Jurnal Kejuruteraan 36(2) 2024: 653–660
https://doi.org/10.17576/jkukm-2024-36(2)-24

Chinese Character Recognition Using Non-negative Matrix Factorization

Chen Huey Voon*, Tang Ker Shin & Ng Wei Shean

Department of Mathematical & Actuarial Sciences, Lee Kong Chian Faculty of Engineering & Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, 43000 Kajang, Selangor, Malaysia

**Corresponding author: chenhv@utar.edu.my*

Received 20 January 2023, Received in revised form 15 January 2024
Accepted 23 February 2024, Available online 30 March 2024

ABSTRACT

Non-negative matrix factorization (NMF) was introduced by Paatero and Tapper in 1994 and it was a general way of reducing the dimension of the matrix with non-negative entries. Non-negative matrix factorization is very useful in many data analysis applications such as character recognition, text mining, and others. This paper aims to study the application in Chinese character recognition using non-negative matrix factorization. Python was used to carry out the LU factorization and non-negative matrix factorization of a Chinese character in Boolean Matrix. Preliminary analysis confirmed that the data size of and and are chosen for the NMF of the Boolean matrix. In this project, one hundred printed Chinese characters were selected, and all the Chinese characters can be categorized into ten categories according to the number of strokes , for . The Euclidean distance between the Boolean matrix of a Chinese character and the matrix after both LU factorization and NMF is calculated for further analysis. Paired t-test confirmed that the factorization of Chinese characters in the Boolean matrix using NMF is better than the LU factorization. Finally, ten handwritten Chinese characters were selected to test whether the program is able to identify the handwritten and the printed Chinese characters. Experimental results showed that 70% of the characters can be recognized via the least Euclidean distance obtained. NMF is suitable to be applied in Chinese character recognition since it can reduce the dimension of the image and the error between the original Boolean matrix and after NMF is less than 5%.

Keywords: Chinese characters recognitions; matrix factorizations; non-negative matrix factorization

INTRODUCTION

To address complex data analysis and representation challenges, machine learning, deep learning, and image processing are employed. Matrix factorization, a foundational mathematical technique that separates a matrix into smaller dimensional matrices, serves as a versatile tool for studying inherent characteristics within matrices (Handschutter et al. 2021). This method, encompassing various types such as LU factorization, QR factorization, and nonnegative matrix factorization (NMF) (Hogben 2013), establishes a crucial link between matrix factorization and character recognition in the realms of image processing, computer vision, and artificial intelligence.

Recent research endeavours, such as Zamani et al. (2023), exemplify the convergence of deep learning in automated image-based stem cell classification. Milad et al. (2023) bridges machine learning techniques for predicting the rheological properties of bitumen-filler mastic, while Alayat and Omar (2023) utilizes digital image processing techniques for pavement surface distress detection. These collective efforts underscore the versatility of advanced methodologies, emphasizing their collaborative potential in diverse applications, from character recognition and stem cell classification to materials science and infrastructure monitoring.

In this paper, the focus shifts to Chinese character recognition due to the complexity and variations inherent in Chinese characters. The application of LU factorization and NMF in Chinese character recognition is explored, comparing the performances of these two matrix factorization techniques.

Character recognition involves recognizing characters using programming algorithms, with characters represented as matrices, highlighting the close connection between character recognition and matrix factorization. Extensive research in Chinese character recognition utilizing matrix

factorization includes Shuai's application of nonnegative matrix factorization (NMF) for radical detection (Shuai, 2016). Lin et al. (2019) proposed a matrix factorization method to decompose Chinese characters for applications in secret sharing. Additionally, Tan et al. (2012) introduced affine sparse matrix factorization (ASMF) as an innovative approach to extracting radicals from Chinese characters.

LU FACTORIZATION

LU factorization is to decompose a given matrix into two matrices which are a lower triangular matrix with a diagonal of one, , and an upper triangular matrix, , i.e. or

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}. \quad (1)$$

In 1938, Tadeusz Banachiewicz introduced LU factorization method (Schwarzenberg-Czerny, 1995). In 2014, Sudipto & Anindya (2014) proved that an invertible matrix A can be factorized into LU if and only if all its leading principal submatrices are invertible and the LU factorization of A is unique. LU factorization has applications such as finding a solution for a system of equations, inverse matrix, determinant of the matrix, etc. This method is useful in dealing with the problem from matrix form. In LU factorization, negative elements are allowed as entries in the matrices L and U (Bernard and David, 2001). A detailed explanation of LU factorization can be found in (Golub & Van Loan, 2022)

NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) is another mathematical process that can be used to find patterns in data. It is also referred to as non-negative matrix decomposition or non-negative least squares. NMF has been applied in many fields such as finance, text mining and computer vision.

Given an m×n matrix V with non-negative elements, it can be decomposed into two matrices W and H with the dimension of $m \times p$ and respectively, such that is approximately equal to WH (as shown in Figure 1), i.e.,

$$V \approx WH. \quad (2)$$

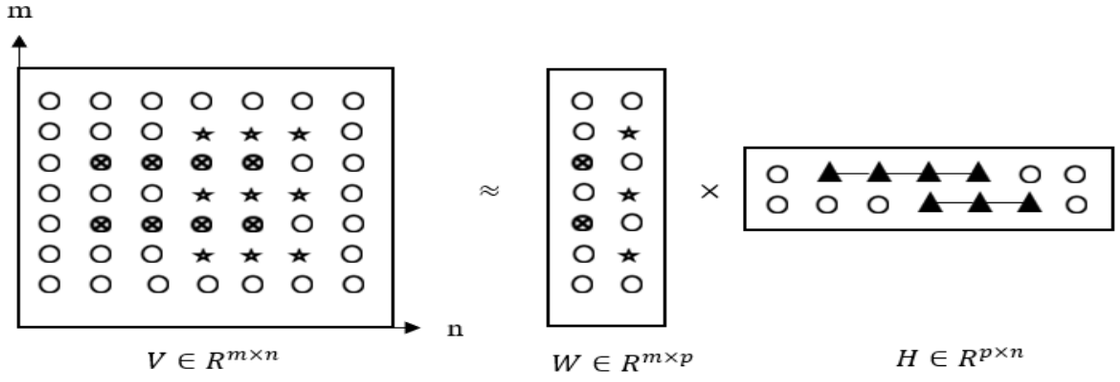


FIGURE 1. General formulation for Non-negative Matrix Factorization

NMF can alter the starting value of matrices W and H to minimize the difference between V and WH until achieving the maximum iterations using the equations below:

$$\left[H^{(k+1)}\right]_{ij} \leftarrow \left[H^{(k)}\right]_{ij} \frac{\left[(W^{(k)})^T X\right]_{ij}}{\left[(W^{(k)})^T W^{(k)} H^{(k)}\right]_{ij}}; \quad (3)$$

$$\left[W^{(k+1)}\right]_{ij} \leftarrow \left[W^{(k)}\right]_{ij} \frac{\left[X(H^{(k+1)})^T\right]_{ij}}{\left[(W^{(k)})H^{(k+1)}(H^{(k+1)})^T\right]_{ij}}, \quad (4)$$

where i, j denote the matrix indices and k is the iteration number. The Euclidean distance between V and WH is $\sqrt{\|V - WH\|^2}$ (Lee and Seung 2000).

In 1994, NMF was introduced by Paatero and Tapper (1994) and it was then popularized by Lee and Seung in 1999. Lee and Seung have shown how an NMF algorithm was used to learn parts of faces and text semantic characteristics. They stated that the use of non-negative constraints distinguishes NMF from other approaches (Lee and Seung 1999). NMF is a way of reducing the dimension, the final matrices after decomposition have fewer elements compared to the original matrix. From this finding, a new method to optimize the rank which is NMF has been introduced (Lin and Boutros 2020). NMF algorithms are a set of algorithms that map data into a low-dimensional representation which can be classified into few categories such as Basic NMF, structured NMF, constrained NMF and generalized NMF according to problem involved and design principles (Wang and Zhang 2013).

NMF is dealing with non-negative elements in the matrices and is elementary when carried out. NMF also becomes a popular method in analyzing data with high dimensions because it automatically brings out sparse and important features from the matrix with all non-negative elements (Gan et al. 2021). According to Langville et al (2014), good initialization could enhance the speed and accuracy of the result of the NMF algorithm. Pauca et al (2004) found that NMF is very useful in text mining, and this technique is applying the method of vector space with non-negative elements to get data representation.

For the application in character recognition for NMF, the input data of Chinese characters have a high dimension and sparse data which consists of a lot of zero values. Text mining allows the process of converting unstructured data into structured data that can be recognized by machine learning (Allahyari at el. 2017). According to Shuai (2016), by giving a database with large handwritten images, handwritten Chinese characters can be recognized by determining the sub-components in Chinese characters such as radical. They have shown the Chinese character in the form of the printed image applied in NMF and how NMF works with the handwritten Chinese character.

METHODOLOGY

CHINESE CHARACTER RECOGNITION USING NMF

Text mining is applied in character recognition to allow the unstructured data of Chinese character in the form of printed images to be converted into structured data with binary values. Decomposing Chinese characters into their graphical components is an important step in understanding the structure of a character. This process can be done without considering the strokes of the character or its semantic meaning using NMF.

The size of the data is one of the important factors to be considered when performing data analysis. The first step is to find the size of the dataset via preliminary analysis. Four Chinese characters '好', '明', '可', and '借' were selected to identify the size of the dataset, and the dimensions for matrices W and H in (2). The sizes of 20×20 and 50×50 are used in the preliminarily analysis and dimensions of 20×K and 50×K for matrix , where are tested to search for the Euclidean distance between the original Boolean matrix and . In Figure 2A, it is evident that the character '明' exhibits the minimum percentage error for K = 3 as opposed to K = 5. Examining Figure 2B, it becomes apparent that '可' attains the lowest percentage error specifically for a dataset size of 50×50. Overall, Figure 2 illustrates that the smallest percentage error occurs when K is higher. Consequently, it is clear that both K = 3 and K = 5 yield superior results in comparison to other values. Therefore, for further analysis, a dataset size of 20×20, along with K values of 3 and 5, is selected.

One hundred printed Chinese characters were collected and each Chinese character was resized into a Boolean matrix. The one hundred Chinese characters (as shown in Figure 3) are further categorized into ten categories according to the number of strokes where . Text mining is applied to the one hundred Chinese characters in the Boolean matrix form. Text mining is a technique to transform unstructured data into structured data that enables Python to recognize the input. Text mining combines machine learning, statistics, and linguistics to find textual patterns in unstructured data. Through this process, more quantitative information can be generated as text data is high dimensional and sparse which consists of many entries zero. Figure 4 shows the process of converting a Chinese character into a Boolean matrix of size

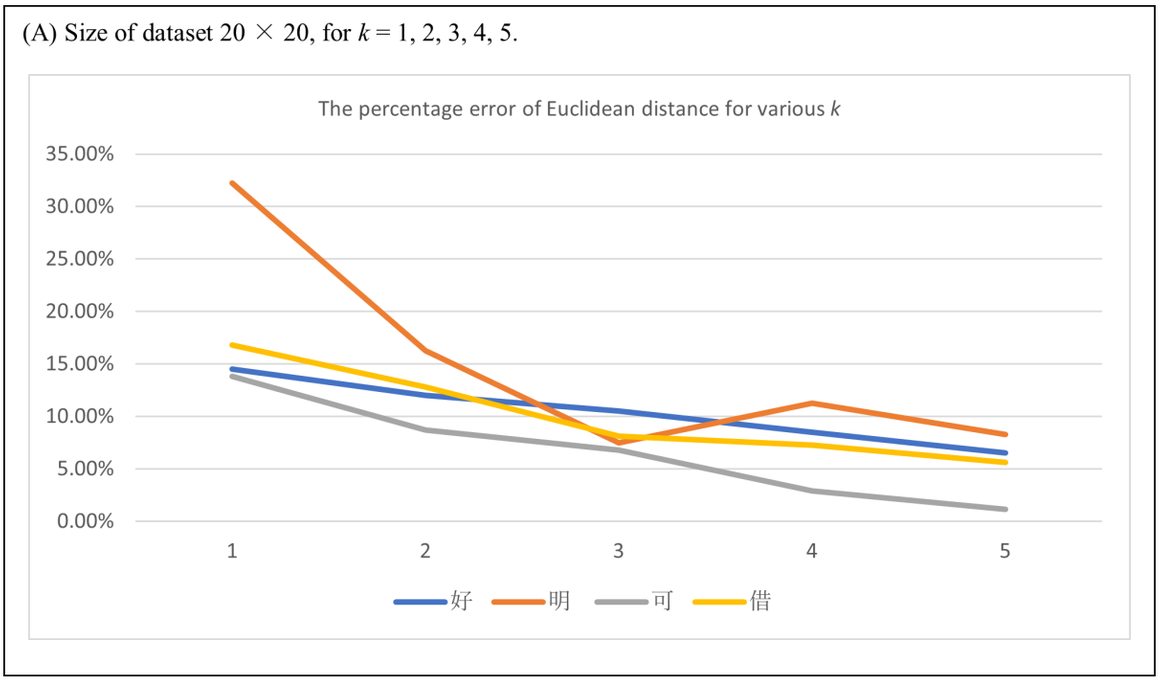


FIGURE 2. The percentage error of Euclidean distance between two matrices, *B* and *WH*, is compared on selected Chinese characters.

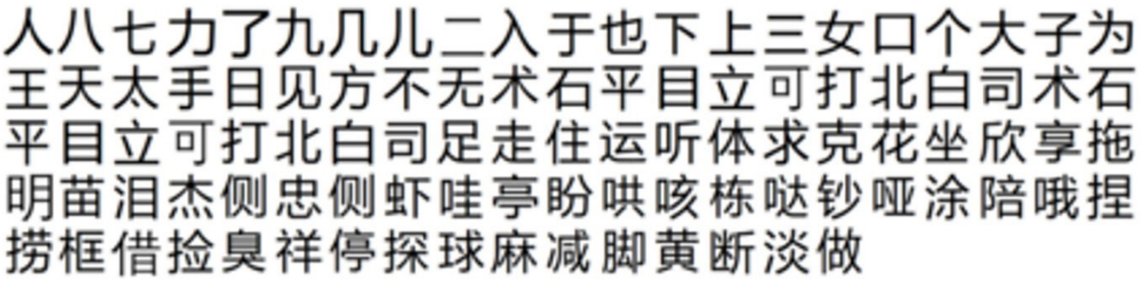


FIGURE 3. The one hundred printed Chinese characters with various strokes

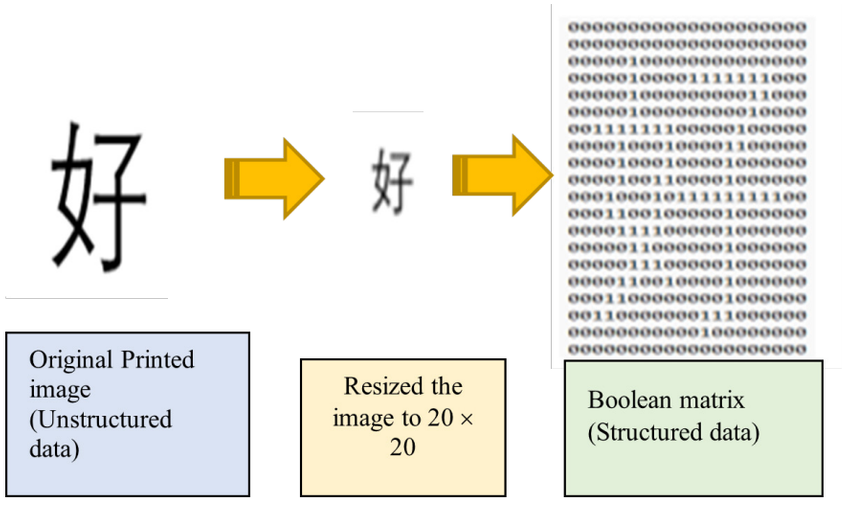


FIGURE 4. The process of converting a printed Chinese character into a Boolean matrix

After converting the printed Chinese character into structured data in a Boolean matrix, B of size 20×20, the matrix B is factorized using both LU factorization and NMF. Table 1 shows the output of a Chinese character in the Boolean Matrix after NMF for K=3 and K=5.

The Euclidean distance between the Boolean matrix of a Chinese character and the matrix, F where $F=LU$ when LU factorization is used and $F=WH$ when NMF is used. Since the matrix $F=(f_{ij})$ (after factorization) might not be a Boolean matrix, the matrix F can be converted into a Boolean matrix $G=(g_{ij})$ by using the formula.

$$g_{ij} = \begin{cases} 1, & if\ f_{ij} \geq 0.5 \\ 0, & if\ f_{ij} < 0.5. \end{cases} \qquad (5)$$

The Boolean matrix facilitates the immediate viewing of the output of the Chinese character following matrix factorizations. The Euclidean distance between the matrices B and F, $\|B - F\|$, and B and G, $\|B - G\|$ is calculated and analyzed, respectively.

FIGURE 5. Ten handwritten Chinese characters with various strokes

To understand the relations between the printed Chinese characters and handwritten Chinese characters in Boolean matrices, ten handwritten Chinese characters with a different number of strokes are selected, as shown in Figure 5. The Euclidean distance of the handwritten and the printed Chinese characters (as shown in Figure 3) from the same category (number of strokes) is calculated and analyzed.

TABLE 1. Output of a Chinese character "好 " Boolean Matrix after NMF for K = 3 and K = 5

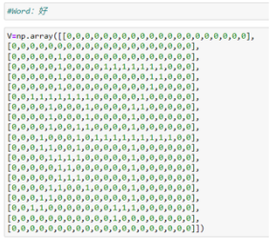
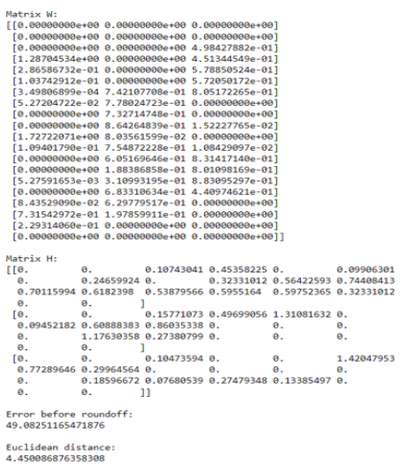
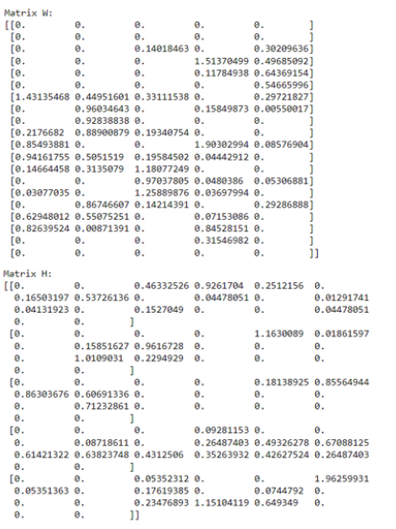
Chinese Character	(NMF)	
	K = 3	K = 5
		

TABLE 2. Comparison between the LU factorization and NMF on the Boolean matrix of selected Chinese characters with various strokes

Chinese Characters	Euclidean distance of $\|B - F\|$ (%)		Euclidean distance of $\|B - F\|$ (%)	
	LU factorization (F)	NMF (F)	LU factorization (G)	NMF (G)
二	14	0.0008	14	0.0007
☒	12.5	6.9523	12.5	3.0881
天	21	8.1097	21	4.5023
可	17.5	6.8081	17.5	1.1249
好	15.5	12.2706	15.5	9.0090

continue ...

... cont.

☒	23.5	11.4497	23.5	6.3005
明	12.5	9.4532	12.5	3.7948
亭	23.5	4.6918	23.5	1.6116
借	21	8.1295	21	5.6605
做	22	10.7280	23	6.8633

RESULTS AND DISCUSSION

Ten printed Chinese characters, each with a different number of strokes, were selected in the form of a Boolean matrix denoted as B. The Boolean matrix for each character underwent factorization using both LU factorization and NMF. The Euclidean distance between the Boolean matrix *B* and the matrices after factorization, denoted as *F* (for both LU factorization and NMF), is calculated and presented in Table 2.

As indicated in Table 2, the Euclidean distance for LU factorization is nearly 18%, significantly higher compared to NMF, which averages around 7.9%. The average Euclidean distance for LU factorization remains approximately 18%, while NMF improves to 4.2% after the transformation of *F* into a Boolean matrix *G*. A paired *t*-test is employed to compare the Euclidean distances resulting from LU factorization and NMF. Let represent the mean calculated using LU factorization, and represent the mean calculated using NMF factorization. The null hypothesis states , while the alternative hypothesis states . With a test statistic , the paired *t*-test confirms that the factorization using NMF is superior to LU factorization when a Chinese character is factorized in Boolean matrix form.

Python programming was utilized for the recognition of handwritten Chinese characters in comparison to their printed counterparts, and the results are presented in Table 3. Ten handwritten Chinese characters were correlated with their printed images based on an equal number of strokes. For instance, the handwritten Chinese character "二," composed of two strokes, undergoes a calculation of the Euclidean distance when compared to the printed character, considering characters such as "人," "八," "七," "力," "了," "九," "几," "儿," "二," and "入." Numerical data reveals that the Euclidean distance error between the handwritten and printed character "二" is the smallest. As indicated in Table 3, 70% of the handwritten Chinese characters could be accurately recognized by obtaining the least Euclidean distance. However, the remaining 30% of handwritten characters, specifically the words '体,' '借,' and '做,' pose challenges for recognition. This is a common issue in handwriting recognition, primarily attributed to the complex construction of Chinese characters involving various strokes. The challenge is closely tied to the way the handwritten text is cropped and positioned.

TABLE 3. Result of handwritten Chinese character recognition

Chinese Character	Euclidean distance between two matrices	The smallest error compared to the printed image under the same number of strokes
二	7.6811	Yes
个	5.7446	Yes
天	6.5574	Yes
可	8.6023	Yes
好	8.4853	Yes
体	9.1104 (Smallest error = 9)	No
明	11.0905	Yes
亭	9.6437	Yes
借	10.6771 (Smallest error = 10)	No
做	9.9499 (Smallest error = 9.6437)	No

CONCLUSION

Non-negative matrix factorization (NMF) can be used to reduce the dimension of the matrix with non-negative entries. In this project, NMF was applied in Chinese character recognition. As a comparison, Python was used to carry out the LU factorization and non-negative matrix factorization of a Chinese character in the Boolean Matrix. The preliminary analysis confirmed that the data size of and and were chosen for the NMF. A sample with one

hundred printed Chinese characters was collected, and all the Chinese characters were categorized into ten categories according to the number of strokes , for . The Euclidean distance between the Boolean matrix of a Chinese character and the matrix after both LU factorization and NMF was calculated for further analysis. Experimental results showed that the factorization using NMF is better than the LU factorization when a Chinese character was factorized in the Boolean matrix form. The average of the Euclidean distance using NMF improved to 4.2% on average whereas the LU factorization still has a high average of around 18%. NMF is suitable to be used in Chinese character recognition since it can reduce the dimension of the image and show the lowest error between the original Boolean matrix and the matrix after factorization.

Ten handwritten Chinese characters were selected to test whether the program is able to identify the handwritten and the printed Chinese characters. The experimental results reveal that 70% of the characters can be accurately recognized through the least Euclidean distance obtained. Looking ahead, there is a possibility for computers to decipher handwritten text and convert it into digital form. This could be achieved by analyzing the position of each stroke and the methodology employed for cropping a handwritten image. For future research, investigating recognition software that requires the text to be presented in a specific orientation, devoid of stroke overlap, and without any additional space on either side of the words holds significant potential.

ACKNOWLEDGEMENT

The authors would like to Universiti Tunku Abdul Rahmanfor support this research.

DECLARATION OF COMPETING INTEREST

None

REFERENCES

Alayat, A. B., & Omar, H. A. 2023. Pavement Surface Distress Detection Using Digital Image Processing Techniques. *Jurnal Kejuruteraan*, *35*(1), 247 – 256.

Allahyari, M., Pouriyeh, S. & Assefi, M. 2017. A brief survey of text mining: classifucation, clustering and extraction techniques. *arXiv:1707.02919.* https://doi.org/10.48550/arxiv.1707.02919

Bernard, K. & David R. H. 2001. *Introductory Linear Algebra with Applications.* Prentice Hall.

Gan, J., Liu, T & Zhang, J. 2021. Non-negative matrix factorization: a survey. *The Computer Journal,* 64(7), 1080–1092.

Golub, G.H. & Van Loan, C.F. 2022. *Matrix Computations*. 4th Edition, Johns Hopkins University Press, Baltimore.

Handschutter, P. D., Gillis, N. & Siebert, X. 2021. A Survey on deep matrix factorizations. *Computer Science Review* 42: 100423.

Hogben, L. (Ed.). 2013. *Handbook of Linear Algebra*. Second edition. Chapman and Hall/CRC.

Langville, A.N., Meyer, C.D., Albright, R., Cox, J., & Duling, D. 2014. Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization. *arXiv preprint arXiv:1407.7299.*

Lee, D. D. & Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature,* 401(6755), 788–791.

Lee, D. D. & Seung, H. S. 2000. Algorithms for non-negative matrix factorization. Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS'00), 535–541.

Lin, C. Y., Kang, L. W., Huang, T. Y. & Chang, M. K. 2019. A novel non-negative matrix factorization technique for decomposition of Chinese characters with application to secret sharing. *EURASIP Journal on Advances in Signal Processing*, 1–8.

Lin, X. & Boutros, P. C. 2020. Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, 21(1), 1–10.

Milad, A., Zaki, A. H. M., Omar, H.A., Ali, I.A., Memon, N.A., Yusoff, N.I.M. 2023. Predicting the Rheological Properties of Bitumen-Filler Mastic Using Machine Learning Techniques *Jurnal Kejuruteraan*, *35*(4), 889–899.

Paatero, P. & Tapper, U. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126.

Pauca, V. P., Shahnaz, F., Berry, M. W. & Plemmons, R. J., 2004. Text Mining using Non-Negative Matrix Factorizations. Proceedings of the 2004 SIAM international conference on data mining, 452–456.

Schwarzenberg-Czerny, A. 1995. On matrix factorization and efficient least squares solution. *Astron. Astrophys. Suppl. Ser.* 110, 405–410.

Shuai, X. 2016. Radical Recognition in Off-Line Handwritten Chinese Characters Using Non-Negative Matrix Factorization. *Senior Projects Spring 2016*. 367. https://digitalcommons.bard.edu/senproj_s2016/367

Tan, J., Xie, X. H., Zheng, W. H. & Lai, J. H. 2012. Radical extraction using affine sparse matrix factorization for printed Chinese characters recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(3), 211–226.

Sudipto, B., & Anindya, R. 2014. *Linear algebra and matrix analysis for statistics*. CRC Press.

Wang, Y. X. & Zhang, Y. J. 2013. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering* 25(6), 1336–1353.

Zamani, N. S. M., Yoon, E.C.H., Huddin, A. B, Zaki, W. M. D. W., & Hamid, Z. A. 2023. Deep Learning for an Automated Image-Based Stem Cell Classification. *Jurnal Teknologi*, *35*(5), 1181–1189.