing resources that are involved in grayscale image classification tasks, as they can utilize this approach without the requirement of independent CUDA devices.

Funding

This research project was funded by the Liaoning Provincial Natural Science Foundation (2021-MS-194) and 111 project (D21008).

Data availability statement

The data used in this manuscript can be accessed at the following links:

- Medical MNIST dataset:
  https://www.kaggle.com/datasets/andrewmvd/medical-mnist
- Chest X-ray Pneumonia dataset:
  https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
- Curated COVID CT dataset:
  https://www.kaggle.com/datasets/mehradaria/covid19-lung-ct-scans
- MIAS dataset:
  http://peipa.essex.ac.uk/info/mias.html
- Ultrasound dataset:
  https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset/data

The framework and the reproducible codes are available on GitHub at https://github.com/xizhou/pixel_embedding.

CRediT authorship contribution statement

Wensu Liu: Methodology, Formal analysis, Conceptualization. **Na Lv:** Writing – review & editing. **Jing Wan:** Writing – review & editing. **Lu Wang:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Data curation. **Xiaobei Zhou:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Zhang, J. Li, Improving the Performance of OpenCL-based FPGA Accelerator for Convolutional Neural Network , n.d. https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/l.

[2] Z. Chen, Z. He, X. Liu, J. Bian, Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases, BMC Med. Inf. Decis. Making 18 (2018) 65, https://doi.org/10.1186/s12911-018-0630-x.

[3] G. Astolfi, D.A. Sant'Ana, J.V. de A. Porto, F.P.C. Rezende, E.C. Tetila, E.T. Matsubara, H. Pistori, An approach for applying natural language processing to image classification problems, Neurocomputing 513 (2022), https://doi.org/10.1016/j.neucom.2022.09.131.

[4] A.P. López-Monroy, M. Montes-y-Gómez, H.J. Escalante, A. Cruz-Roa, F.A. González, Bag-of-visual-ngrams for histopathology image classification, in: IX International Seminar on Medical Information Processing and Analysis, 2013, https://doi.org/10.1117/12.2034113.

[5] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2017, https://doi.org/10.18653/v1/e17-2068.

[6] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston, StarSpace: embed all the things, in: 32nd AAAI Conference on Artificial Intelligence, 2018, https://doi.org/10.1609/aaai.v32i1.11996. AAAI 2018.

[7] L. Wang, N. Xu, J. Song, Decoding intra-tumoral spatial heterogeneity on radiological images using the Hilbert curve, Insights Imaging 12 (2021) 154, https://doi.org/10.1186/s13244-021-01100-8.

[8] H. You, S. Tian, L. Yu, Y. Lv, Pixel-level remote sensing image recognition based on bidirectional word vectors, IEEE Trans. Geosci. Rem. Sens. 58 (2020), https://doi.org/10.1109/TGRS.2019.2945591.

[9] P. Kulkarni, A. Stranieri, S. Kulkarni, J. Ugon, M. Mittal, Hybrid technique based on N-GRAM and neural networks for classification of mammographic images, in: Second International Conference on Signal, Image Processing and Pattern Recognition, 2014, https://doi.org/10.5121/csit.2014.4225.

[10] S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, M. Zhou, Y. Liu, O. Gevaert, K. Wang, Y. Zhu, H. Zhou, Z. Liu, J. Tian, Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning, Eur. Respir. J. 53 (2019) 1800986, https://doi.org/10.1183/13993003.00986-2018.

[11] V. Turbé, C. Herbst, T. Mngomezulu, S. Meshkinfamfard, N. Dlamini, T. Mhlongo, T. Smit, V. Cherepanova, K. Shimada, J. Budd, N. Arsenov, S. Gray, D. Pillay, K. Herbst, M. Shahmanesh, R.A. McKendry, Deep learning of HIV field-based rapid tests, Nat. Med. 27 (2021) 1165–1170, https://doi.org/10.1038/s41591-021-01384-9.

[12] S.A. Harmon, T.H. Sanford, S. Xu, E.B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, M. Blain, M. Kassin, D. Long, N. Varble, S. M. Walker, U. Bagci, A.M. Ierardi, E. Stellato, G.G. Plensich, G. Franceschelli, C. Girlando, G. Irmici, D. Labella, D. Hammoud, A. Malayeri, E. Jones, R. M. Summers, P.L. Choyke, D. Xu, M. Flores, K. Tamura, H. Obinata, H. Mori, F. Patella, M. Cariati, G. Carrafiello, P. An, B.J. Wood, B. Turkbey, Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets, Nat. Commun. 11 (2020) 4080, https://doi.org/10.1038/s41467-020-17971-2.

[13] Z. Feng, Q. Yu, S. Yao, L. Luo, W. Zhou, X. Mao, J. Li, J. Duan, Z. Yan, M. Yang, H. Tan, M. Ma, T. Li, D. Yi, Z. Mi, H. Zhao, Y. Jiang, Z. He, H. Li, W. Nie, Y. Liu, J. Zhao, M. Luo, X. Liu, P. Rong, W. Wang, Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics, Nat. Commun. 11 (2020) 4968, https://doi.org/10.1038/s41467-020-18786-x.

[14] S. Asif, Q. Ain, R. Al-Sabri, M. Abdullah, LitefusionNet: boosting the performance for medical image classification with an intelligent and lightweight feature fusion network, J. Comput. Sci. 80 (2024) 102324, https://doi.org/10.1016/j.jocs.2024.102324.

[15] S. Rana, M.J. Hosen, T.J. Tonni, MdA.H. Rony, K. Fatema, MdZ. Hasan, MdT. Rahman, R.T. Khan, T. Jan, M. Whaiduzzaman, DeepChestGNN: a comprehensive framework for enhanced lung disease identification through advanced graphical deep features, Sensors 24 (2024) 2830, https://doi.org/10.3390/s24092830.

[16] J.B. Thomas, S.K. V, S.M. Sulthan, A. Al-Jumaily, Deep feature meta-learners ensemble models for COVID-19 CT scan classification, Electronics (Basel) 12 (2023) 684, https://doi.org/10.3390/electronics12030684.

[17] M.A. Karagoz, O.U. Nalbantoglu, A self-supervised learning model based on variational autoencoder for limited-sample mammogram classification, Appl. Intell. 54 (2024) 3448–3463, https://doi.org/10.1007/s10489-024-05358-5.

[18] M.F. Aslan, A hybrid end-to-end learning approach for breast cancer diagnosis: convolutional recurrent network, Comput. Electr. Eng. 105 (2023) 108562, https://doi.org/10.1016/j.compeleceng.2022.108562.

[19] E. Hassan, M.S. Hossain, A. Saber, S. Elmougy, A. Ghoneim, G. Muhammad, A quantum convolutional network and ResNet (50)-based classification architecture for the MNIST medical dataset, Biomed. Signal Process Control 87 (2024) 105560, https://doi.org/10.1016/j.bspc.2023.105560.

[20] S. Dash, P. Parida, J.R. Mohanty, Illumination robust deep convolutional neural network for medical image classification, Soft Comput. (2023), https://doi.org/10.1007/s00500-023-07918-2.

[21] M. Shahlaei, A. Fassihi, L. Saghaie, E. Arkan, A. Madadkar-Sobhani, A. Pourhossein, Computational evaluation of some indenopyrazole derivatives as anticancer compounds; application of QSAR and docking methodologies, J. Enzym. Inhib. Med. Chem. 28 (2013), https://doi.org/10.3109/14756366.2011.618991.

[22] N. Ketkar, Stochastic gradient descent bt - deep learning with Python: a hands-on introduction, in: Deep Learning with Python, 2017.

[23] S. Kus, Medical MNIST, vol. 1, Mendeley Data, 2022, https://doi.org/10.17632/8hdt269s7r.1.

[24] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A.N. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (2018), https://doi.org/10.1016/j.cell.2018.02.010.

[25] J. Suckling, C.R.M. Boggis, I. Hutt, S. Astley, D. Betal, N. Cerneaz, N. Karrsemeijer, A. Clark, The mini-MIAS database of mammograms, the mammographic image analysis society digital mammogram database exerpta medica, Int. Congr. 1069 (1994).

[26] K. Li, Y. Xu, M.Q.H. Meng, Automatic recognition of abdominal organs in ultrasound images based on deep neural networks and K-Nearest-Neighbor classification. 2021 IEEE International Conference on Robotics and Biomimetics, ROBIO, 2001, https://doi.org/10.1109/ROBIO54168.2021.9739348.

[27] D. Li, The mnist database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[28] R. Jain, P. Nagrath, G. Kataria, V. Sirish Kaushik, D. Jude Hemanth, Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning, Measurement 165 (2020) 108046, https://doi.org/10.1016/j.measurement.2020.108046.

[29] M.A.K. Raiaan, N.M. Fahad, M.S.H. Mukta, S. Shatabda, Mammo-Light: a lightweight convolutional neural network for diagnosing breast cancer from mammography images, Biomed. Signal Process Control 94 (2024) 106279, https://doi.org/10.1016/J.BSPC.2024.106279.

[30] M. Kaya, Y. Çetin-Kaya, A novel ensemble learning framework based on a genetic algorithm for the classification of pneumonia, Eng. Appl. Artif. Intell. 133 (2024), https://doi.org/10.1016/J.ENGAPPAI.2024.108494.

[31] L. Wang, H. Wang, Y. Huang, B. Yan, Z. Chang, Z. Liu, M. Zhao, L. Cui, J. Song, F. Li, Trends in the application of deep learning networks in medical image analysis: evolution between 2012 and 2020, Eur. J. Radiol. 146 (2022) 110069, https://doi.org/10.1016/j.ejrad.2021.110069.

[32] G.T. Tufa, F.A. Andargie, A. Bijalwan, Acceleration of deep neural network training using field programmable gate arrays, Comput. Intell. Neurosci. 2022 (2022) 1–11, https://doi.org/10.1155/2022/8387364.

[33] N.A. Mohammed, M.H. Abed, A.T. Albu-Salih, Convolutional neural network for color images classification, Bulletin of Electrical Engineering and Informatics 11 (2022), https://doi.org/10.11591/eei.v11i3.3730.

[34] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, Y. Miao, Review of image classification algorithms based on convolutional neural networks, Remote Sens (Basel) 13 (2021), https://doi.org/10.3390/rs13224712.

[35] P. Kulkarni, A. Stranieri, S. Kulkarni, J. Ugon, M. Mittal, Visual character N-grams for classification and retrieval of radiological images, Int. J. Multimed. Appl. 6 (2014), https://doi.org/10.5121/ijma.2014.6204.