

Swin Wavelet Transformer (SWT): Mixing Tokens with Wavelet and Multiwavelet Transforms

Waleed A. Mahmoud Al-Jawher

College of Engineering, Uruk University, Baghdad, Iraq

profwaleed54@gmail.com

Abstract The Swin Transformer possess a hierarchical structure, a robust structure, and an efficient self-attention mechanism. This makes it superior in performance for a wide variety of artificial intelligence and machine training tasks. It has revolutionized the field of digital signal processing and its applications by creating vision in multiple fields. It uses non-overlapping windows which leads to creating a distinct perspective and understanding of the context of the digital signals. However, it possess a very complicated hierarchical structure an complexities together. By reducing calculations and complexities together. This will lead to robust performance that make it a powerful tool for a wide range of computer vision tasks. In this research, simplified symbol mixing methods were developed for coding structures similar to transformers, by forming various semantics in the text through linear mixing transformation and in combination with nonlinearity in the feed-forward layers. Hence, we were able to successfully propose a Swin Wavelet Transformer (SWT) model in which the self-attention sublayer was replaced by a Wavelet transform. In a second attempt, the wavelet transform was replaced by the multi-wavelet transform. These two proposed models are achieved a better performance from their FNETs and BERT counterparts and are highly competitive with traditional and efficient transformers.



 Crossref  [10.36371/port.2024.3.14](https://doi.org/10.36371/port.2024.3.14)

Keywords: Swin Transformer, Deep Learning, Wavelet transform, Multiwavelet Transform, Multi Head Attention, BERT, CNN.

1. INTRODUCTION

A newly suggested neural network architecture called Swin Transformer is intended to be efficient and effective for computer vision applications [1-3]. Based on the Transformer architecture, Swin Transformer makes some changes to make it more effective for processing image input. The Swin Transformer uses shifted windows, which is one of its main improvements. The attention mechanism is used for every pixel in a picture in conventional Transformer systems. The attention method is only used to focus on a portion of pixels at a time in Swin Transformer's shifting window technique. As a result, the Swin Transformer is more effective since it only has to calculate the attention weights for a portion of the pixels. Language-to-vision adaptation issues result from distinctions between the two fields, such as the size of visual things and pixels in high-resolution images compared to words in the text [4-12]. A hierarchical Transformer with representation generated via a sliding window is suggested as a solution to overcome inconsistencies. By confining the computation of self-attention to small, non-overlapping windows while allowing connections between windows, the sliding window technique increases efficiency. This layered design offers linear computing costs concerning picture size and the ability to simulate at different sizes. These characteristics of the Swin Transformer make it suitable for a variety of visual tasks [13-24].

2. SWIN TRANSFORMER

Swin Transformers (ST) is a subcategory of Vision Transformers. It constructs hierarchical feature maps by merging image patches into deeper layers and has a linear computational complexity proportional to the size of the input image due to self-attention processing occurring only within each local window. As a result, it can be used as a general-purpose backbone for picture classification and dense recognition applications. In comparison, earlier vision Transformers generate feature maps with a single low resolution and have a quadratic computational complexity proportional to the size of the input image due to global self-attention processing. [25-33]

Swin Transformer (the name Swin stands for Shifted window operation shown in Fig. 1) is initially consists of four stages. Each stage contains two Swin transformer blocks except the stage three that contains six Swin transformer blocks. It capably serves as a general-purpose backbone for computer vision. It is basically a hierarchical Transformer whose representation is computed with shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. Swin Transformer achieves strong performance on COCO object detection, semantic segmentation, surpassing previous models by a large margin [34-47].

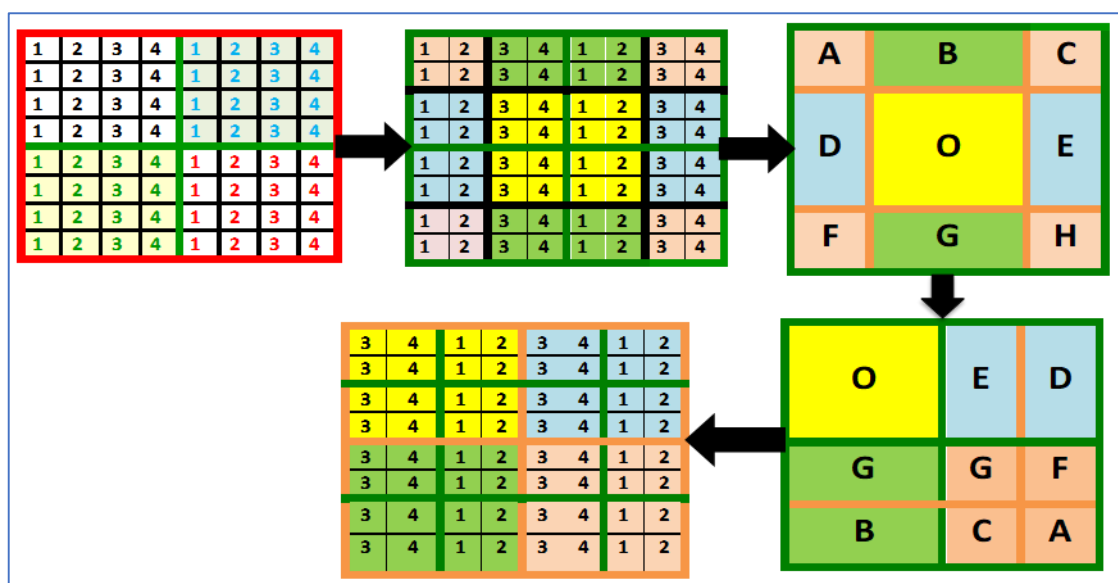


Fig. 1 Graphical representation of Shifted window operation

Appendix A gives numerical example that demonstrated the computations involved in the Swin transformer stages. As well as it shows the computational expense involved in calculating attention shown in Fig. 2 for images of small dimensions.

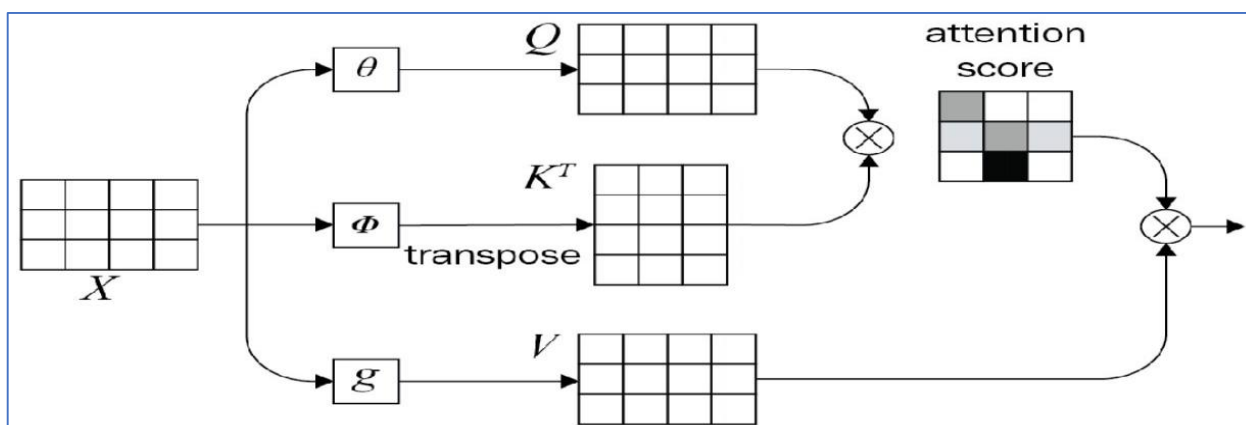


Fig. 2. The self-Attention of Swin Transformer

3. FROM FOURIER TRANSFORM TO MULTIWAVELET TRANSFORM

Two methods exist for analyzing signals: the spatial and frequency domain methods. A suitable transform model connects these two equivalent domains. In general, all transformations decompose a spatial signal to reveal the frequency of waves required to simulate that spatial signal. These frequencies can reveal difficult-to-manage details and characteristics of the signal in the spatial domain and provide tools for analyzing and designing signal processing systems [48-62].

In many signal processing applications, it is preferable to use discrete functions rather than functions defined at all time values; for instance, it is preferable to sample images or video signals at discrete time intervals before processing them. For this reason, the Discrete Fourier Transform (DFT) was

developed to accommodate all instrumental and experimental calculation requirements.

It is easy for unauthorized individuals to decipher, resulting in Weak security. However, its speed and ease of use are unmatched [63-70]. Due to its simplicity and speed when dealing with 2D images, the Arnold Transform was a critical stage in the proposed image encoding algorithms. Although the DFT can efficiently extract valuable information about the signal's frequency components, it requires a massive computational effort, which prevents it from being used in most practical applications. This prompted Cooley and Tukey to propose a numerical algorithm that significantly reduces the number of calculations required for DFT evaluation. This algorithm was dubbed the Fast Fourier Transform (FFT). It is regarded as a turning point in digital signal processing, as it has been credited with facilitating the application of the Fourier



transform in real-world scenarios [52-68]. The Discrete Hartley Transform (DHT), introduced by Hartley in [58], is closely related to the Fourier transform of a real-valued sequence. The Hartley transform can obtain a Fourier power spectrum for a given function using only real-valued computation [6]. The following equations define the two-dimensional DHT: Noting that the only difference between the forward transform and the DFT is the absence of the " -j " multiplier on the sine term, the DHT of a real-valued sequence is simply the imaginary part of the DFT subtracted from the real part [59]. Arnold Transform, also known as Arnold cat map, is a chaotic system utilized extensively in image encryption to transform pixels from position (x, y) to position (x1, y1) straightforwardly and efficiently [60]. Wavelet transformations are a relatively new phenomenon that has captivated researchers in the physical and mathematical sciences with their diverse application potential. Due to connections with multi-rate filtering, quadrature mirror filters, and sub-band coding, wavelet application fields have expanded rapidly over the past few years. It has been utilized by the digital signal processing community to eliminate noisy signals and compress data and images. One of the primary reasons for detecting waveforms and wavelet transforms is that the Fourier transform analysis lacks local information about the signals. Therefore, the Fourier transform cannot be used to analyze signals in the time and frequency domain. As the wavelet transform computes the internal products of a signal with a family of wavelets, a wealth of new mathematical results has been produced. Notable is that the DWT produces very high time resolution at low frequencies and low time resolution at high frequencies, where high frequencies represents a signal's precise details, and low frequencies means the signal's summary, both of which are required for signal analysis and representation [44-70]. Multi-Wavelets Transform (MWT) is a special case of wavelets in which a vector scaling function is used instead of a single scaling function. This allows the multi-resolution analysis spaces to be described in terms of translates of linear combinations of the scaling vector's functions. corresponds to the Multiwavelet function. It has been demonstrated that scalar wavelets cannot simultaneously possess orthogonality, symmetry, compact support, and vanishing moments, whereas MWT does. This ensures that the signal energy and decomposition are preserved without slipping and redundancy, effectively avoiding the error caused by signal truncation and allowing for the correct signal reconstruction [69]. In this research, one of the most significant MWT filters, the GHM filter proposed by Geronimo et al, is utilized. MWT satisfy two-scale refinement equations. It is easy for unauthorized individuals to decipher, resulting in Weak security. However, its speed and ease of use are unmatched [94-98]. Due to its simplicity and speed when dealing with 2D images, the Arnold Transform was a critical stage in the proposed image encoding algorithms. The algorithms were augmented with additional multidimensional chaotic systems to circumvent the security issues caused by the intrinsic periodicity of the Arnold transform and improve the image's security and reliability. Although the DFT can

efficiently extract valuable information about the signal's frequency components, it requires a massive computational effort, which prevents it from being used in most practical applications. This prompted Cooley and Tukey to propose a numerical algorithm that significantly reduces the number of calculations required for DFT evaluation. This algorithm was dubbed the Fast Fourier Transform (FFT). It is regarded as a turning point in digital signal processing, as it has been credited with facilitating the application of the Fourier transform in real-world scenarios.[57 - 55] The Discrete Hartley Transform (DHT), introduced by Hartley in [58], is closely related to the Fourier transform of a real-valued sequence. The Hartley transform can obtain a Fourier power spectrum for a given function using only real-valued computation [6]. The following equations define the two-dimensional DHT: Noting that the only difference between the forward transform and the DFT is the absence of the " -j " multiplier on the sine term, the DHT of a real-valued sequence is simply the imaginary part of the DFT subtracted from the real part.[59] . Arnold Transform, also known as Arnold cat map, is a chaotic system utilized extensively in image encryption to transform pixels from position (x, y) to position (x1, y1) straightforwardly and efficiently [60]. Wavelet transformations are a relatively new phenomenon that has captivated researchers in the physical and mathematical sciences with their diverse application potential. Due to connections with multi-rate filtering, quadrature mirror filters, and sub-band coding, wavelet application fields have expanded rapidly over the past few years. It has been utilized by the digital signal processing community to eliminate noisy signals and compress data and images. One of the primary reasons for detecting waveforms and wavelet transforms is that the Fourier transform analysis lacks local information about the signals. Therefore, the Fourier transform cannot be used to analyze signals in the time and frequency domain. As the wavelet transform computes the internal products of a signal with a family of wavelets, a wealth of new mathematical results has been produced. Notable is that the DWT produces very high time resolution at low frequencies and low time resolution at high frequencies, where high frequencies represents a signal's precise details, and low frequencies means the signal's summary, both of which are required for signal analysis and representation.[67 ,66 ,57 ,56 ,34 ,28] . Multi-Wavelets Transform (MWT) is a special case of wavelets in which a vector scaling function is used instead of a single scaling function. This allows the multi- resolution analysis spaces to be described in terms of translates of linear combinations of the scaling vector's functions.[68] corresponds to the Multiwavelet function. It has been demonstrated that scalar wavelets cannot simultaneously possess orthogonality, symmetry, compact support, and vanishing moments, whereas MWT does. This ensures that the signal energy and decomposition are preserved without slipping and redundancy, effectively avoiding the error caused by signal truncation and allowing for the correct signal reconstruction [69]. In this research, one of the most significant MWT filters, the GHM filter proposed by Geronimo et al [70], is utilized. MWT satisfy two-scale refinement equations.

4. THE PROPOSED SWIN WAVELET TRANSFORMER (SWT) MODEL.

The success in applying attention with NLP tasks and images reached the state of art in these days. These achievements are due to their flexibility and ability to create understanding of the attention. They may use most of these techniques in various treatments. Although we have endeavored to achieve a lot of work to gain a more distinct understanding of attention, they

successful dazzling concerts derive from work patterns based on compositions in different layers.

However, there is a cost associated with this interest that must be paid in order to achieve somewhat successful results. It has been proven that the attention weights adopted in their complex model are very successful and expressive, but are not necessary to achieve accurate NLP models

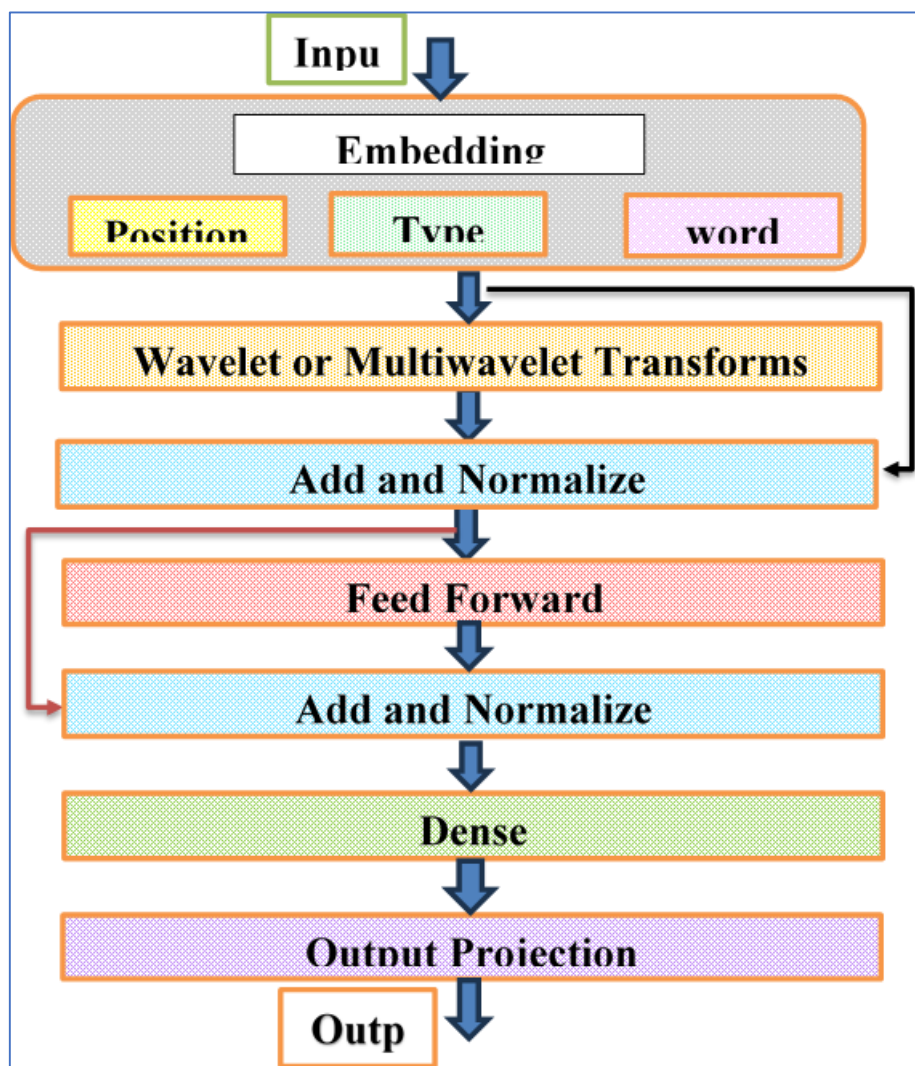


Fig. 3 Proposed Swin Wavelet-Multiwavelet Transformer Architecture.

Acceptable results were obtained that are not far from the exact ones when retaining learnable mutual attention weights. The standard attention mechanism requires many times more time and creates a bottleneck in memory in the case of long sequences, which makes it difficult to apply in such cases. Note that such a mechanism requires great efforts to improve attention efficiency by relying on dispersing the attention matrix.

This research replaces all these methods of self-attention with a different combination and mixture through the use of the Wavelet transform. This development leads to outstanding

performance and reducing the size of the model by dispensing with learnable attention parameters as well as in the simplicity of the new model.

The proposed model is a care-free transformation structure, where the main layers consist of Wavelet mixing sub-levels followed by a feed-forward sub-level. Figure 2 shows the structure of the new structure. Note that the self-interest sublayer of each sublayer is replaced by Wavelet, which applies a 2D Wavelet transformer to the embedding input. The first wavelet will be along the serial dimension and the second wavelet will be along the hidden dimension.

Since the Wavelet parameters are real numbers, we do not need to modify the feed-forward sublayers or the output layers here. After testing the Wavelet model, it became clear to us that it gives the best results compared to FNet. It is worth noting that here we combine word embeddings, absolute position embeddings for tokens, and type embeddings for sentences. This is necessary because we secure placement inclusions to allow for clear, consistent comparison with BERT.

The simplest explanation for Wavelet transformation is that it is a particularly efficient token shuffling mechanism, which provides the feed-forward sublayers with sufficient access to all tokens.

Because of the duality of the Wavelet transform, we can also view each alternating cipher block as applying alternating Wavelet transforms and inverse Wavelet transforms, transforming the input back and forth between the time domain and the frequency domain.

Because multiplication by feed-forward sublayer coefficients in the frequency domain is equivalent to convolution (with a related set of coefficients) in the time domain, the new model can be viewed as an alternation between multiplication and convolution.

5. DATASET USED IN THIS WORK

The datasets that were used in the proposed model are explained in this section. A dataset for X-ray, MRI, and CT scan medical images datasets were used. Below are the explanations for each:

Dataset A: As part of the study, MRI scans of brain tumors were obtained from six different Kaggle datasets to obtain the data used for this study. Several datasets provide annotated images that can be used for analysis in these datasets. Dataset 1 contained 253 MRI brain tumor images categorized into two labels: 98 images without tumors and 155 images with tumors. In Dataset 2, 3,264 MRI-labeled images were divided into four parts. The first part consists of 926 glioma images, the second part consists of 937 meningioma images, and the third part consists of 500 images with no tumor labels, while the fourth part consists of 901 images with pituitary labels. Dataset 3 contains 3,060 MRI brain images categorized into three categories: 1,500 images with tumors, 1,500 images without tumors, and 7060 unlabeled images for testing. The fourth dataset, Dataset 4, included 7,023 labeled images divided into four categories. The first part contained 1,621 images of glioma tumors, the second part contained 1,645 images of meningioma tumors, the third part contained 2,000 images without tumors, and the last part contained 1,757 images of pituitary tumors. In Dataset 5, 400 MRI-labeled images were categorized into two groups: 170 normal images (without tumors) and 230 abnormal images. The last dataset of brain MRI images, Dataset 5, contained 2,501 images classified into two categories: 1,551 normal images and 950 abnormal images. Our final classification is based on two types of dataset: normal and

abnormal. There were 5,819 images in the normal class and 10,622 images in the abnormal class, the number of observations in each class. As a result, this dataset contained 16,441 MRI images of brain tumors. A comparison was made on the above data between the achievement of the proposed model SWT and BERT.

BERT Transformer encoder model.

SWT encoder: by the replacement of every self-attention sublayer with a Wavelet sublayer.

SMWT encoder: by the replacement of every self-attention sublayer with a Multiwavelet sublayer.

Linear encoder: by the replacement of each self-attention sublayer with two learnable, dense, linear sublayers, one applied to the hidden dimension and the other to the sequence dimension.

The accuracy of these systems are BERT the best record of classification of average value of 88.3%, the SWT encoder achieved under the same conditions gave an accuracy of 87.6%, SMWT gave 88.1% and the linear encoder 74.9%.

6. A COMPARISON OF FOURIER TRANSFORM AND WAVELET TRANSFORM IN THE CONTEXT OF TRANSFORMER MODELS

The analogy of Fourier Transform (FT) and Wavelet Transform (WT) to the Transformer architecture is a fascinating one, though it's important to note that the direct application of these transforms within the model isn't straightforward.

6.1 Fourier Transform and Transformer Layers

The Fourier Transform is a mathematical tool used to decompose a function into its constituent frequencies. In the context of Transformer models, the self-attention mechanism can be seen as a form of frequency analysis, where the model learns to attend to different parts of the input sequence based on their relevance.

However, the Fourier Transform's limitation lies in its inability to capture time-frequency information. It's better suited for stationary signals, where the frequency content remains constant over time. In contrast, natural language is a non-stationary signal, with varying frequencies and patterns across different parts of the text.

6.2 Wavelet Transform and Transformer Layers

The Wavelet Transform, on the other hand, offers a time-frequency representation of a signal. It's particularly useful for analyzing non-stationary signals, like natural language, where frequencies and patterns change over time.

While Transformer layers don't directly implement the Wavelet Transform, they can be seen as capturing a similar idea of multi-scale analysis. The self-attention mechanism, combined

with the hierarchical structure of the Transformer, allows the model to attend to information at

different levels of granularity, from individual words to longer dependencies.

6.3 Key Differences and Similarities

Feature	Fourier Transform	Wavelet & Multi-Wavelet Transform	Transformer
Time-Frequency Analysis	Poor time-frequency resolution	Good time-frequency resolution	Good time-frequency resolution through self-attention
Signal Processing	Primarily for stationary signals	Suitable for non-stationary signals	Handles non-stationary language data effectively
Computational Complexity	Can be computationally expensive, especially for large datasets	More computationally efficient than the Fourier Transform	Efficient due to the attention mechanism and parallel processing capabilities

It is clear that while the Wavelet Transform and Multiwavelet Transform provide valuable insights into signal processing, the Transformer architecture offers a more flexible and powerful approach to natural language processing. By combining the strengths of both worlds, Transformer models can effectively capture the complex patterns and dependencies in language data.

7. CONCLUSIONS

In this research, a new structure for the Swin Transformer was proposed by replacing the attention part with a Wavelet transform. This means that simplified symbol mixing units for coding structures similar to transformers were studied. This

proposed model is similar to a transformer, where the self-attention sublayer is replaced by an unparameterized Wavelet transform. The mixing mechanism using Wavelet is efficient and effective, which is compatible with the accuracy of the most accurate models, in addition to its speed and requiring much less storage in memory.

During its application, it gave the impression that it is an easy alternative to the attention mechanism in all classification tasks. The experience of replacing Wavelet with Multiwavelet was also very successful, and this means that there will be a great demand for using the new structure in the classification task in the future.

REFERENCES

- [1] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
- [2] Rasha Ali Dihin, E AlShemmary and Waleed Al-Jawher "Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet" Journal of Kufa for Mathematics and Computer, Vol. 10, Issue 2, PP. 167-172, 2023.
- [3] Rasha Ali Dihin, Ebtesam N AlShemmary, Waleed AM Al-Jawher, "Wavelet-Attention Swin for Automatic Diabetic Retinopathy Classification" Baghdad Science Journal, 2024.
- [4] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei and Baining Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution" We2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022.
- [5] James Lee-Thorp and Joshua Ainslie and Ilya Eckstein and Santiago Ontañón "FNet: Mixing Tokens with Fourier Transforms" Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4296 – 4313, 2022.
- [6] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tok- enizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, 2018.
- [7] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural lan- guage understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, 2020.
- [8] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. 2020. Fast transformers with clustered at- tention. Advances in Neural Information Processing Systems, 33:21665–21674, 2020.

- [9] Zhenyou Zhang, Yi Wang, and Kesheng Wang. “ Fault diagnosis and prognosis using wavelet packet decomposition, fourier transform and artificial neu- ral network”. Journal of Intelligent Manufacturing, 24(6):1213–1227, 2013.
- [10] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher” Hybrid image encryption algorithm based on compressive sensing, gray wolf optimization, and chaos” , Journal of Electronic Imaging, Volume 32, Issue 4, Pages 043038-043038, 2023.
- [11] Maryam I Mousa Al-Khuzayy, Waleed A Mahmoud Al-Jawher, “New Proposed Mixed Transforms: CAW and FAW and Their Application in Medical Image Classification” International Journal of Innovative Computing, Volume 13, Issue 1-2, Pages 15-21, 2022.
- [12] AHM Al-Helali, W. A. Mahmoud, HA Hali, AF Fadhel “Multispectral Image Fusion using Walidlet Transform” Advances in Modelling and Analysis B, Volume 52, Issu. 1-2, pp. 1-20, 2009.
- [13] Walid Amin Mahmoud-Jawher “Computation of Wavelet and Multiwavelet Transforms using Fast Fourier Transform” Journal Port Science Research, Vol. 4, Issue 2, PP. 111-117, 2021.
- [14] Maryam I Al-Khuzayie, Waleed A Mahmoud Al-Jawher “Enhancing Medical Image Classification: A Deep Learning Perspective with Multi Wavelet Transform” Journal Port Science Research, Vol. 6, Issue 4, PP. 365-373, 2023.
- [15] Waleed. A. Mahmoud & I.K. Ibraheem "Image Denoising Using Stationary Wavelet Transform” Signals, Inf. Patt. Proc. & Class. Vol. 46, Issue 4, Pages 1-18, 2003.
- [16] W. A. Mahmoud & Z. Ragib “Face Recognition Using PCA and Optical Flow” Engineering Journal, Vol. 13, Issue 1, PP. 35-47, 2007.
- [17] SM Saadi, WAM Al-Jawher, “Proposed DeepFake Detection Method Using Multiwavelet Transform,” International Journal of Innovative Computing 13 (1-2), 61-66, 2022
- [18] Qutaiba Kadhim, Waleed Ameen Mahmoud Al-Jawher “A new multiple-chaos image encryption algorithm based on block compressive sensing, swin transformer, and wild horse optimization” Multidisciplinary Science Journal, Vol. 7, Issue 1, PP. 2025012-2025012, 2024.
- [19] Waleed A Mahmoud Al-Jawher, Shaimaa A Shaaban “K-Mean Based Hyper-Metaheuristic Grey Wolf and Cuckoo Search Optimizers for Automatic MRI Medical Image Clustering” Journal Port Science Research, Volume 7, Issue 3, Pages 109-120, 2024.
- [20] Walid A Mahmoud, Majed E Alneby, Wael H Zayer “Multiwavelet Transform and Multi-Dimension-Two Activation Function Wavelet Network Using For Person Identification” Iraqi Journal Of Computers, Communications, Control And Systems Engineering, Vol 11, Issue 1, 2011.
- [21] Waleed A Mahmoud, Ahmed S Hadi “Systolic Array for Realization of Discrete Wavelet Transform “Journal of Engineering, Vol. 13, Issue 2, PP. 1-9, 2007.
- [22] W. A. Mahmoud & Omama Razaq “Speech recognition using new structure for 3D neural network” University of Technology, 1st Computer Conference, PP. 161-171, 2010.
- [23] W. A. Mahmoud & Z. J. M. Saleh “ An Algorithm for Computing Multiwavelets & Inverse Transform Using an Over-Sampled Scheme of Pre& Post processing respectively” Engineering Journal, Vol. 10, Issue 2, PP. 270-288, 2004.
- [24] W. A. Mahmoud Z Jalal & N. K. Wafi “A New Method of Computing Multi-wavelets Transform using Repeated Row Preprocessing.” Al-Rafidain Engineering Journal, Vol. 12, Issue 2, PP. 21-31., 2004.
- [25] E. Dihin, R. Al-Jawher, Waleed and Al-Shemmary “Implementation of The Swin Transformer and Its Application In Image Classification” Journal Port Science Research, vol. 6, Issue 4, PP. 318-331. 2023.
- [26] Waleed Ameen Mahmoud Al-Jawher, A. Barsoum and Entather Mahos “Fuzzy Wavenet (FWN) classifier for medical images” Al-Khwarizmi Engineering Journal, Vol. 1, Issue 2, PP. 1-13, 2005.

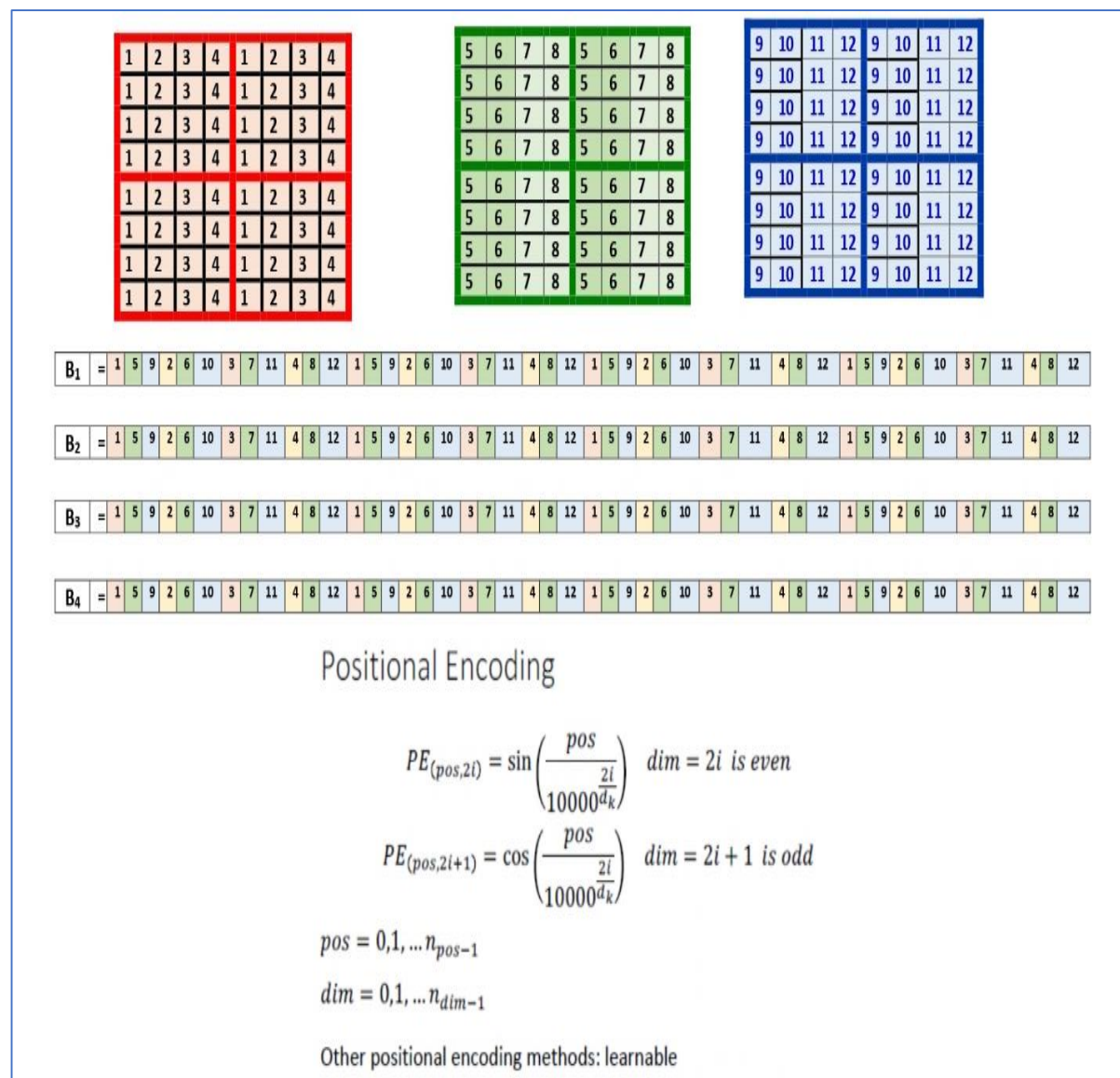
- [27] Waleed A. Mahmoud, MS Abdulwahab, HN Al-Taai: "The Determination of 3D Multiwavelet Transform" IJCCCE, vol. 2, issue 4, 2005.
- [28] Waleed A. M. Al-Jawher, T Abbas – "Feature combination and mapping using multiwavelet transform" IASJ, AL-Rafidain, Issue 19, Pages 13-34, 2006.
- [29] Waleed A Mahmoud, MR Shaker "3D Ear Print Authentication using 3D Radon Transform" proceeding of 2nd International Conference on Information & Communication Technologies, Pages 1052-1056, 2006.
- [30] H. Al-Taai, Waleed A. Mahmoud & M. Abdulwahab "New fast method for computing multiwavelet coefficients from 1D up to 3D", Proc. 1st Int. Conference on Digital Comm. & Comp. App., Jordan, PP. 412-422, 2007.
- [31] Waleed Ameen Mahmoud "A Smart Single Matrix Realization of Fast Walidlet Transform" Journal of Research and Reviews in Computer Science, Volume 2, Issue, 1, PP 144-151, 2011.
- [32] WA Mahmoud, AI Abbas, NAS Alwan "Face Identification Using Back-Propagation Adaptive Multiwavelet" Journal of Engineering 18 (3), 2012
- [33] WA Mahmoud, AS Hadi, TM Jawad "Development of a 2-D Wavelet Transform based on Kronecker Product" - Al-Nahrain Journal of Science, Vol. 15, Issue 4, PP. 208-213, 2012.
- [34] H. M Hasan, Waleed A. Mahmoud Al- Jawher, M. A Alwan "3-d face recognition using improved 3d mixed transform" Journal International Journal of Biometrics and Bioinformatics, Vo. 6, Issue 1, PP. 278-290, 2012.
- [35] Waleed A. Mahmud Al-Jawher, Talib M. J. Abbas Al-Talib, R. Hamudi A. Salman "Fingerprint Image Recognition Using Walidlet Transform" Australian Journal of Basic and Applied Sciences, Australia, 2012.
- [36] Saleem MR Taha, Walid A Mahmood "New techniques for Daubechies wavelets and multiwavelets implementation using quantum computing "Journal Facta universitatis-series: Electronics and Energetics, Volume 26, Issue 2, Pages 145-156, 2013.
- [37] Waleed A Mahmoud, Dheyaa J Kadhim "A Proposal Algorithm to Solve Delay Constraint Least Cost Optimization Problem" Journal of Engineering, Vol. 19, Iss 1, PP 155-160, 2013.
- [38] SMR Taha, WA Mahmood "New techniques for Daubechies wavelets and multiwavelets implementation using quantum computing" Facta universitatis-series: Electronics and Energetics 26 (2), 145-156, 2013.
- [39] Waleed A. Mahmoud, J J. Stephan and A. A. Razzak "Facial Expression Recognition Using Fast Walidlet Hybrid Transform" Journal port Science Research, Volume3, No:1, Pages 59-69 2020.
- [40] R. A. Dihin, W. A Mahmoud Al-Jawher, Ebtesam N AlShemmary "Diabetic Retinopathy Image Classification Using Shift Window Transformer", Journal of Innovative Computing, Vol. 13, Issue 1-2, PP. 23-29, 2022.
- [41] Rasha Ali Dihin, Ebtesam N. AlShemmary and Waleed A. Mahmoud Al-Jawher "Automated Binary Classification of Diabetic Retinopathy by SWIN Transformer" Journal of Al-Qadisiyah for computer science and mathematics (JQCM), Vol 15, Issue 1, PP. 169-178, 2023.
- [42] L R. Hussssein and J. M. A. Al-Sammarie W. A. Mahmoud "Image Identification using Minimum Distance Classifier with Multi-Wavelet Transform" Journal of Advances in Modelling and Analysis B, Volume 46, Issue (5-6), pages 1-22, 2003.
- [43] Waleed A Mahmoud, MR Shaker "3D Ear Print Authentication using 3D Radon Transform" proceeding of 2nd International Conference on Information & Communication Technologies, Pages 1052-1056, 2006.
- [44] W Al-Jowher, N Al-Ramahi, M. Alfaouri "Image Identification And Labeling Using Hybrid Transformation And Neural Network" Neural Network World: International Journal on Neural and Mass - Parallel Computing and Information Systems; Prague, Volume 17, Issue 4, Pages 377-395, 2007.
- [45] Ibraheem Al-Jadir, Waleed A Mahmoud "A Grey Wolf Optimizer Feature Selection Method and its Effect on the Performance of Document Classification Problem" Journal Port Science Research, Vol. 4, Issue 2, Pages 125-131, 2021.

- [46] W. A. Mahmoud, Jane Jaleel Stephan and A. A. W. Razzak “Facial Expression Recognition from Video Sequence Using Self Organizing Feature Map” *Journal port Science Research, Transaction on Engineering, Technology and Their Applications*, Vol. 4, Issue 2, Pages 53-68, 2021.
- [47] AAR Sakran, SM Hadi, WAM Al-Jawher “A New Approach for DNA Sequence Analysis Using Multiwavelet Transform (MWT)” *Journal of Physics: Conference Series* 2432 (1), 012022.
- [48] Q. K Abed, W. A Mahmoud Al-Jawher “A Robust Image Encryption Scheme Based on Block Compressive Sensing and Wavelet Transform” *International J. of Innovative Computing*, Vol. 13, I. 1-2, PP. 7-13, 2022.
- [49] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher, “Image Encryption Algorithm Based on Arnold Transform and Chaos Theory in the Multi-wavelet Domain”, *International Journal of Computers and Applications*, Vol. 45, Issue 4, pp. 306-322, 2023
- [50] L. F. Katran, Ebtesam N AlShemmary, Waleed Ameen Al Jawher “Deep Learning's Impact on MRI Image Analysis: A Comprehensive Survey” *Texas Journal of Engineering and Technology*, Vol. 25, PP. 63-80, 2023.
- [51] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher “A Hybrid Domain Medical Image Encryption Scheme Using URUK and WAM Chaotic Maps with Wavelet–Fourier Transforms” *Journal of Cyber Security and Mobility*, Pages 435–464-435–464, 2023.
- [52] Zahraa A Hasan, Suha M Hadi, Waleed A Mahmoud, “Speech scrambler with multiwavelet, Arnold Transform and particle swarm optimization” *Journal Pollack Periodica*, Volume 18, Issue 3, Pages 125-131, 2023.
- [53] Zahraa A Hasan 1, Suha M. Hadi, Waleed A. Mahmoud al-Jawher “Speech scrambling based on multiwavelet and Arnold transformations”, *Indonesian Journal of Electrical Engineering and Computer Science*, 30 , 2023.
- [54] M. I Al-Khuzai, Waleed A Mahmoud Al-Jawher “Enhancing Medical Image Classification: A Deep Learning Perspective with Multi Wavelet Transform” *Journal Port Science Research*, Vol. 6, Issue 4, PP. 365-373, 2023.
- [55] Lamyaa Fahem Katran, Ebtesam N AlShemmary, Waleed Ameen Al Jawher “Deep Learning's Impact on MRI Image Analysis: A Comprehensive Survey” *Texas Journal of Engineering and Technology*, Vol. 25, PP. 63-80, 2023.
- [56] L. Fahem Katran, Ebtesam N AlShemmary, Waleed AM Al-Jawher “A Review of Transformer Networks in MRI Image Classification” *Al-Furat Journal of Innovations in Electronics & Computer Engineering*, PP. 148-162, 2024.
- [57] S. Saadi, W. Al-Jawher” Enhancing image authenticity: A new approach for binary fake image classification using DWT and swin transformer” *Global Journal of Engineering & Technology Advances*, 19(3), 2024.
- [58] Maryam I Mousa Al-Khuzai, Waleed A Mahmoud Al-Jawher “Enhancing Brain Tumor Classification with a Novel Three-Dimensional Convolutional Neural Network (3D-CNN) Fusion Model” *Journal Port Science Research*, Volume 7, Issue 3, Pages 254-267, 2024.
- [59] L. Katran, EN AlShemmary, WAM Al-Jawher “Integrating Swin Transformer with Fuzzy Gray Wolf Optimization for MRI Brain Tumor Classification” *Journal of Intelligent Engineering & Systems*, Vol. 17, Issue 6. 2024.
- [60] Waleed A Mahmoud, MR Shaker “3D Ear Print Authentication using 3D Radon Transform” *proceeding of 2nd International Conference on Information & Communication Technologies*, Pages 1052-1056, 2006.
- [61] WA Mahmoud, ALM Rasheed “3D Image Denoising by Using 3D Multiwavelet” *AL-Mustansiriya J. Sci* 21 (7), 108-136, 2010.
- [62] H. M Hasan, Waleed A. Mahmoud Al- Jawher, M. A Alwan “3-d face recognition using improved 3d mixed transform” *Journal International Journal of Biometrics and Bioinformatics*, Vo. 6, Issue 1, PP. 278-290, 2012.
- [63] Waleed A. Mahmud Al-Jawher, Talib M. J. Abbas Al-Talib, R. Hamudi A. Salman “Fingerprint Image Recognition Using Walidlet Transform” *Australian Journal of Basic and Applied Sciences*, Australia, 2012.
- [64] SMR Taha, WA Mahmood “New techniques for Daubechies wavelets and multiwavelets implementation using quantum computing” *Facta universitatis-series: Electronics and Energetics* 26 (2), 145-156, 2013.

- [65] Waleed A. M. Al-Jawher, T Abbas – “Feature combination and mapping using multiwavelet transform” IASJ, AL-Rafidain, Issue 19, Pages 13-34, 2006.
- [66] Walid A Mahmoud, Majed E Alneby, Wael H Zayer “2D-multiwavelet transform 2D-two activation function wavelet network-based face recognition” J. Appl. Sci. Res, vol. 6, issue 8, 1019-1028, 2010.
- [67] Waleed Ameen Mahmoud “A Smart Single Matrix Realization of Fast Walidlet Transform” Journal of Research and Reviews in Computer Science, Volume 2, Issue, 1, PP 144-151, 2011.
- [68] W. Mahmoud, D Kadhim “A Proposal Algorithm to Solve Delay Constraint Least Cost Optimization Problem” Journal of Engineering, Vol. 19, Iss 1, PP 155-160, 2013.
- [69] M. Al-Khuzae, W. Al-Jawher “Enhancing Brain Tumor Classification with a Novel Three-Dimensional Convolutional Neural Network Fusion Model” Journal Port Science Research, Volume 7, Issue 3, Pages 254-267, 2024.
- [70] W. A. Mahmoud & I. A Al-Akialy “A Tabulated Method of Computation Multiwavelet Transform” Al-Rafidain University College, Vol. 15, PP. 161-170, Iraq, 2004.

Appendix A : Demonstrated Example on the comutation of Swin Transformer

$PE(0, 0) = \sin(0/10000. \wedge (2^*0)/12) = 0$, $PE(0, 1) = \cos(0/10000. \wedge (2^*1)/12) = 1$,



B1= 1 5 7 2 6 8 9 3 2 10 14 10 + PE1= 0 1 0 1 0 1 0 1 0 1 0 1 =1 6 7
3 6 9 9 4 2 11 14 11

B2= 3 7 9 4 8 4 1 5 4 12 16 1 + PE2= 0.8415 0.9769 0.0464 1.0000 0.0022 1.0000 0.0001
1.0000 0.0000 1.0000 0.0000 1.0000 = 3.8415 7.9769 9.0464 5.0000 8.0022 5.0000 1.0001 6.0000 4.0000
13.0000 16.0000 2.0000

B3= 7 8 7 11 12 9 5 9 6 6 3 4 + PE3= 0.9093 0.9086 0.0927 0.9998 0.0043 1.0000 0.0002
1.0000 0.0000 1.0000 0.0000 1.0000 = 7.9093 8.9086 7.0927 11.9998 12.0043 10.0000 5.0002 10.0000 6.0000
7.0000 3.0000 5.0000

B4= 9 3 14 2 4 15 7 15 3 8 2 2

+ PE4= 0.1411 0.7983 0.1388 0.9996 0.0065 1.0000 0.0003 1.0000 0.0000 1.0000 0.0000 1.0000

= 9.1411 3.7983 14.1388 2.9996 4.0065 16.0000 7.0003 16.0000 3.0000 9.0000 2.0000 3.0000

Swin Transformer Block

Layer Norm $LN(z) = \frac{z - \mu}{\delta} \circ \gamma + \beta$

B1= 1 6 7 3 6 9 9 4 2 11 14 11 $\mu = \text{mean} = 6.9167$, $\delta = \text{standard deviation} = 4.0104$, $\gamma = 1$, $\beta = 0$

LN (B1) = -1.4753 -0.2286 0.0208 -0.9766 -0.2286 0.5195 0.5195 -0.7273 -1.2260 1.0182 1.7662 1.0182

B2= 3.8415 7.9769 9.0464 5.0000 8.0022 5.0000 1.0001 6.0000 4.0000 13.0000 16.0000 2.0000 $\mu = \text{mean} = 6.7389$, $\delta = \text{standard deviation} = 4.3829$, $\gamma = 1$, $\beta = 0$

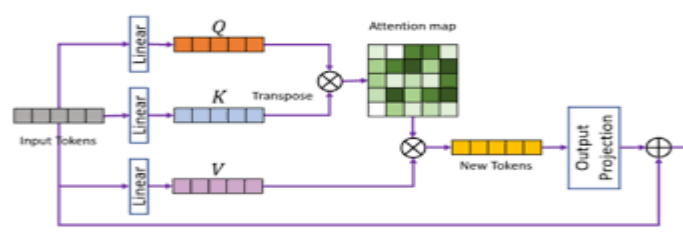
LN (B2) = -0.6611 0.2825 0.5265 -0.3968 0.2882 -0.3968 -1.3094 -0.1686 -0.6249 1.4285 2.1130 -1.0812

B3= [7.9093 8.9086 7.0927 11.9998 12.0043 10.0000 5.0002 10.0000 6.0000 7.0000 3.0000 5.0000] $\mu = \text{mean} = 7.8262$, $\delta = \text{standard deviation} = 2.8497$, $\gamma = 1$, $\beta = 0$

LN (B3) = 0.0291 0.3798 -0.2574 1.4646 1.4662 0.7628 -0.9917 0.7628 -0.6409 -0.2899 -1.6936 -0.9918

B4= [9.1411 3.7983 14.1388 2.9996 4.0065 16.0000 7.0003 16.0000 3.0000 9.0000 2.0000 3.0000] $\mu = \text{mean} = 7.5070$, $\delta = \text{standard deviation} = 5.3144$, $\gamma = 1$, $\beta = 0$

LN (B4) = 0.3075 -0.6979 1.2479 -0.8482 -0.6587 1.5981 -0.0954 1.5981 -0.8481 0.2809 -1.0362 -0.8481



B1Q= B1*WQ, LN (B1) = -1.4753 -0.2286 0.0208 -0.9766 -0.2286 0.5195 0.5195 -0.7273 -1.2260 1.0182 1.7662 1.0182

WQ=

1.5270	-0.1623	-1.0642	-0.9480	0.1825	2.0237	0.0662	0.9111	1.0205	0.8577	-1.2706	-2.1321
0.4669	-0.1461	1.6035	-0.7411	-1.5651	-2.2584	0.6524	0.5946	0.8617	-0.6912	-0.3826	1.1454
-0.2097	-0.5320	1.2347	-0.5078	-0.0845	2.2294	0.3271	0.3502	0.0012	0.4494	0.6487	-0.6291
0.6252	1.6821	-0.2296	-0.3206	1.6039	0.3376	1.0826	1.2503	-0.0708	0.1006	0.8257	-1.2038
0.1832	-0.8757	-1.5062	0.0125	0.0983	1.0001	1.0061	0.9298	-2.4863	0.8261	-1.0149	-0.2539
-1.0298	-0.4838	-0.4446	-3.0292	0.0414	-1.6642	-0.6509	0.2398	0.5812	0.5362	-0.4711	-1.4286
0.9492	-0.7120	-0.1559	-0.4570	-0.7342	-0.5900	0.2571	-0.6904	-2.1924	0.8979	0.1370	-0.0209
0.3071	-1.1742	0.2761	1.2424	-0.0308	-0.2781	-0.9444	-0.6516	-2.3193	-0.1319	-0.2919	-0.5607
0.1352	-0.1922	-0.2612	-1.0667	0.2323	0.4227	-1.3218	1.1921	0.0799	-0.1472	0.3018	2.1778



1.6742 -1.0329 -0.4619 0.0032 -1.0649 0.6737 -2.5644 -0.4807 1.0989 -0.7158 0.0675 1.9278
0.1250 -0.3233 0.8836 0.3656 -1.7684 -0.6691 0.4659 0.8368 0.1472 -0.2805 -0.1871 -0.1762
0.5301 0.7665 0.4359 3.5267 -0.4229 -0.4003 1.8536 2.5383 2.2957 1.1665 0.2917 -0.2438
-0.9521 1.7447 0.8967 -0.1124 -1.0531 -0.6718 1.0393 -1.3233 2.7526 1.2128 0.9877 -0.8976
0.8540 -1.1605 0.5047 -1.5566 0.6478 0.5756 0.9109 0.1283 0.1383 0.4855 0.3929 -0.7923
0.3891 2.3774 -0.4009 1.9151 -0.3176 -0.7781 -0.2397 -1.4424 -1.9071 1.0260 0.1946 -0.9530
-1.1560 1.5261 -0.5138 0.6098 1.7690 -1.0636 0.1810 1.3025 -0.3650 0.8707 0.2798 0.3539
0.0397 0.1685 0.7964 -0.6479 1.5106 0.5530 0.2442 1.4099 -0.8481 -0.3818 0.0512 1.5970
B2k=[-3.3516 5.9687 -3.3248 2.0356 1.2611 -2.3460 -4.2297 -4.8985 -4.8895 0.9802 -0.9803 0.9389]

$$B2v=B2*Wv, \quad Wv=$$

0.5275 -0.9606 -1.7423 -0.3020 0.7449 -0.3086 0.2360 2.3652 -0.6300 -2.2751 1.3419 -0.5817
0.8542 -1.6338 0.2053 1.8136 -0.8282 0.4567 -0.8352 -0.4822 -0.0469 -1.6333 -0.9884 -1.8301
1.3418 0.7612 1.1929 0.9149 0.5745 -0.2751 -1.2760 0.6474 2.6830 0.4155 1.8179 -0.4491
-2.4995 1.1933 -0.8028 -0.0571 0.2818 0.4431 0.6170 -1.0344 -1.1467 -0.6548 -0.3744 0.9493
-0.1676 1.6321 -1.2656 1.3094 1.1393 -0.1348 0.6127 1.3396 0.5530 -0.2963 -1.4517 0.7174
0.3530 -1.5322 -0.1493 -1.0447 -0.4259 -0.0183 0.2894 -0.9691 -1.0765 -1.4969 -0.6187 2.2878
0.7173 -1.3369 -1.6364 -0.3483 0.6361 0.4608 0.3953 0.2087 1.0306 -0.9048 0.9345 0.1667
-1.3049 -1.4738 0.0173 1.4126 0.7932 1.3623 -0.8706 -0.6186 0.3275 -0.4042 1.0559 -2.1565
-1.0059 -0.0417 0.8284 1.5024 -0.8984 0.4519 -0.4977 0.5120 0.6521 -0.7258 0.1602 1.6894
0.7907 -0.6155 0.2177 0.7304 0.1562 1.6484 -0.1067 0.0114 -0.2789 -0.8665 0.2874 1.2823
-0.1166 1.3142 -1.9092 0.4908 1.5973 -2.0284 -0.6878 -0.0440 0.2452 -0.4218 0.6329 -0.5826
0.5531 -1.4551 -0.5368 -0.5861 0.1124 -0.4493 0.3319 2.9491 1.4725 -0.9427 -1.4590 0.2226
B2V=[1.5971 6.6749 0.3304 4.0013 3.0329 -2.5801 -3.2713 -3.9322 -0.4272 2.6264 1.5901 -1.9971]

B3Q=B3*WQ, LN (B3) = [0.029 0.3798 -0.257 1.4646 1.4662 0.7628 -0.9917 0.7628 -0.6409 -0.2899 -1.6936
-0.9918]

$$WQ=$$

0.7795 -0.0910 0.8751 1.7710 0.2719 1.3373 0.4340 1.2033 -0.2192 0.6201 -0.6305 -0.5708
0.3847 -0.2528 1.3954 0.2213 1.4896 2.1257 0.5201 0.5220 -0.8798 -0.2867 0.6096 0.4942
0.6964 1.1948 0.3210 2.7304 1.4371 0.0540 -1.0922 0.3970 -0.3208 0.5980 0.7823 0.9914
-0.1127 0.6061 1.6234 -0.2962 -0.0276 0.1630 -0.2258 -0.4828 -0.7844 -0.2455 2.4366 1.0771
-0.0388 0.5405 1.0624 0.5643 0.9239 -0.6327 -0.4049 -0.2315 -0.3650 -1.7807 0.3024 0.7768
0.0881 -1.4449 0.2141 1.5826 -0.3213 1.6120 0.5279 0.6134 0.1173 -2.3472 0.0583 -2.2598
-0.7897 -0.9677 0.8768 2.7292 0.6611 -0.0754 -1.0070 1.6829 0.1743 -1.7136 -0.5741 -0.5644
1.4230 0.2021 0.1944 0.3036 1.9153 -0.4732 1.0890 0.5684 -0.2157 -0.2371 -0.1952 0.9015
0.0063 -0.3479 -0.4149 -0.7903 0.1568 2.1842 1.7849 -1.2060 -0.1526 -0.6196 -0.0505 0.3947
0.6865 1.2901 0.3585 0.8034 -0.3005 0.8099 -0.3038 0.4331 0.0337 -0.7202 -1.7558 0.0049
-0.8549 1.3412 0.0362 -1.3199 -0.5000 0.7163 -0.0087 -0.0921 0.4583 0.0407 -0.2574 0.4369
-1.0752 -0.5808 -0.3646 -0.2738 0.7165 -1.0056 0.5066 -0.2441 1.2816 -0.6590 0.7495 1.1301
B3Q= [4.0145 -0.5609 4.3126 1.3392 2.2012 -0.7636 0.2553 -0.6374 -4.1491 -2.2967 4.7226 0.0409]

$$B3k=B3*Wk, \quad Wk=$$

0.1538 0.1789 1.7553 -2.2220 0.1812 2.0099 0.7742 0.3606 -0.8595 -0.3918 -0.0460 -1.6848
-0.7586 0.9276 0.9315 0.4488 0.0543 1.4167 0.2657 1.8851 -1.4130 -1.4847 -0.4657 0.4131
-0.1802 -0.1102 0.8253 0.0006 0.6878 0.0115 -0.2351 -0.3068 -0.5160 -0.3446 0.0759 0.5017
-0.2078 1.5724 -0.8148 -0.7562 -1.3934 -0.9390 1.8773 -1.0127 -1.0851 -1.3239 -0.9188 0.0831
0.8967 0.5605 -0.5342 0.4043 1.4254 -1.7389 0.6092 -0.2385 -0.7072 -0.5748 -1.9195 0.1578
0.4123 -0.4203 0.2426 -0.7939 -0.8939 0.0170 -0.1103 -0.4840 -0.3017 -0.7407 -0.0364 -0.5279
0.5475 -0.1539 -0.1006 0.8598 0.0378 0.2192 0.2775 -0.3273 -1.2004 -1.1850 -1.2251 0.7231
0.1478 -0.2752 -1.6250 0.0669 -0.3636 1.0458 0.0870 0.4755 -0.0194 -2.0649 -1.9024 -0.8499
-0.3623 0.2411 -1.5144 -1.6394 0.1496 -0.9510 0.1736 -0.1300 0.3404 0.5651 2.3742 -0.7964
0.0611 0.7547 1.0262 -2.4247 -1.9445 0.7948 0.4536 -0.5944 -0.9632 2.0001 -0.2333 0.7253
0.2167 -0.2919 -0.7581 -0.2838 1.5239 0.0714 -0.6035 -0.4438 1.1139 2.2233 0.4039 1.6865
-1.3981 0.4584 2.0783 1.1458 0.5458 -0.7737 -0.7970 -1.3225 -1.5861 -0.4922 1.1925 -0.3864
B3K=[1.8916 2.7991 -2.8432 -0.7184 -3.7530 -1.7125 5.1037 1.6095 -2.3617 -8.4526 -7.9425 -3.6093]

$$B3v=B3*Wv, \quad Wv=$$

-0.5051	-0.4949	3.5699	0.0053	0.8254	-1.4150	-0.6607	-0.5352	0.4121	-0.3936	-1.2203	0.3535	
0.4081	-0.6248	3.4075	1.1391	-0.2293	-0.0334	0.3276	-0.2620	-0.9371	-2.1763	-0.3762	-1.2568	
1.0724	-0.9336	1.1469	0.4274	0.2623	0.2611	1.5081	-1.0954	0.8352	-1.1957	0.0524	-1.0360	
0.9680	-0.2787	0.7873	0.1810	-0.8980	0.5230	-0.8889	0.5492	1.0945	0.3653	-1.9539	-0.4270	
0.2701	-0.2005	-1.2781	0.6575	-2.1573	0.6768	-0.9593	-2.1746	0.3212	2.4124	0.5654	0.0777	
-0.6344	0.1367	-0.5846	0.5842	0.0939	0.6744	-0.4927	0.1446	0.0151	0.7402	0.0778	1.7909	
0.6298	-0.8833	-0.6194	-1.6165	-0.9512	0.6784	0.4411	0.6690	-1.1645	0.3097	0.0807	0.7973	
-0.0792	0.0825	0.5353	0.0188	1.1726	-0.3126	-0.2048	-0.8488	0.4948	-0.5420	0.7947	-0.4425	
1.3769	-0.6039	0.4830	-0.4260	1.7351	0.3792	-0.9778	-0.0571	-1.2158	0.1223	1.0437	-0.6291	
-1.4082	-0.3687	0.1085	-2.0340	-0.3717	-2.2529	-1.5729	0.2918	1.5908	0.9623	0.8022	1.5338	
0.1412	-0.8382	-0.3403	-1.3286	1.1929	-1.1684	-0.3703	-0.6146	0.5265	-0.2540	-0.4688	2.7335	
1.2897	-0.2825	-0.9258	-0.3199	0.9536	1.2020	-0.2211	-0.0136	0.2648	1.8918	-0.0466	0.1674	
B3v=[	-1.4834	2.5233	2.1121	7.0449	-6.6697	2.4371	-2.0318	-2.4113	2.2227	1.5825	-1.7003	-5.3109]
B4Q=B3*WQ, LN (B4) = 0.3075	-0.6979	1.2479	-0.8482	-0.6587	1.5981	-0.0954	1.5981	-0.8481	0.2809	-1.0362		
	-0.8481,	WQ=										
-0.4789	0.9940	-1.5438	2.7485	0.5447	-1.9201	-0.8127	0.2195	-0.9890	-0.2631	0.7151	-2.0676	
-1.4127	0.4284	-1.5518	-1.5130	-0.1379	0.6254	-0.4384	-0.1149	-1.5159	1.0821	1.0486	-1.1546	
0.9750	-1.0042	0.8671	0.4340	0.6199	0.7530	0.8586	0.0686	-2.2285	0.9824	0.1775	-0.1072	
0.2897	-0.6327	-0.1454	-0.2298	-0.0056	0.2135	0.1952	0.7515	-0.1515	-0.1114	0.1533	-1.7129	
1.1654	-1.0457	-0.3862	-0.8271	1.1072	-0.7702	0.8889	-0.6894	1.1627	1.2960	-1.2543	-0.4599	
-0.9087	0.5057	1.3162	-0.8320	-0.1856	-0.0071	0.0692	0.4508	-0.1819	1.7813	-1.1728	1.0910	
0.2406	0.3320	-0.7965	0.4979	-1.1214	0.0932	2.4868	-1.5650	-0.2340	-0.0656	-1.4614	-0.1639	
-0.6540	-1.6354	0.1354	2.3156	0.2464	0.9353	-1.6656	-0.0788	-1.0467	-0.2557	-1.2443	-0.0672	
1.2750	-1.9068	0.4178	-0.7938	1.5610	0.6635	-0.4159	-0.9418	1.5833	0.4046	-0.1551	-1.0016	
-0.5512	-0.0404	0.8199	0.5410	-1.1966	-0.3502	-0.0842	-0.6529	-0.2494	0.3518	-1.6149	-0.5557	
-0.0874	0.6972	-0.8544	-0.5591	-0.2423	1.6199	0.0893	0.2825	1.2940	0.8075	2.2461	-1.3516	
-0.3478	0.1688	0.3573	1.9766	1.0048	-0.0508	1.4561	-1.1251	0.3620	-0.2556	0.7393	0.3642	
B4Q=[	-2.3290	-1.1178	4.9221	5.2340	-1.7442	-0.5832	-3.4104	2.0706	-7.6665	1.2107	-6.5939	5.2287]
B4K=B3*WK, LN (B4) = 0.3075	-0.6979	1.2479	-0.8482	-0.6587	1.5981	-0.0954	1.5981	-0.8481	0.2809	-1.0362		
	-0.8481											
	WK=											
-0.4522	0.8639	-1.5180	-1.5163	-1.1821	0.9143	-0.5278	-0.2228	-0.6888	0.1056	-0.5051	-0.4484	
0.7834	-1.4169	-1.0749	-0.0683	0.5838	1.1077	1.2416	0.2721	-0.8568	1.1284	-0.4760	0.1684	
1.2372	0.2000	-3.0722	0.7824	-0.5800	0.8205	-0.1576	-1.1685	0.0484	0.7425	-2.0516	-1.1205	
1.0772	0.1958	0.5214	-1.4207	-0.5614	-0.8176	-1.3736	-1.2242	-0.6649	1.1436	-0.4483	0.4007	
-0.0348	-0.1511	-0.9910	0.6694	-1.5559	-0.1265	0.8708	-2.0995	1.4527	-0.9147	-1.5512	0.7391	
-0.5040	0.8928	-0.2531	0.6832	1.1002	0.2641	-1.5685	-0.3902	1.3798	0.1798	0.9298	0.9000	
-1.2240	-0.3642	1.0093	-0.8821	0.1751	3.1585	-1.8443	0.6643	0.0951	-0.9833	0.9019	-1.5312	
0.1185	-0.8262	0.0510	-1.5045	1.0036	1.2266	0.2884	-0.7023	-0.4271	0.3848	0.1383	0.5046	
0.9595	0.2760	-0.4404	0.4331	1.5110	2.3206	-0.9509	0.5013	0.5108	0.3257	-0.3784	-0.8642	
-0.3409	0.4598	-0.8485	0.8083	-1.1372	0.4145	-0.9107	0.5403	-0.6563	1.2963	0.1431	-0.3766	
-0.1733	-0.2175	-0.2404	0.5789	0.6430	0.2118	-0.1619	0.9908	-0.1250	1.0992	1.6050	0.7880	
0.6163	0.7961	0.6029	0.7628	-0.0128	0.6132	-0.4889	0.9894	-0.5305	0.6532	1.3491	0.2982	
B4k=	-1.7844	1.0239	-3.8862	-1.2935	1.0953	0.7986	-1.3708	-3.2437	1.5289	-0.8049	-1.8065	-0.5318
B4V=B3*WV , LN (B4) = 0.3075	-0.6979	1.2479	-0.8482	-0.6587	1.5981	-0.0954	1.5981	-0.8481	0.2809	-1.0362		
	-0.8481											
	WV=											
-0.1637	1.0996	1.0666	-0.2676	0.4735	-0.1299	1.2481	-0.1325	0.7636	0.9463	-1.5485	-0.6271	
0.6067	-0.8556	-2.0992	0.1866	1.3656	0.7337	2.8092	1.4393	1.1274	0.8182	1.8632	0.4402	
1.6345	0.0076	0.6385	0.9509	-1.6378	0.1203	-0.2322	2.0313	0.1416	1.5890	0.1340	-1.5026	
-0.6235	-0.9376	0.3715	-0.7905	2.0237	1.1363	0.2870	-0.9782	1.9281	0.5260	-1.5460	-0.2082	
-1.3501	-0.6816	-0.3742	-0.4895	0.7778	-0.6868	-0.4646	-0.6410	0.6706	-2.4652	0.4333	-1.5051	
-1.1622	-0.2601	0.6953	2.9745	-0.5489	0.4717	0.3843	0.5205	-0.1101	-0.8525	0.1030	1.8097	



-0.9443 -0.2288 0.8776 -0.6226 -0.1260 0.2883 -0.3798 -0.3807 0.3710 0.5117 -0.5703 -0.1170
-0.6712 -0.5248 1.0336 1.9203 0.2996 1.3919 -0.1018 1.9929 1.0966 0.2578 0.4931 1.2262
0.5767 1.1283 0.4198 0.9611 0.2962 -1.3455 1.6182 1.6359 -0.3982 1.9624 -0.7075 0.4949
-2.0858 0.5501 0.6011 -0.5578 1.2008 0.0007 -0.8472 0.5590 -0.0855 1.4063 -1.3480 0.9086
0.2360 1.8551 -0.6740 -0.1066 1.0902 0.0535 -0.5759 0.5636 2.3733 0.4968 -1.7543 -1.1542
-0.7784 -0.2773 -1.0952 -0.2152 -0.3587 -2.3419 -0.0759 -0.3384 -0.4739 0.0828 -0.3638 1.2576
B4v=[-0.5152 -1.5333 6.6406 9.1699 -6.2059 5.1094 -2.2654 5.2670 -2.6550 0.0269 2.7714 3.6207]

$$f1=B1K'*B1Q+B2k'*B1Q+B3K'*B1Q+B4k'*B1Q;$$

21.2019 -9.0369 -9.1473 -12.4043 16.0557 40.7409 -9.0605 48.3425 -0.8501 22.3347 2.1428 10.6174
-43.0607 18.3539 18.5780 25.1930 -32.6088 -82.7441 18.4017 -98.1828 1.7265 -45.3614 -4.3520 -21.5637
24.0723 -10.2604 -10.3857 -14.0836 18.2293 46.2565 -10.2871 54.8871 -0.9651 25.3584 2.4329 12.0548
1.5336 -0.6537 -0.6617 -0.8973 1.1614 2.9470 -0.6554 3.4969 -0.0615 1.6156 0.1550 0.7680
-4.1834 1.7831 1.8049 2.4475 -3.1680 -8.0387 1.7878 -9.5386 0.1677 -4.4069 -0.4228 -2.0950
15.1012 -6.4366 -6.5152 -8.8350 11.4357 29.0180 -6.4534 34.4322 -0.6055 15.9080 1.5262 7.5623
-15.0263 6.4047 6.4829 8.7912 -11.3790 -28.8741 6.4214 -34.2615 0.6025 -15.8291 -1.5187 -7.5248
6.9265 -2.9523 -2.9884 -4.0524 5.2453 13.3098 -2.9600 15.7932 -0.2777 7.2966 0.7000 3.4686
15.9144 -6.7832 -6.8660 -9.3108 12.0515 30.5805 -6.8009 36.2863 -0.6381 16.7646 1.6084 7.9695
27.0876 -11.5456 -11.6866 -15.8478 20.5127 52.0506 -11.5757 61.7624 -1.0860 28.5348 2.7377 13.5648
30.5710 -13.0303 -13.1895 -17.8858 23.1506 58.7443 -13.0643 69.7050 -1.2257 32.2044 3.0897 15.3092
-1.2221 0.5209 0.5273 0.7150 -0.9255 -2.3484 0.5223 -2.7866 0.0490 -1.2874 -0.1235 -0.6120

$$ff1=\text{softmax}(f1)$$

0.0001 0.0000 0.0000 0.0000 0.0008 0.0000 0.0000 0.0000 0.0308 0.0001 0.1186 0.0075
0.0000 1.0000 1.0000 1.0000 0.0000 0.0000 1.0000 0.0000 0.4046 0.0000 0.0002 0.0000
0.0015 0.0000 0.0000 0.0000 0.0068 0.0000 0.0000 0.0000 0.0274 0.0010 0.1585 0.0315
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0677 0.0000 0.0162 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0851 0.0000 0.0091 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0393 0.0000 0.0640 0.0004
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1315 0.0000 0.0030 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0545 0.0000 0.0280 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0380 0.0000 0.0695 0.0005
0.0297 0.0000 0.0000 0.0000 0.0662 0.0012 0.0000 0.0004 0.0243 0.0248 0.2149 0.1428
0.9687 0.0000 0.0000 0.0000 0.9262 0.9988 0.0000 0.9996 0.0211 0.9741 0.3056 0.8173
0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0756 0.0000 0.0123 0.0000

$$C1=B1V*ff1+B2V*ff1+B3v*ff1+B4v*ff1$$

C1=5.3631 7.7881 7.7881 7.7882 5.3854 5.3463 7.7881 5.3458 2.2473 5.3601 4.3198 5.4166

$$B1=[1 \ 6 \ 7 \ 3 \ 6 \ 9 \ 9 \ 4 \ 2 \ 11 \ 14 \ 11]$$

C1= [5.3631 7.7881 7.7881 7.7882 5.3854 5.3463 7.7881 5.3458 2.2473 5.3601 4.3198 5.4166]

Z=C1+B1= [6.3631 13.7881 14.7881 10.7882 11.3854 14.3463 16.7881 9.3458 4.2473 16.3601 18.3198 16.4166]

LN (Z) = [6.3631 13.7881 14.7881 10.7882 11.3854 14.3463 16.7881 9.3458 4.2473 16.3601 18.3198 16.4166]

$$MLP, W=$$

-0.4264 -0.3119 0.6477 -0.7091 0.1152 -0.0670 1.2828 -2.7918 0.3457 -0.8132 -0.2164 0.0867
0.8896 0.2214 -0.3035 0.0293 -0.6282 -0.7173 -0.4749 0.6022 -0.3447 0.1320 -0.3161 0.5003
0.6498 -0.2973 -0.2268 0.5504 0.3876 0.4173 0.6234 0.0404 0.0308 -1.1847 1.3853 -0.3630
-1.8062 -1.6972 -0.8716 -0.7255 3.0743 1.1246 -0.3477 -1.0309 -0.1333 0.3404 -0.3483 -0.3507
1.0247 -0.5987 1.1258 0.0008 0.0976 0.6826 -0.7583 -1.3010 -0.6260 -1.5921 1.2055 -1.3629
0.8252 -0.5233 0.9916 -0.1875 -0.0456 1.3961 -0.7372 -1.3180 1.1114 -2.4538 1.2372 1.7228
-0.2363 -1.7158 0.6864 1.2414 0.8583 -1.3886 -1.4510 -0.2785 1.2022 -0.2081 2.5197 0.2014
-0.9813 0.8857 0.5313 0.4725 -1.5636 0.0193 1.5208 -2.0221 0.0724 -0.9819 -0.3795 -0.5477
-0.7696 -2.1666 0.7264 1.0945 1.1282 -1.3504 -2.4472 0.4067 -1.0267 1.5311 -0.4753 0.4018
-1.2361 0.4221 0.6943 0.6227 -0.8476 0.1146 -1.2900 -0.0898 -1.1730 1.3727 1.6757 0.8344
-0.0502 -0.2948 0.3941 1.0092 0.5675 0.5416 -1.9846 -0.3423 0.2510 -0.6121 -0.7384 -1.0483
-0.3887 -0.3643 0.6316 -0.0945 0.2766 -0.7677 -1.3429 0.6522 -0.2646 -0.9667 -1.8155 0.2253

$b=0$, $MLP = Z*W+b$ [-20.7486 -70.1690 62.7434 50.5536 37.0841 6.1162 -112.17 -72.583 2.8434 -81.32 63.28 5.6099]

$zz=z+mlp$ [-14.3855 -56.3809 77.5315 61.3418 48.4695 20.4625 -95.3881 -63.2376 7.0907 -64.9645 81.6089 22.0265]

-14.3855 -56.3809 77.5315 61.3418
48.4695 20.4625 -95.3881 -63.2376
7.0907 -64.9645 81.6089 22.0265

Sw-MSA, $S=[20.4625 -95.3881 -63.2376 48.4695 -64.9645 81.6089 22.0265 7.0907 -56.3809 77.5315 61.3418 -14.3855]$ $zzz=zz+s$, $zzz=[20.4625 -95.3881 -63.2376 48.4695 -64.9645 81.6089 22.0265 7.0907 -56.3809 77.5315 61.3418 -14.3855]$ MLP , $zmlp= zzz*W+b$, $W=$

-0.4824	-0.1611	-0.3412	1.6937	1.2849	-0.7249	1.9352	0.7122	-1.0676	-0.5317	1.6723	1.2721
1.6507	-0.6578	-0.6334	-0.3315	1.0007	-0.4653	0.2617	0.1356	-0.4535	-0.0473	0.8678	0.6678
0.1589	0.7068	0.3947	-1.1501	1.0734	-0.3870	1.7792	-0.9469	0.4582	-1.0338	0.4381	-0.9108
-1.1247	1.5403	-0.8184	-0.9907	1.8940	-0.6620	1.8927	-0.8469	1.0218	0.6037	0.8266	-1.0527
-0.1111	-0.0219	-1.3617	0.3642	1.5344	0.8699	-0.6203	0.2923	-0.4591	0.2492	0.3895	0.4272
0.4491	1.0548	-0.3021	-0.4203	0.1756	-0.1871	-1.1907	-2.1843	-0.0046	0.4415	0.9318	-0.6888
2.1006	-0.6968	1.5992	0.9836	-0.4761	1.8515	0.8582	-1.1744	1.1859	-0.0269	1.2058	-1.1983
0.0396	0.1593	-0.3923	-0.2758	-0.2797	-0.0061	-0.8151	-1.5658	0.2237	-0.7421	-1.0213	-0.2475
0.5716	-0.1765	0.2208	-2.0855	0.7026	1.8663	-0.1572	-0.3279	0.0804	-0.0853	-0.1910	-0.8447
1.2007	1.6303	-0.5937	-0.4395	1.2082	-2.4300	-0.5712	0.4852	-0.3311	-1.2841	-0.0272	0.8496
0.5088	0.2940	-2.0504	0.7119	-0.2476	0.3618	0.1008	2.3648	-0.3846	-0.3386	-0.5551	2.3066
-0.8698	0.7677	0.0004	1.0265	1.5127	0.2835	0.5964	-0.0041	1.0580	2.3072	-0.9942	-1.7010

$Zmlp=-36.8920$ 306.0381 -99.2044 165.1588 -125.9542 -284.6081 -87.7064 -12.5268 30.1080 -46.4483 22.7705 136.2421

$Zend= Zmlp+zzz$, $=-36.892$ 306.038 -99.204 165.158 -125.954 -284.608 -87.7064 -12.5268 30.108 -46.4483 22.77 136.2421