A Transformer-based Approach for aAnomaly Detection in Wire eElectrical Discharge

W. Hammed[1*], A. H. Al-Rubaye[2], B. S. Bashar[3], M. k. Imran[4], M. Gh. Rzooki[5], A. M. Hashesh[6]
1- Medical technical college, Al-Farahidi University, Baghdad, Iraq.
Email: waleed.hammed@gmail.com (Corresponding author)
2- Department of Petroleum Engineering, Al-Kitab University, Altun Kupri, Iraq.
Email: amir.hazim@uoalkitab.edu.id
3- Al-Nisour University College, Baghdad, Iraq.
Email: bashar.s.eng@nuc.edu.iq
4- Building and Construction Engineering Technology Department, AL-Mustaqbal University College, Hillah 51001, Iraq.
Email: Merzah.kareem@Mustaqbal-college.edu.iq
5- Medical Device Engineering, Ashur University College, Baghdad, Iraq.
Email: mustafa.ghanim@au.edu.iq
6- Al-Hadi University College, Baghdad,10011, Iraq.
Email: ali.mohammed@huc.edu.iq

Received: 8 August 2022 Revised: 12 October 2022 Accepted: 5 November 2022

ABSTRACT:
Although theoretical models of manufacturing processes are useful for understanding physical events, it can be challenging to apply them in real-world industrial settings. When huge data are accessible, artificial intelligence approaches in the context of Industry 4.0 can offer effective answers to real production challenges. Deep learning is increasingly being used in the realm of artificial intelligence to address a variety of issues relating to information and communication technology, but it is still limited or perhaps nonexistent in the industrial sector. In this study, wire electrical discharge machining—a sophisticated machining technique primarily used for computer hardware components—is applied to effectively forecast unforeseen occurrences. By identifying hidden patterns in process signals, anomalies, such as changes in the thickness of a machined item, may be efficiently anticipated before they occur. In this study, a model for anomaly detection in the sequence of thickness change in the machined component based on transformers is suggested. Our method is able to achieve 94.32 % and 94.16 % accuracy in Z 135 and Z 15 datasets, respectively. Also, it forecasts the abnormalities inside the sequence 1.1 seconds in advance, according to our tests on a dataset that has been introduced.

KEYWORDS: Anomaly Detection, Transformers, Wire Electrical Discharge, Anomaly Detection.

1. INTRODUCTION

When it comes to process optimization, industrial machinery and manufacturing sectors generally have a history of depending on empirical methods [1]. The actual implementation of theoretical models is challenging due to the numerous occurrences and variables that are involved in each activity [2]. In actuality, despite the fact that theoretical models are highly valuable for comprehending the underlying physical processes, they frequently have significant drawbacks for practical use in industry. This reality is especially clear when it comes to producing parts for industries with significant added values, like the production of airplanes [3]. Recent years have seen an exponential growth in the aerospace sector. By 2032, it's anticipated that there will be twice as many airplanes in the world [4]. To adapt their goods to the rising tolerance and accuracy standards expected by this industry, manufacturing businesses have made significant investments as a result of this development. Wire electrical discharge machining has received a lot of attention as non-traditional machining techniques have become more popular [5]. Using this technique, significantly rigid materials may be processed with incredibly tight precision and an outstanding surface polish. However, because theoretical models are only partially accurate, trial and error methods are still necessary for increasing efficiency. If enormous volumes of data can be gathered from the process, artificial intelligence and more especially deep learning

Paper type: Research paper
DOI: 10.30486/mjee.2022.696524
How to cite this paper: W. Hammed, A. H. Al-Rubaye, B. S. Bashar, M. k. Imran, M. Gh. Rzooki and A. M. Hashesh, “A Transformer-based Approach for aAnomaly Detection in Wire eElectrical Discharge”, Majlesi Journal of Electrical Engineering, Vol. 16, No. 4, pp. 131-138, 2022.

techniques seem to be an intriguing solution in this situation [6].

In very challenging learning tasks including image identification [7], handwriting recognition [8], natural language processing [9], picture description [10], and industrial applications [11-12-13], deep learning utilizing Deep Neural Networks (DNNs) has produced amazing state-of-the-art results. In a DNN, a series of hidden layers extract abstract information from a sequence of pictures, in contrast to shallow neural networks. This has led to an enormous amount of new information and communication technologies applications being created, such as speech recognition [14] and machine translation [15], which has increased attention from both academics and business. The paragraphs that follow provide a quick description of the primary network designs for deep learning. Recurrent neural networks (RNNs) [16] are a popular method in many domains for processing sequential input. Rumelhart et al. [17] used back-propagation across time in their initial attempts to train RNNs. Later, Elman et al. [18] presented the Elman network with feedback from the hidden layer's output to its input. Due to disappearing and ballooning gradients, many training techniques and architectures are unable to handle long-term temporal dependencies. Thus, in 1997 Hochreiter & Schmidhuber [19] developed the long short-term memory networks to address the vanishing gradients problem (LSTMs). An LSTM employs gates instead of the traditional RNN to determine whether or not to maintain the current memory.

Convolutional Neural Networks (CNNs) do remarkably well at extracting features from data sequences [20]. Local receptive fields, subsampling, and shared weights are three concepts that are combined in CNNs, which are feed-forward neural networks. The neural network model put forward by Fukushima already included the concepts of local receptive fields and subsampling. Neurons may extract features from pictures (2D structures), sequences, or time series by utilizing CNNs with local receptive fields (1D structures). Multiple futures can be derived from a convolutional layer using various future maps. CNNs may also recognize higher-order characteristics by merging these features in the next layers [21].

DNNs have historically been employed in defect diagnosis for a variety of industries when considering industrial applications beyond the scope of ICTs. For instance, Yin et al. [22] offered an innovative approach to the manual defect diagnostic process currently used in high-speed railroads. The authors specifically suggested an automated diagnosis network to find equipment problems in vehicle on-board systems. The findings demonstrate that a deep belief network outperforms other trained networks and increases fault diagnostic accuracy by up to 95%. The choice of several methods for enhancing failure diagnostics in rolling bearings is another example.

DNNs can only be trained effectively if a large amount of labeled data is available to use back-propagation training procedures, which is frequently not feasible in industrial contexts [23]. However, there are several intriguing methods in the scientific literature. The majority of published research focuses on employing shallow neural networks to optimize process parameters in advanced machining processes. The application of deep learning in machining has only been the subject of a relatively small number of research. Wang et al. [24] created a deep learning-based approach to material removal rate prediction in polishing technologies in a very intriguing recent study. In order to create an intelligent laser welding equipment, Gunter has also built a pattern recognition, identification, and process control system.

Encoders and decoders, both of which are made up of modules that can talk repeatedly over top of one another, are used to construct transformers. Since we cannot utilize this directly, the inputs and outputs are first embedded into n-dimension space [25]. Therefore, it goes without saying that we must encrypt whatever inputs we provide. The positioning and coding of various words is a small but significant component of this paradigm.

Natural Language Processing (NLP), computer vision, and speech processing have all embraced Transformer, a well-known deep learning model. Transformer was first put up as a machine translation sequence-to-sequence paradigm [26]. Later studies demonstrate that pre-trained models based on Transformers can perform at the cutting edge on a variety of tasks. Transformer has thus become the preferred architecture in NLP. Transformer has been used in computer vision, audio processing, and even other fields including chemistry and the natural sciences in addition to activities relating to language [27].

A stack of L identical blocks makes up each encoder and decoder in the basic Transformer, which is a sequence-to-sequence model. A position-wise feed-forward network and a multi-head self-attention module make up the majority of each encoder block. A residual connection is used around each module to help develop a deeper model, which is then followed by the Layer Normalization module. Decoder blocks, in contrast to encoder blocks, also insert cross-attention modules between the position-wise feed-forward networks and the multi-head self-attention modules [28]. Additionally, the decoder's self-attention modules have been modified to stop each position from paying attention to later positions.

In this study, we propose a transformer-based algorithm that is able to detect anomalous data in wire electrical discharge. Our approach is applied on sequential data and utilizes the mechanism of multi-head self-attention in order to find the rare events in the electrical discharge load. The approach is validated against all collected samples from real-world situations in industrial context. Our contributions are as follows:

1. A transformer-based approach for detecting anomalous behavior in electrical discharge load.
2. A dataset is introduced which is collected from industrial facilities.
3. Our approach is able to reach 94.32% and 94.16% accuracy.

2. MATERIALS AND METHODS

2.1. Transformers

One of the most notable recent developments in deep learning and deep neural networks is the transformer model. It is mostly utilized for sophisticated natural language processing applications. It is being utilized by Google to improve search engine outcomes. Transformers were employed by OpenAI to develop its well-known GPT-2 and GPT-3 models [29].

The transformer architecture has developed and branched out into several forms since its introduction in 2017, going beyond language problems to other domains [30]. They have been applied to forecast time series. They are the main technological advancement underpinning DeepMind's protein structure prediction model, AlphaFold. Transformers serve as the foundation for Codex, an OpenAI source code creation paradigm. Transformers have more recently made their way into the field of computer vision, where they are gradually taking the place of CNN in a variety of challenging applications [31]. Fig. 1 demonstrates the overall process in feed forwards neural networks and recurrent neural networks.

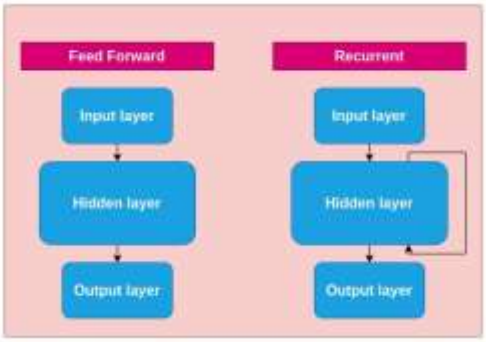


Fig. 1. Feed Forward vs. Recurrent process.

The traditional feed-forward neural network converts each input into an output but is not intended to keep track of consecutive data [32]. This is successful when categorizing photographs but fails when dealing with text or other sequential data. When processing text, a machine learning model must calculate each word as well as take into account the order and relationships of the words. Depending on the words that follow before and after a word in a phrase, a word's meaning might vary [33].

RNNs were the preferred method for natural language processing before Transformers. An RNN processes the first word in a series of words and sends the outcome back to the layer that processes the following word. As a result, it can track the complete phrase rather than just the individual words.

Recurrent neural networks have drawbacks that restrict how effective they might be. They started out moving very slowly. They were unable to benefit from parallel computing gear or graphics processing units (GPU) for training or inference since they had to handle input sequentially. Second, they had trouble reading lengthy text blocks [34]. The impacts of the initial words of a phrase steadily diminished as the RNN read farther into a text fragment. This issue, referred to as "vanishing gradients," was a difficulty when two related words were spread out across a large portion of the text. Thirdly, they simply recorded how a word is related to the ones that came before it. The truth is that words' meanings are influenced by the ones that follow before and after them.

The successor to RNNs, Long Short-Term Memory (LSTM) networks, were able to handle longer text sequences and partially resolve the vanishing gradients problem [35]. However, LSTMs were still unable to fully utilize parallel processing and were significantly slower to train than RNNs. They continued to rely on text sequences being processed serially.

Transformers made two significant contributions that were described in the 2017 paper "Attention Is All You Need." First, they made it feasible to analyze whole sequences in parallel, enabling sequential deep learning models to grow their speed and capacity at previously unheard-of rates. Additionally, they included "attention mechanisms" that allowed for the forward and reverse tracking of word relationships over extremely lengthy text sequences [36]. Fig 2 shows the architecture prototype of variants in sequence-based models.

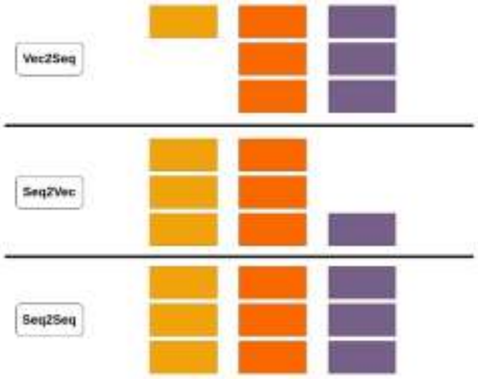


Fig. 2. Vec2Seq, Seq2seq, and Seq2Seq architecture prototype.

It is important to describe the kinds of issues that sequential neural networks are capable of solving before we go into the mechanics of the transformer model. A "vector to sequence" paradigm creates a series of data, like a description, from a single input, like an image [37]. A "sequence to vector" model produces a single value, such as a sentiment score, from a sequence of inputs, such as a social media post or a series of product reviews. A "sequence to sequence" model converts an input sequence, such as an English sentence, into an output sequence, such as the text's French translation.

All of these kinds of models have a trait despite their diversity. They learn how to depict things. A neural network's function is to convert one type of data into another. The neural network's hidden layers—the layers between the transmitter and receiver send their variables during training in order to best reflect the characteristics of the input data type and translate it to the output [38].

A convolutional neural network-based approach is developed to cope with the complicated behavior of arcing current and to identify the arcing state from the regular state of the load current. This network also includes the categorization, creating an end-to-end structure [39]. It is important to note that CNN may be used on the unprocessed current data without the need for any preprocessing operations such discrete wavelet, Fourier, or Chirp Zeta transformations. Given these benefits of CNN, we developed the 1D CNN-based ArcNet model to identify arc and non-arc states as well as different load kinds. A CNN is a trainable hierarchical network with both forward and backward transmission that is made up of several stages.

2.2. Mediums and Calculated Variables

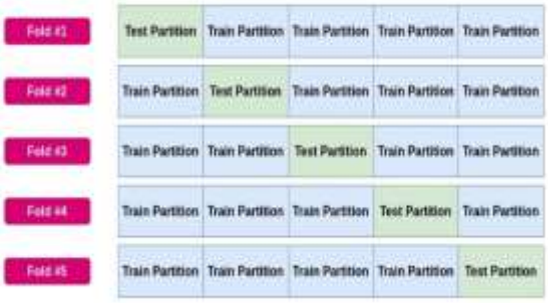
The machine generator supplied the transparent voltage. Ionization then began (voltage signal was constant and current was zero amperes). The local dielectric circumstances, rather than the machine generator, were in charge of this time, also referred to as the ionization time. Ionization time, in all other words, was not a machine parameter but rather dependent on the unique electric field characteristics of the membrane for each discharging [40]. Ionization time was lengthy if flushing was successful and the gap was clean. On the other hand, ionization time was low or even negligible if flushing was challenging and debris was present in the gap.

Each discharge included useful data regarding the operation of the process. Other writers have lately tried to link process signals with end quality products, as was demonstrated in the preceding section. Sophisticated pattern recognition, however, may be far more effective in assessing performance levels. Due to the massive quantity of data that can be gathered (with sampling rates as high as 10.0 MS/s, as discussed below), it is now possible to train DNNs that have previously demonstrated their superiority in other domains.

2.3. Cross Validation

The approach of model evaluation known as cross validation is superior than residuals. The drawback of residual assessments is that they cannot anticipate how well a learner will perform when asked to make new predictions for material that it has not previously seen. One solution to this issue is to train a learner without using the complete data set. Prior to training, some of the data is eliminated. After training is complete, the deleted data may be used to assess how well the learnt model performs on "fresh" data. This is the fundamental principle behind the entire family of model assessment techniques known as cross validation.

One technique to improve upon the holdout method is to use K-fold cross validation. The holdout approach is used for each of the k subgroups of the data set [41]. The other k-1 subsets are combined to create a training set, and one of the k subsets is utilized as the test set each time. The average error for all k trials is then calculated. The benefit of this approach is that the manner in which the data are separated is less important. Each data point appears precisely once in the test set and k times in the training set. As K is raised, the variance of the resultant estimate decreases. The drawback of this approach is that the evaluation process requires k times as much computing since the training procedure must be repeated k times from the beginning. This approach may be modified by randomly dividing the data k times into a test and training set [42]. This has the benefit that you may individually select the size of each test set and the number of trials you average across. Fig. 3 illustrates the overall procedure of cross validation partitioning of data.



Fold #1	Test Partition	Train Partition	Train Partition	Train Partition	Train Partition
Fold #2	Train Partition	Test Partition	Train Partition	Train Partition	Train Partition
Fold #3	Train Partition	Train Partition	Test Partition	Train Partition	Train Partition
Fold #4	Train Partition	Train Partition	Train Partition	Test Partition	Train Partition
Fold #5	Train Partition	Train Partition	Train Partition	Train Partition	Test Partition

Fig. 3. Cross-validation partitioning.

3. RESULTS AND DISCUSSION

3.1. Experimental Setup

Performance evaluation tries to investigate and analyze how well classifiers work in accurately recognizing the instance by using assessment metrics like accuracy, true positive rate, and false positive rate. Our classifier was tested using 5-fold cross-validation. The K-fold cross-validation technique, which separates the original sample into a training set and a test set, is used to assess predictive models. For 5-fold cross-validation, the data is split into 5 subgroups, with the last subset acting as the test set. The classifier is trained using the remaining nine subsets.

3.2. Evaluation Metrics

To evaluate the system's actual values to its expected values, performance evaluation functions as a flexible technique. The objective of our study is to assess and evaluate a classifier's performance in identifying hazardous code. In order to achieve high results from the proposed methods, accuracy—which is defined as equation 1—is something we are really worried about.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{The\ number\ of\ samples} \times 100 \quad (1)$$

A false positive situation occurs when the attack detection technique incorrectly interprets a lawful code as hazardous code. In a specific method, a false negative occurs when harmful code is not discovered while engaging in prohibited behavior. A confusion matrix or error matrix is used to evaluate false positives and false negatives in order to calculate the detection rate. The false positive and false negative detection rates are determined by equations 2 and 3, respectively.

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative} \quad (2)$$

$$False\ Positive\ Rate = \frac{False\ Negative}{False\ Negative + True\ Positive} \quad (3)$$

Where, FN stands for false negative, TN for true negative, and TP for true positive, and FPR stands for false positive rate.

A false positive situation occurs when the attack detection technique incorrectly interprets a lawful code as hazardous code. In a specific method, a false negative occurs when harmful code is not discovered while engaging in prohibited behavior. A confusion matrix or error matrix is used to evaluate false positives and false negatives in order to calculate the detection rate. The false positive and false negative detection rates are determined by equations 2 and 3, respectively.

3.3. Classification Performance

Table 1's findings make it abundantly evident that for all of the evaluated measures, the design with a convolutional layer and GRU network performed better than other models. As a result, this model was utilized to assess the model's performance and the complexity of the other datasets (z 135 and z 15).The evaluation process is done using 5 fold cross validation with the aim of validating the proposed approach in a more reliable fashion.

Table 1. The results achieved for z 135.

Fold	Accuracy	Precision	Recall	F1-Score
1	94.20	96.00	92.66	94.30
2	94.15	96.21	92.73	94.34
3	93.90	95.80	91.67	93.92
4	95.10	96.11	92.80	94.41
5	94.32	96.10	92.43	94.58

Table 2 shows that the outcomes obtained with fewer zones were significantly better than those obtained with all the zones. The F1 Score for the Z 135 dataset was 0.9169, while it was 1 for the Z 15 dataset. These results were excellent since they show how precisely the CGRU (Convolutional Gated Recurrent Unit) network can categorize voltage sequences. The models using GRU units performed significantly better than those with CNN. This made sense because recent studies have shown that GRU units correctly handle sequences. In fact, spark sequences with an F1 score of less than 60% cannot be classified adequately by CNN in the absence of a gate unit. Therefore, using DNNs with GRU units to categorize spark sequences is strongly advised.

Table 2. The results achieved for z 15.

Fold	Accuracy	Precision	Recall	F1-Score
1	93.21	92.01	91.98	92.21
2	94.16	92.09	91.73	92.35
3	92.90	92.80	92.67	92.92
4	92.11	93.21	92.43	92.41
5	93.31	93.15	93. 15	93.21

When focusing on designs with GRU units, it's noteworthy to note that the model with bidirectional GRU fared better than the GRU in terms of recall and F1 Score but practically exactly equal in terms of accuracy. This is intriguing since the input layer of the BiGRU model has less GRU units (10 in the BiGRU and 50 in the GRU). In order to conclude that a BiGRU model would consistently beat the GRU model for sequence classification, the data were not sufficiently apparent.

Similarly, analyzing the results from Table 2, it appears that adding a convolutional layer in the input of a GRU model helped to classify WEDM spark sequences. Thus, the first convolutional layer helped to extract features from spark sequences and then GRU units modeled these new sequences generated by the convolutional layer. Therefore, the results show that a CGRU model works accurately when classifying WEDM sequences with high precision (0.7260). Moreover, Table 3 shows that this model is classified almost perfectly when dealing with less complicated datasets. Indeed, the model was capable of achieving 100% precision for classifying sequences of Zones 1 and 5.

Similar to Table 3, it seems that adding a convolutional layer to a GRU model's input assisted in classifying spark sequences after careful analysis of the findings. In order to model the latest wave created by the convolutional layer, GRU units first analyzed the spark sequences that the first convolutional layer had managed to extract attributes from. As a consequence, the outcomes demonstrate that a CGRU model is effective in classifying sequences with high precision (0.7260). Additionally, Table 3 demonstrates that when working with less difficult datasets, our model categorized nearly flawlessly. In fact, the algorithm was able to classify Zones 1 and 5 sequences with 100 percent accuracy.

Table 3. The achieved TP, TN, FN, FP for each fold

Fold	TP	TN	FN	FP
1	4123	4432	100	162
2	4103	4442	104	122
3	4190	4231	114	123
4	4100	4201	112	133
5	4150	4101	116	109

According to the resutls, the proposed model possess the superior performance with respect to various aspects. First of all, the high accuracy proves the better performance of the model in classifying the correct samples. On the other hand, high recall claims the excellency of the proposed model in terms of fetching the relevant data for both benign and malicious samples. This shows the reliablity of the model in imbalanced situations. High precision shows the good ability of the classifier for detecting benign samples.

4. CONCLUSION

The purpose of this study was to assess the likelihood that using DNNs to identify hidden patterns from process raw voltage signals, one could predict an unexpected event occurring during an industrial wire electrical discharge machining process. For several DNN models and datasets, a comparison of precision, recall, and F1 scores was given. The findings unmistakably demonstrated that the model with the best performance was one with a first convolutional layer and two GRU layers. Additionally, this model delivered exceptional results for the remaining datasets, with a precision of almost 100%. Confusion matrices showed that thickness fluctuation may be predicted, at least 2 mm in advance, giving enough time to adjust machining settings from a process standpoint.

REFERENCES

[1] Wang, Jun, José A. Sánchez, Borja Izquierdo, and Izaro Ayesta. **"Experimental and numerical study of crater volume in wire electrical discharge machining."** *Materials*, Vol. 13, No. 3, p. 577, 2020.

[2] Abu Qudeiri, Jaber E., Ahmad Saleh, Aiman Ziout, Abdel-Hamid I. Mourad, Mustufa Haider Abidi, and Ahmed Elkaseer. **"Advanced electric discharge machining of stainless steels: Assessment of the state of the art, gaps and future prospect."** *Materials*, Vol. 12, No. 6, p. 907, 2019.

[3] Barrios, Sonia, David Buldain, María Paz Comech, Ian Gilbert, and Iñaki Orue. **"Partial discharge classification using deep learning methods—Survey of recent progress."** *Energies*, Vol. 12, No. 13, p. 2485, 2019.

[4] El-Bahloul, Sara Ahmed. **"Optimization of wire electrical discharge machining using statistical methods coupled with artificial intelligence techniques and soft computing."** *SN Applied Sciences*, Vol. 2, No. 1, pp. 1-8, 2020.

[5] Khan, Sarmad Ali, Mudassar Rehman, Muhammad Umar Farooq, Muhammad Asad Ali, Rakhshanda Naveed, Catalin I. Pruncu, and Waheed Ahmad. **"A detailed machinability assessment of DC53 steel for die and mold industry through wire electric discharge machining."** *Metals*, Vol. 11, No. 5, p. 816, 2021.

[6] Umar Farooq, Muhammad, Mohammad Pervez Mughal, Naveed Ahmed, Nadeem Ahmad Mufti, Abdulrahman M. Al-Ahmari, and Yong He. **"On the investigation of surface integrity of Ti6Al4V ELI using Si-mixed electric discharge machining."** *Materials*, Vol. 13, No. 7, p. 1549, 2020.

[7] Fujiyoshi, Hironobu, Tsubasa Hirakawa, and Takayoshi Yamashita. **"Deep learning-based image recognition for autonomous driving."** *IATSS research*, Vol. 43, No. 4, 244-252, 2019.

[8] Pastor-Pellicer, Joan, María José Castro-Bleda, Salvador Espana-Boquera, and Francisco Zamora-Martinez. **"Handwriting recognition by using deep learning to extract meaningful features."** *Ai Communications*, Vol. 32, No. 2, pp. 101-112, 2019.

[9] Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. **"Pre-trained models for natural language processing: A survey."** *Science China Technological Sciences*, Vol. 63, No. 10, pp. 1872-1897, 2020.

[10] Kinghorn, Philip, Li Zhang, and Ling Shao. **"A hierarchical and regional deep learning architecture for image description generation."** *Pattern Recognition Letters*, Vol. 119, pp. 77-85, 2019.

[11] Sharghi, Elnaz, Vahid Nourani, Hessam Najafi, and Amir Molajou. **"Emotional ANN (EANN) and wavelet-ANN (WANN) approaches for Markovian and seasonal based modeling of rainfall-runoff process."** *Water resources management*, Vol. 32, No. 10, pp. 3441-3456, 2018.

[12] Nourani, Vahid, Zahra Razzaghzadeh, Aida Hosseini Baghanam, and Amir Molajou. **"ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method."**

Theoretical and Applied Climatology, Vol. 137, No. 3, pp. 1729-1746, 2019.

[13] Nourani V, Molajou A, Tajbakhsh AD, Najafi H. **“A wavelet based data mining technique for suspended sediment load modeling. Water Resources Management”** , Vol. 31, pp. 1769-84, 2019.

[14] Nassif, Ali Bou, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. **"Speech recognition using deep neural networks: A systematic review."** *IEEE access*, Vol. 7, pp. 19143-19165, 2019.

[15] Stahlberg, Felix. **"Neural machine translation: A review."** *Journal of Artificial Intelligence Research*, Vol. 69, pp. 343-418, 2020.

[16] Ackerson, Joseph M., Rushit Dave, and Naeem Seliya. **"Applications of recurrent neural network for biometric authentication & anomaly detection."** *Information*, Vol. 12, No. 7, p. 272, 2021.

[17] McClelland, James L., and David E. Rumelhart. **"Distributed memory and the representation of general and specific information."** In *Connectionist psychology: A text with readings*, Vol. 43, pp. 75-106, 2020.

[18] Wang, Yaoli, Lipo Wang, Fangjun Yang, Wenxia Di, and Qing Chang. **"Advantages of direct input-to-output connections in neural networks: The Elman network for stock index forecasting."** *Information Sciences*, Vol. 547, pp. 1066-1079, 2021.

[19] Smagulova, Kamilya, and Alex Pappachen James. **"A survey on LSTM memristive neural network architectures and applications."** *The European Physical Journal Special Topics* Vol. 228, No. 10, pp. 2313-2324, 2019.

[20] Kiranyaz, Serkan, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. **"1D convolutional neural networks and applications: A survey."** *Mechanical systems and signal processing*, Vol. 151, p. 107398, 2021.

[21] Sarıgül, Mehmet, Buse Melis Ozyildirim, and Mutlu Avci. **"Differential convolutional neural network."** *Neural Networks*, Vol. 116, pp. 279-287, 2019.

[22] Yin, Jiateng, and Wentian Zhao. **"Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach."** *Engineering Applications of Artificial Intelligence*, Vol. 56 , 250-259, 2016.

[23] Canizo, Mikel, Isaac Triguero, Angel Conde, and Enrique Onieva. **"Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study."** *Neurocomputing*, Vol. 363, pp. 246-260, 2019.

[24] Wang, Peng, Robert X. Gao, and Ruqiang Yan. **"A deep learning-based approach to material removal rate prediction in polishing."** *Cirp Annals*, Vol. 66, No. 1, pp. 429-432, 2017.

[25] Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. **"Transformers in vision: A survey."** *ACM Computing Surveys* , Vol. 363, pp. 246-260, 2021.

[26] Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. **"A survey of multilingual neural machine translation."** *ACM Computing Surveys (CSUR)*, Vol. 53, No. 5, pp. 1-38, 2020.

[27] Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham et al. **"Big bird: Transformers for longer sequences."** *Advances in Neural Information Processing Systems*, Vol. 33, pp. 17283-17297, 2020.

[28] Pan, Xuran, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. **"On the integration of self-attention and convolution."** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vol. 32, pp. 815-825. 2022.

[29] Floridi, Luciano, and Massimo Chiriatti. **"GPT-3: Its nature, scope, limits, and consequences."** *Minds and Machines*, Vol. 30, No. 4, 681-694, 2020.

[30] Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen. **"Transformer models for text-based emotion detection: a review of BERT-based approaches."** *Artificial Intelligence Review*, Vol. 54, No. 8, pp. 5789-5829, 2021.

[31] Bhatt, Dulari, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. **"CNN variants for computer vision: History, architecture, application, challenges and future scope."** *Electronics*, Vol. 10, No. 20, p. 2470, 2021.

[32] Currey, Anna, and Kenneth Heafield. **"Incorporating source syntax into transformer-based neural machine translation."** In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Vol.12, pp. 24-33. 2019.

[33] Grechishnikova, Daria. **"Transformer neural network for protein-specific de novo drug generation as a machine translation problem."** *Scientific reports*, Vol. 11, No. 1, pp. 1-13, 2021.

[34] Basodi, Sunitha, Chunyan Ji, Haiping Zhang, and Yi Pan. **"Gradient amplification: An efficient way to train deep neural networks."** *Big Data Mining and Analytics*, Vol. 3, No. 3, pp. 196-207, 2020.

[35] Jha, Dipendra, Vishu Gupta, Logan Ward, Zijiang Yang, Christopher Wolverton, Ian Foster, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. **"Enabling deeper learning on big data for materials informatics applications."** *Scientific reports*, Vol. 11, No. 1, pp. 1-12, 2021.

[36] Gumaei, Abdu, Mohammad Mehedi Hassan, Abdulhameed Alelaiwi, and Hussain Alsalman. **"A hybrid deep learning model for human activity recognition using multimodal body sensing data."** *IEEE Access*, Vol. 7, pp. 99152-99160, 2019.

[37] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. **"Swin transformer: Hierarchical vision transformer using shifted windows."** In *Proceedings of the IEEE/CVF on Computer Vision*, Vol.65, pp. 10012-10022. 2021.

[38] Liu, Zhenhua, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. **"Post-training quantization for vision transformer."** *Advances in Neural Information Processing Systems*, Vol. 34, pp. 28092-28103, 2021.

[39] Deepak, S., and P. M. Ameer. **"Automated categorization of brain tumor from mri using cnn**

features and svm." *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 8, 8357-8369, 2021.

[40] Choi, Youngkwon, Gayathri Naidu, Long D. Nghiem, Sangho Lee, and Saravanamuthu Vigneswaran. **"Membrane distillation crystallization for brine mining and zero liquid discharge: opportunities, challenges, and recent progress."** *Environmental Science: Water Research & Technology*, Vol. 5, No. 7, 1202-1221, 2019.