Jurnal Teknologi

Full Paper

A DEEP LEARNING APPROACH TO COLLISION DETECTION IN NPP COOLANT SYSTEM USING VIBRATION DATA

Chua Kai Zer, Nahrul Khair Bin Alang Md Rashid

School of Chemical Engineering, Faculty of Engineering, UTM, 81310 Skudai, Johor, Malaysia

Article history
Received

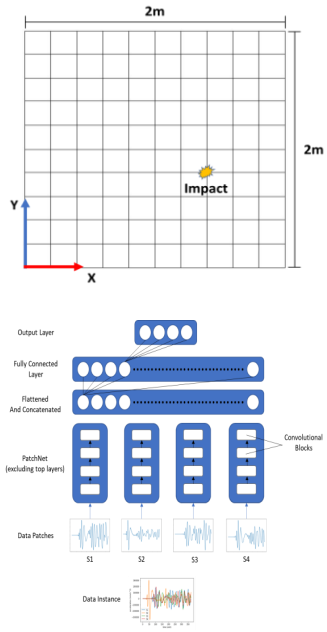
Received in revised form

Accepted

Published Online

*Corresponding author
nahrulkhair@utm.my

Graphical abstract



Abstract

In recent years, the adoption of intelligent systems in fault detection and diagnosis (FFD) has expanded into the fields of nuclear engineering especially in structural damage identification, defect tracking of instruments, and other signal processing operations. In this study, the proposed Patch-Based 1D Convolutional Neural Network is subject to two training phases, namely local pre-training and global fine-tuning that allow the model to learn both local and generic features for all-at-once collision parameters prediction in reactor coolant system by using the waveform simulation acceleration data generated by four accelerometers. Besides, a multi-model approach including the data aggregation aided machine learning (ML) algorithms and simplified 1D Convolutional Neural Network for many-to-one prediction of each collision parameter are also implemented as additional model options to achieve the main objective of the study which is to select the best performing model under the percentage mean absolute error metric.

Keywords: Artificial Intelligence, Fault Detection and Diagnosis, Artificial Neural Network, Convolutional Neural Network, Local-specialized.

Abstrak

Kebelakangani ini, penerapan sistem kecerdasan dalam bidang kejuruteraan nuklear melalui kaedah-kaedah pengesanan dan diagnosis kesalahan semakin berkembang dalam pengesanan dan penjejakan kerosakan struktur, kecacatan komponen dan juga operasi pemprosesan isyarat yang lain. Dalam kajian ini, model Rangkaian Saraf Konvolusi 1D Berpetak telah menjalani dua fasa latihan, iaitu pra latihan secara tempatan dan juga penalaan secara global yang membolehkan pemebelajaran kedua-dua ciri tempatan dan generik dalam peramalan parameter perlanggaran secara keseluruhan dalam sistem penyejukan reactor dengan menggunakan data gelombang yang dihasilkan oleh empat akselerometer. Selain itu, satu pendekatan multi-model yang merangkumi algoritma pembelajaran mesin dibantu agregasi data dan Rangkaian Saraf Konvolusi 1D dipermudahkan demi peramalan setiap parameter perlanggaran secara banyak ke satu juga diimplementasikan sebagai model pilihan untuk mencapai objektif utama kajian dalam memilih model yang berprestasi terbaik bawah metrik peratusan min ralat mutlak.

Kata kunci: Kecerdasan Buatan, Pengesanan dan Diagnosis Kesalahan, Rangkaian Saraf Tiruan, Rangkaian Saraf Knovolusi, Pengkhususan Tempatan

1.0 INTRODUCTION

With the rapid advancements in computational performance, computer data storage capacity, big data analytics, and information technologies such as the internet of things (IoT), artificial intelligence (AI), and cloud computation enables the realization of the vision and the potential of Industry 4.0 to achieve a higher level of automation in various industries including the nuclear industry. In the past decades, there exists a variety of risks and major concerns for current nuclear power plants (NPP). Firstly, the complex system itself creates a relatively high barrier to its operators. According to [1] in their study, various failures can occur may occur in equipment, instruments, or even system processes in NPP, greatly impacting both the performance and security of the NPP. Due to the insufficient automation level of NPP, system control and management is generally carried out manually by human operators under great pressures because of high control requirements, thus making NPP systems vulnerable to human errors [1] which may lead to fatal transients and accidents. However, the problems above can be addressed by the application of AI technologies as auxiliary systems in the management of NPP systems. This study aims to develop an auxiliary neural network-based framework to predict important parameters of a collision event in RCS from accelerometer vibration data provided by Korea Atomic Energy Research Institute (KAERI) [2]. The main objective of the project is to reduce the reliance of NPP safety systems, particularly in RCS on human intervention so to mitigate the operational errors due to varying degrees of external influence on operator performance in decision making. The proposed modelling approach is expected to extract, convert and assimilate the raw vibration data from accelerometers to bring meaningful knowledge to the operators in NPP.

2.0 LITERATURE REVIEW

2.1 Biological Neurons and the Perceptrons

As part of the evolutionary history of deep learning models, the early study of biological neural networks in human brain had propelled the artificial intelligence research frontier forward. Biologically, neural networks are made up of special biological cells called neurons that are responsible for processing information. Interconnected by neuronal junctions called synapse, a biological neuron has a cell body responsible for processing incoming signals from the dendrite. The processed signals are transmitted to relevant cells by axon through the secretion of neurotransmitters. During the process, the massively connected neurons can effectively adjust themselves and learn from human activities in which they participate. A perceptron (neuron or node) is a bio-inspired single mathematical information-processing unit. It takes in one or more input signals from synapses of nodes situated at the previous layer which are characterized by specific weights and bias. The bias is applied externally as an auxiliary node which can promote better data fitting. At the summing junction as shown in Figure 1, the sum of the product of input values and their associated synaptic weights is computed. The calculated weighted sum is then fed into an appropriate activation function to make modifications to the learned pattern [3]. This is done by squeezing the value to a finite permissible amplitude range. The typical target intervals are [0, 1] and [-1, 1]. The activation function returns a final value that is then passed as an input value to the succeeding layer or as the final output in the case where the perceptron is the output node of the network. In mathematical terms, the relationship between the weighted sum, and the output, can be shown in equations (1) and (2) as:

$$z = w_1x_1 + w_2x_2 + \ldots + w_{m-1}x_{m-1} + w_mx_m + b \quad (1)$$

$$y = \varphi(z) \quad (2)$$

where,

x_n = the n-th input value of the current instance
w_n = the connection weight between the n-th input neuron and the current output neuron
b = the current layer bias
φ = the activation function

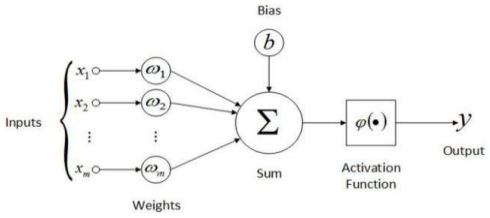


Figure 1 Simplified Model of a Perceptron

2.2 Artificial Neural Network

Artificial Neural Network (ANN) is an interconnection of multiple bio-inspired perceptrons that are ordered into layers in which its process of activation computation is usually done in a single forward direction. ANN comprises processing elements that receive information and produce output based on their learned parameters such as weights, biases, and their pre-defined network architectures and activation functions. ANN is very famous together with backpropagation as its training algorithm [4] and typically consists of hundreds or thousands of interconnecting simple processing units

called the artificial neurons (nodes) which are structured and programmed in such a way to mimic the working of the human biological nervous system. Generally, the ANN is organized and arranged in three types of layers, namely the input, hidden and putput layer. The nodes in the input layer will supply respective elements from the input vector to nodes in the second layer (hidden layer). The output values of the second layer will be the inputs of the third layer and so on for the rest of the network. In typical neural network design as shown in Figure 2, more than one hidden layer is used and the network is fully connected in which every node within a layer and its preceding layer is interconnected. Thus, the activation of each predecessor reflects certain properties from its ancestors in the previous layers which is controlled by the degree of significance for given connections, the weight coefficients [4].

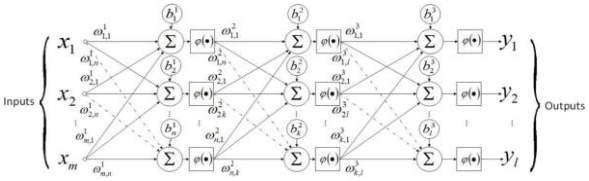


Figure 2 A Multilayer Artificial Neural Network

In ANN, activation functions represent a set of mathematical functions used to process and appropriate the weighted sum of the input values to a finite range as a way to manipulate how activated the neuron should be and to decide if the neuron can be fired or not. The role of an activation function is highly dependent on its position in the neural network. It performs conversion of linear mappings into suitable non-linear forms for propagation when placed after any hidden layer while in the output layer, it makes prediction [3]. Although deep neural networks develop better performances with the aid of activation functions, common problems like vanishing gradients and exploding gradients may arise, thus necessitating thorough activation function selection.

2.3 Deep Learning

Learning is a fundamental trait or ability of intelligence. Unlike humans, learning in artificial neural networks can be viewed as a process of updating the values of connection weights or network architecture in some models so to perform a specific task. During the learning process, the performance of a network is improved over time and its direction of weight updates are supervised and governed by the loss function set by human which maps the deviation of its output from the desired results. The weights and biases of the network are tuned in such a way that the loss function is reduced as much as possible. Given a collection of representative data, artificial neural networks appear to be able to learn underlying patterns connecting the inputs and outputs through deep layers in which some of them do not make sense and are impossible for a human to fully understand.

2.4.1 Backpropagation Training Algorithm

The backpropagation algorithm is well-known for its ability to monitor neural network learning and store the intrinsically learned mapping relations without the need to specifically disclose the related mathematical equation in advance. Depending on the types of neural networks involved, the training algorithm typically utilizes gradient descent to regulate and modify the weight and bias values of a network to achieve a minimum sum of squares error [7]. In this algorithm, the weights and biases are initialized by a pre-defined kernel initializer before the process starts. The algorithm handles one mini-batch of 32 instances by default at a time until the full training set is gone through, indicating the passage of an epoch. According to [7], the backpropagation learning process can be described by two main operations: the forward propagation of the operating signal and the backward propagation of the error signal. During the forward propagation, the outputs of all input neurons are computed and passed to the next layer as described before until the output of the last layer is obtained. The algorithm then compares the desired output, y, and the predicted output, $\hat{y}$ of the jth output neuron of the network using the typical sum of the squares loss function, C in equation (3) as

$$C = \frac{1}{2}\sum_{j}(\hat{y}_j - y_j)^2 \tag{3}$$

The vector of loss function values for each instance in the mini-batch is calculated and averaged. According to the minimization method of steepest descent as illustrated in Figure 3, the fastest decline in loss function can be achieved by some modifications to the weight and bias values along with their respective negative gradient direction [7]. In association with that, the chain rule is applied to compute the loss function changes with respect to the weight, w and bias, b of the kth training step which will be used in equations (4) and (5) along with the learning rate, η for weights and biases updates.

$$w_{k+1} = w_k - \eta\frac{\partial C}{\partial w_k} \tag{4}$$

$$b_{k+1} = b_k - \eta\frac{\partial C}{\partial b_k} \tag{5}$$

During the process, the error contributions from each connection are measured. The weights and biases remain constant during the loss function gradients computation with regards to each parameter of each instance within the mini-batch. Finally, during the error signals backpropagation, a gradient descent step is performed in order to tweak all the connection weights and biases in the network by using the gradients of loss functions computed. The next mini-batch repeats the same process but with newly updated weights and biases.

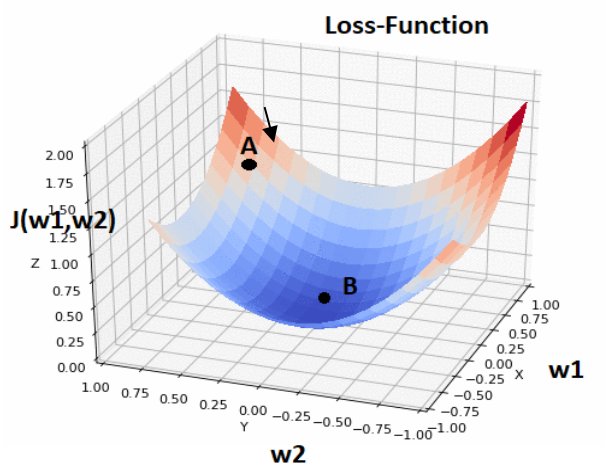


Figure 3 The Steepest Descent $(-\nabla J(w1, w2))$ from Point A to B

2.4.2 Avoiding Vanishing & Exploding Gradients, Dead Neurons, and Overfitting

Although the backpropagation algorithm shows promising efficiency in updating connection parameters, the loss function gradients usually get smaller and smaller as the gradients propagate towards the lower layers. As a result, the connection weights especially of the lower layers will take forever to train and converge to a desirably decent solution. The scenario is known as the vanishing gradients problem.

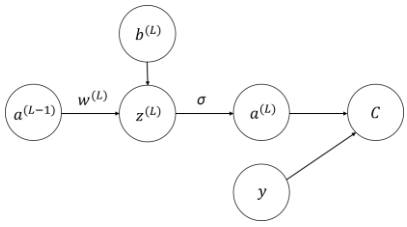


Figure 4 Neural Illustration of Output Node and its Preceding Node

Figure 4 illustrates the relationship between an output node of layer *L* and one of its preceding neurons, the output neuron activation, $a^{(L)}$ is determined by equations (6) and (7), where the transformation of the result, $z^{(L)}$ of weight, $w^{(L)}$, bias, $b^{(L)}$ of layer *L* and the previous neuron's activation, $a^{(L-1)}$ by the activation function, σ along with the desired output, *y* is used to compute the sum of squares loss, *C*.

$$z^{(L)} = w^{(L)}a^{(L-1)} + b^{(L)} \tag{6}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{7}$$

The error gradient or the change of loss both with respect to weight and bias used for connection parameters update is computed by using the chain rule as shown in equations (8) and (9).

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial C}{\partial a^{(L)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial z^{(L)}}{\partial b^{(L)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial C}{\partial a^{(L)}} \tag{9}$$

The derivative $(\partial a^{(L)}/\partial z^{(L)})$ of sigmoid and tanh function lie in the intervals of (0,1) and (-1,1) respectively. Since the maximum modulus value of the derivatives is always less than 1, repetitive multiplication of smaller numbers will result in a very small error gradient especially at the region of lower layers as the chain rule progresses. In some cases, as opposed to the case of vanishing gradient problem, the gradient may grow bigger as the algorithm propagates backward. This is the exploding gradients problem. By further expanding the chain rule applied,

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial a^{(L-1)}}{\partial w^{(L)}}\frac{\partial z^{(L)}}{\partial a^{(L-1)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial C}{\partial a^{(L)}} \tag{10}$$

$$\frac{\partial z^{(L)}}{\partial a^{(L-1)}} = \frac{\partial}{\partial a^{(L-1)}}(w^{(L)}a^{(L-1)} + b^{(L)}) = w^{(L)} \tag{11}$$

From equation (11), the term $\partial z^{(L)}/\partial a^{(L-1)}$ depends solely on the value of the $w^{(L)}$ and if the weight is inappropriately initialized as a big number, the result of the chain rule in equation (10) will be a very huge number, thus causing the error gradients to be divergently updated from the desired set of values. In the case of "dying ReLU", the dead neuron problem is caused by the zero gradients on the left side of the ReLU activation function derivative. The zero value of $(\partial a^{(L)}/\partial z^{(L)})$ will result in complete deactivation of certain connection areas where $z^{(L)} < 0$, which in turn brings the network into a state of coma. These problems show the significance of appropriate weight initialization and activation function selection before training an artificial neural network to ensure the convergence of a neural network. Glorot and Bengio [9] experimented and proposed the very first initialization analysis and it was found that by assuming

equal layer inputs (fan_{in}) and outputs (fan_{out}) neurons in the case where the activation function is differentiable at 0, the variance of input and output gradients are equal both in forward and backward flow of the algorithm. The widely used initialization method has proven to work very well in practice and is known as the Glorot Initialization. In the initialization method, the weights in a neural network are sampled and initialized from a normal distribution N(0, σ^2) where the variance,

$$\sigma^2 = \frac{1}{fan_{avg}} \quad (12)$$

With fan_{avg} in equation (12) replaced by fan_{in} comes the initialization strategy named LeCun initialization. LeCun initialization is identical to that of Glorot when $fan_{in} = fan_{out}$. In the case where the neural network is built entirely with dense layers and SELU activation function, LeCun initialization works extremely well with the self-normalizing nature of the neural network. However, despite Glorot's promising weight initialization method, it was found that the technique did not perform well with activation functions that are not differentiable at 0 like ReLU and its variants. In association with that, He et al. [10] proposed an initialization of weights sampled from variance,

$$\sigma^2 = \frac{2}{fan_{in}} \quad (13)$$

and justified their method with an experiment on a 30 layers deep ReLU function neural network in which He initialization method surpassed Glorot's method in terms of convergence. The activation functions and their corresponding weight initialization methods are summarized in Table 1.

Table 1 Initialization Methods for Different Activation Functions

Initialization Method	Activation Functions	σ^2 (Normal)
Glorot	Tanh, Logistic, Softmax	$\frac{1}{fan_{avg}}$
He	ReLU and variants	$\frac{2}{fan_{in}}$
LeCun	SELU	$\frac{1}{fan_{in}}$

Beside weight initialization method, batch normalization (BN) is an important technique to address the vanishing and exploding gradients problems by reducing the internal covariate shift of the neural network. Internal covariate shift is defined as the activation distribution shift due to the constant changing of network parameters during training [11]. In the BN technique, an operation of zero-centering and normalizing activation is added and performed before or after a hidden layer. By using two new parameter vectors, BN scales and shifts the input result to let the model learn optimally and not chasing the moving dot. These actions create the whitening effect of inputs produced by the lower layer, thus making a big step towards achieving non-moving input distribution to mitigate the negative impact of internal covariate shift [11].

With the inclusion of dropout layers in model top layers, one of the main deep neural network concern, the overfitting problem can be lightened. There is a structure-wise resemblance of dropout layer to a dense layer, the only thing different is that at every training step, every neuron of the dropout layer has a probability or dropout rate, *p* of being "shut down", leaving the weights associated to the neuron untrained and ineffective to the preceding and proceeding layer for the particular step. Applying the dropout layer to a neural network can be thought of as sampling vast amounts of sub-networks that contain survived dropout neurons. Consider just adding a dropout layer with *n* units of neuron, there are $2^n - 1$ possible sub-networks to be sampled and presented for weights training [8].

2.5 Convolutional Neural Network (CNN) in Fault Detection and Diagnosis (FDD)

CNN has recently been applied in the development of state-of-the-art methods and proven to be efficient in vibration-based problems by recent studies. In a recent study. In 2018, Abdeljaber et al. [13] designed and trained a 1-dimensional CNN (1DCNN) based model by using Silicon Drift Detector (SSD) vibration data to detect and predict the structural damage of a civil structure. The vibration response of 12 sensors used in a total of 31 scenarios was recorded and used to train a 1DCNN model for each sensor which is only responsible for predicting the local average damage value PoD_{avg} at the corresponding location. It was found that the result of the approach was a success and had unprecedentedly achieved successful predictions on completely unseen damage scenarios [13]. In another study, Yu et al. [14] conducted an experiment on damage identification of smart buildings. Instead of using 1DCNN, the waveform acceleration data from 14 accelerometers were concatenated into a 2-dimensional matrix for 2DCNN model training. Among

three models used in the regression problem, the proposed 2DCNN based model showed the least root mean square error (RMSE) percentage of predicted damage severity in both validating (0.98%) and testing (1.6%) phase as compared to the other models namely General Regression Neural Network (GRNN) and Adaptive Neuro-Fuzzy Inference System (ANFIS). In studies conducted by [13] and [14], both 1-dimensional and 2-dimensional CNN have shown superior results and performance over other architectures in the field of vibration-based FDD. Besides, the studies conducted by Chae et al. [15] in 2020 showed a detection accuracy of 99.1% with their proposed CNN model in NPP coolant system pipe damage detection due to fast accelerated corrosion (FAC) [15]. The wide usage of CNN as FDD methods in the field of nuclear engineering can also be exemplified by Chen et al. [16] in their study on crack detection of underwater metallic surfaces for NPP components inspection. While 1DCNN has a relatively low computational cost for training and is preferable for isolated applications [12], 2DCNN outperforms 1DCNN in cases involving a massive number of detectors through data concatenation.

2.6 Proposed Work

In this study, a variant of 1DCNN called the Patch-Based 1DCNN is proposed. While being able to retain the advantage of low computational cost as a 1DCNN, the patch networks design in the main structure allows the model to focus on the importance of deep regional or local features and hold them together for prediction. The proposed model is expected to be as precise as 2DCNN with data concatenation in vibration analysis and outperform conventional CNNs in terms of learnable parameters' tuneability and flexibility under an appropriate regularization framework. In association with that, a few alternatives of different modelling apporach which includes machine learning algorithms and also CNN of different architecture are also proposed as additional options for model selection. The techniques used, architecture, training, validating, and the evaluation processes of the Patch-Based 1DCNN and other models will be further discussed in the coming chapters.

3.0 METHODOLOGY

3.1 Experimental Flowchart

Figure 5 shows the experimental flowchart to achieve the objectives of the study. In the following sections, each process in the flow diagram will be discussed in more detail.

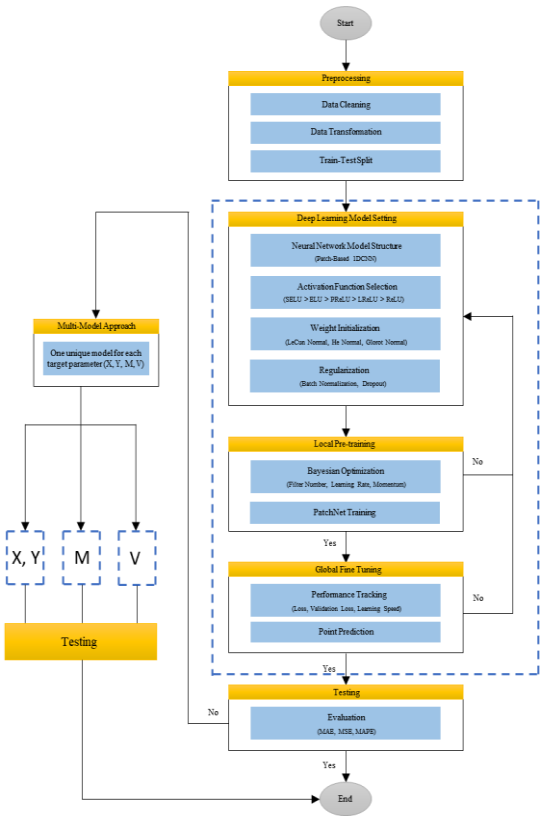


Figure 5 Experimental Flowchart

3.2 Data Preprocessing

In this study, the dataset used consists of simulation time-domain vibration data of 2800 collision events generated by 4 accelerometers. Each collision event is associated with collision parameters such as collision point, (X, Y) in mm, the mass, M (g), and velocity, V (m/s) of the body upon collision to be trained and predicted. The locations of the accelerometers in the simulator are enclosed to avoid any intended selection bias. Since the acceleration data of all 2800 impact events have the same uniform distribution and number of time instances which is 375, the specific time instances are of no importance and can be represented as a simple unit. For each impact event, the 4 one-dimensional arrays of acceleration wave data from 4 accelerometers are concatenated to form a 4 x 375 two-dimensional array representing a single training data as illustrated in Figure 6 (a). Alternatively, instead of using the raw waveform of the data, fast Fourier transform (FFT) is performed to generate another set of experimental data by computing discrete Fourier transform (DFT) of the data sequences as shown in Figure 6 (b). By using the Scikit-Learn library, specifically the train_test_split method, the data arrays are split into random train and test subsets. In this experiment, the test_size is set to 0.2, giving a train-test ratio of 8:2, equivalent to 2240 training and 560 testing instances respectively.

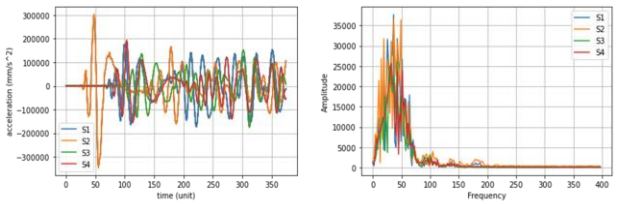


Figure 6 (a) Raw Waveform and (b) its DFT Plot

3.3 Patch-Based 1DCNN Architecture

The fundamental Patch-Based 1DCNN presents four convolutional blocks at the bottom, with each block comprising a one-dimensional convolutional layer immediately followed by a batch normalization, a SELU activation layer, and a max pooling layer as shown in Figure 7 (a). The batch normalization layer serves as a normalizer to the output feature maps from the convolutional layer. Since the network is to be trained sequentially, the SELU activation function is preferred for better training performance. At the top of the network, output feature maps of the last convolutional block are flattened and connected to a dense layer of 375 nodes which is also followed by a batch normalization, a SELU activation, as well as a dropout layer with a dropout rate of 0.5. The final layer of the model consists of four nodes that return four continuous numerical values, of the collision parameters.

Table 2 Model Layer Specifications

Layer	Kernel Size	Stride	Input Maps	Output Maps
1D Convolutional 1	3 × 3	1	1 × (375)	Scaler*2 × (373)
Max Pooling 1	2 × 2	2	Scaler*2 × (373)	Scaler*2 × (187)
1D Convolutional 2	3 × 3	1	Scaler*2 × (187)	Scaler*3 × (185)
Max Pooling 2	2 × 2	2	Scaler*3 × (185)	Scaler*3 × (93)
1D Convolutional 3	3 × 3	1	Scaler*3 × (93)	Scaler*4 × (91)
Max Pooling 3	2 × 2	2	Scaler*4 × (91)	Scaler*4 × (46)
1D Convolutional 4	3 × 3	1	Scaler*4 × (46)	Scaler*5 × (44)
Max Pooling 4	2 × 2	2	Scaler*5 × (44)	Scaler*5 × (22)
Fully Connected	-	-	Scaler*5 × (22)	1 × (375)
Dropout	-	-	-	-
Output	-	-	1 × (375)	1 × (4)

3.4 Local Pre-training

In order to effectively initialize the whole Patch-Based 1DCNN, the model is made specialized in deep features from each signal region by splitting the two-dimensional vibration data and Patch-Based 1DCNN architecture into four patches and PatchNets respectively. The architecture of a PatchNet is shown in Figure 7 (b). Each PatchNet expects a patch or one-dimensional array of 375 elements from a respective vibration signal of the accelerometer. Before proceeding to the pre-training phase, hyperparameters such as filter number of one-dimensional convolutional layer, momenta of batch normalization layer, and the learning rate of PatchNet are fine-tuned. While maintaining a 2:3:4:5 ratio for the filter number from the first to the last convolutional layer, an integer scaler ranges from 5 to 10 is introduced. The scaler introduced is applied by direct multiplication with the ratio, resulting in the number of filters as shown in Table 2. Meanwhile, the batch normalization momenta and model learning rate are assigned with two sets of values respectively, [0.9, 0.99, 0.999] and [0.1, 0.01, 0.001]. By using the training set and a suitable tuning method, the three sets of hyperparameter values are implemented into a defined HyperModel and run for 20 epochs to select the best performing hyperparameter combination to achieve the objective function, in this case, the combination with the lowest validation error. In this experiment, the Bayesian Optimization method is preferred as it is a probability model that keeps track and learns from its past evaluations in contrast to other methods like random and grid search. In correspondence to the SELU activation used, the LeCun Normal method together with the best hyperparameter combination is then used to initialize each PatchNet for pre-training.

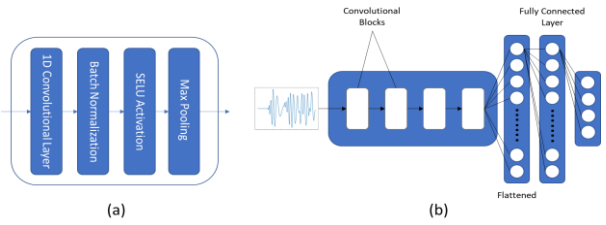


Figure 7 (a) Convolutional Block and (b) PatchNet Architecture

3.5 Global Fine Tuning and Testing

After training the four PatchNets in their respective signal region, their final weights and biases, excluding that of top layers are used to construct the bottom layers of the whole Patch-Based 1DCNN with shared top layers as shown in Figure 8 for global training on full vibration data training. The top layers of the main model are randomly initialized so to further improve the generalizability of the model. After the initialization, the same training data are split into patches and fed into their respective channels to fine-tune the weights of the model as a whole, allowing it to explore more global features which were not locally learned by each PatchNet. Figure 8 shows the ultimate architecture of the Patch-Based 1DCNN during and after global fine-tuning.

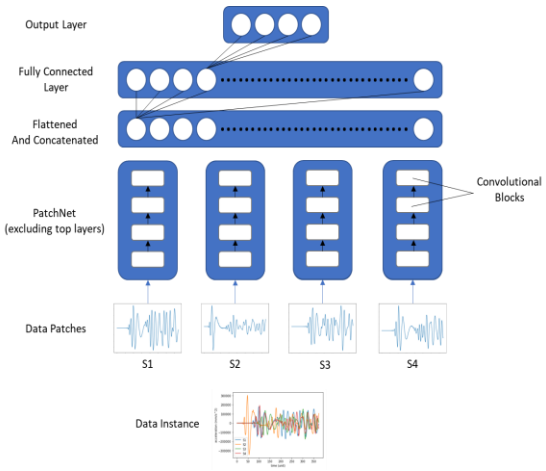


Figure 8 Patch-Based 1DCNN Architecture

Once the model is completely trained, the tuned weights and biases are frozen and remain uninfluenced during the testing phase. In this experiment, the model is evaluated based on the percentage Mean Absolute Error (MAE) of the 560 instances to be tested.

3.6 Multi-Model Approach

Depending on the result of the proposed approach, a multi-model approach may necessarily be implemented where each target parameter is associated with only one predictive model. In this approach, both machine learning algorithms and deep learning models will be considered to look for the best model for each parameter. In contrast to the nature of the proposed model, there will be 3 separate models used to individually predict the impact point (X, Y), the velocity (V), and the mass (M) of the collision body. Depending on the collision parameters' time-frequency dependencies, a suitable type of dataset generated is used in their respective training, validating, and testing processes.

4.0 Results and Discussion

4.1 Patch-Based 1D Convolutional Neural Network

By using Python 3.9, the PatchNet architecture are constructed and defined using TensorFlow-Keras in Jupyter Notebook, an interactive web-based computing platform that is suitable for model training. A total of 4 PatchNet are defined, each representing an input channel for wave data of each sensor. During hyperparameter tuning, Bayesian Optimization algorithms is used and the best working combination of parameters is found to be ['scaler' : 9, 'momentum' : 0.99, 'learning_rate' : 0.001]. The technique of early stopping is applied and its parameters are set to halt the model training once 50 non-progressive epochs based on validation mean absolute error are reached, returning the best performing model weights throughout the training history. By using the best parameters from the tuner, each PatchNet is initialized and trained for 500 epochs. The training history of all PatchNet is shown in Figure 9.

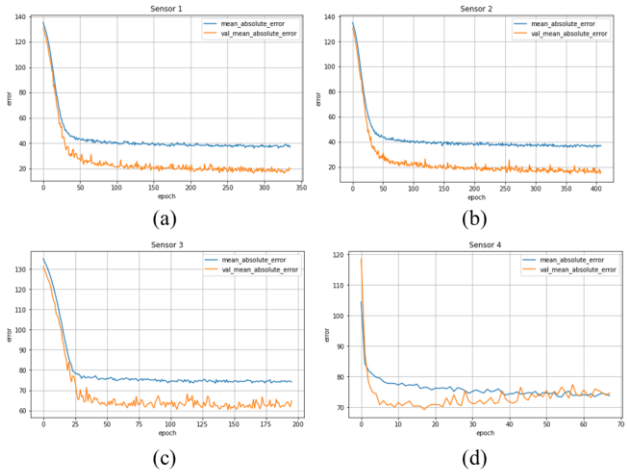


Figure 9 Training History of PatchNet for (a) Sensor 1, (b) Sensor 2, (c) Sensor 3, and (d) Sensor 4

It can be noticed that both training and validation mean absolute error dropped significantly in the first 20 epochs. From Figure 9 (a) and (b), PatchNet 1 and 2 managed to make their long ways to more than 350 epochs, indicating a good training with fewer encounter of non-progressive epochs. On the other hand, PatchNet 3 and 4 shown in Figure 9 (c) and (d) were called back and halted earlier due to their fluctuating validation mean absolute error, inferring the lack of useful information from the wave data of Sensor 3 and 4 for generalization purpose and all-at-once prediction of collision parameters. Also, almost all models have their training mean absolute error higher than that of validation. In most cases, training error would always underestimate validation error. This is a bad sign of unknown fits as the models are not predicting based on what they are trained for.

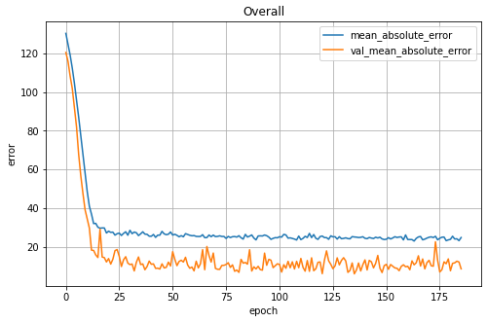


Figure 10 The Overall Model Training History

After the pre-training process, the 4 PatchNets are concatenated and the resulting model architecture is then subject to global tuning. Having the exact same characteristic of unknown fit as illustrated in Figure 10, the final model is halted at epoch 186, recording the best overall validation mean absolute error of 6.2713 (0.6596%). As expected, the final model did not perform well on testing sets, marking a mean absolute error of 110.5015 (11.6219%) which is also a clear sign of overfitting. Before proceeding to the multi-model approach, another attempt was made by switching to the dataset after FFT. However, the dataset was found to be very poor and lack of meaningful information for collision parameters prediction both in single and multi-model cases. Thus, the FFT dataset is eliminated from the experimental procedures afterwards.

4.2 Multi-model Approach

In this approach, machine learning algorithms are used for modelling based on the sample data with a different feature engineering approach. Instead of using the raw waveform as the input data, the amplitude maximum, minimum, mean, median, standard deviation and the distribution skewness of each sensor data array for data instances are computed and made into a data frame as shown in Figure 11. The resulting table has a dimension of 2800 x 24, which consists of 2800 instances of 1D arrays to be fed to the machines.

id	S1_max	S2_max	S3_max	S4_max	S1_min	S2_min	S3_min	S4_min	S1_mean	S2_mean	...
0	235080.9	235080.90	231773.4	286557.2	-315471.5	-315471.50	-168789.1	-293865.7	-308.304553	-308.304553	...
1	1069688.0	1244737.00	951349.6	1095499.0	-826482.9	-2390317.00	-1137225.0	-1306526.0	12780.176893	981.778780	...
2	383092.2	150770.40	148724.0	168649.2	-364591.8	-161006.00	-150458.2	-201313.7	-674.289628	-1575.681938	...
3	379650.3	378475.30	394448.7	487880.4	-329680.7	-726961.00	-333712.8	-463918.9	2164.216981	-11274.728589	...
4	754833.0	321731.20	373526.8	397334.0	-592148.9	-378091.60	-423593.5	-472361.4	-980.480585	-6966.912566	...
...	...	...	...	...	...	...	...	...	...	...	...
2795	698079.1	1027622.00	1031316.0	1037711.0	-793812.4	-720545.40	-1155664.0	-1703724.0	5600.614397	5885.824925	...
2796	290787.9	144759.00	119504.7	157180.4	-493104.6	-179902.70	-100318.6	-180438.7	-1215.328099	2702.879683	...
2797	109273.7	98441.96	109713.6	162800.1	-127440.0	-90686.41	-101889.3	-123130.7	436.239993	276.327514	...
2798	576629.1	550360.40	803728.8	878377.7	-587942.2	-742554.80	-811196.2	-1196096.0	-2394.515047	6230.349355	...
2799	648335.8	965234.70	499031.7	645815.6	-571597.8	-627908.20	-492273.2	-621268.9	1059.619352	28.166706	...

2800 rows x 24 columns

Figure 11 Transformed Sensors Wave Data

From the Scikit-learn machine learning library, regression models like Linear Regression (LR), Elastic Net (EN) (representing Lasso and Ridge regression), K-Nearest Neighbours Regressor (KNN), Decision Tree Regressor (CART), and Gradient Boosting Regressor (GBM) are imported and appended into a pipeline for k-fold cross validation to predict the collision parameters using 10 splits within the training set. During the validation process, the dataset will be split into 10 sets (9 for training and 1 for validating) and validated for 10 trials, with each split acting as the validation set per trials. The resulting mean absolute errors of each model after the 10 trials are averaged with their respective trial standard deviation computed and tabulated as shown in Table 3.

Table 3 Performance of ML Models on Each Collsion Parameters

Model	X, Y [-1,1]		M [0,1]		V [0,1]	
	Mean	Std	Mean	Std	Mean	Std
LR	0.328200	0.013658	0.207814	0.006412	0.109594	0.004585
EN	0.556603	0.008279	0.290931	0.009985	0.301392	0.009794
KNN	0.176049	0.011449	0.181503	0.006838	0.080335	0.005614
CART	0.128181	0.011692	0.153274	0.021667	0.070647	0.009285
GBM	nan	nan	0.166211	0.005929	0.075040	0.003554

In cases of M and V prediction, CART achieved the lowest mean absolute error and highest standard deviation among all the models. Due to its relatively unsteady results throughout the trials, the second performing GBM will also be brought to the next tuning phase because of its steady performance. In X and Y prediction, the nan values indicate the incompatibility of 2D array outputs with the nature of GBM which combines "weak learners" sequentially to improve from its error. Thus, the cross-validations of X and Y are excluded from the experimental procedure due to the concerns of information impairments if parameter X and Y carrying the same locational information are to be treated separately. Thus, only CART will be tuned and tested in the next phase. Grid search was performed and a model was built for every possible combination of hyperparameter for evaluation. After the tuning process, the best performing combination will be used to initialize the model to be tested on the testing sets. For parameter X and Y, CART records a mean absolute error in percentage of 25.558% while for parameter M and V, giving {"CART" : 24.9399%, "GBM" : 15.6316%} and {"CART" : 11.0310%, "GBM" : 6.8796%} respectively. The overall performance of using only CART model is 20.5096%, while for CART+GBM, giving a moderately lower mean absolute error of 15.8564%.

Aside from ML algorithms, a simplified 1D convolutional neural network was also proposed in the multi-model approach. Instead of separating the wave data into patches based on their sensor origin, the data are transposed and arranged in such a way to directly feed the model with 4 channels of 1D arrays. A very simple and direct 1DCNN architecture as illustrated in Figure 12 is used for training. In order to ease the training process, XY and M target values, originally ranged at [-400,400] and [25,175] are scaled to [-1,1] and [0,1] using a min-max scaler, governed by tanh and sigmoid respectively as the final output activation function of their respective model. A sigmoid final output activation function is also directly applied to Model (V) without the need of scaling as its target values at range [0.2,1] are already within the sigmoid "sweet spot". By applying the same number of training epochs and callback, each model is trained individually.

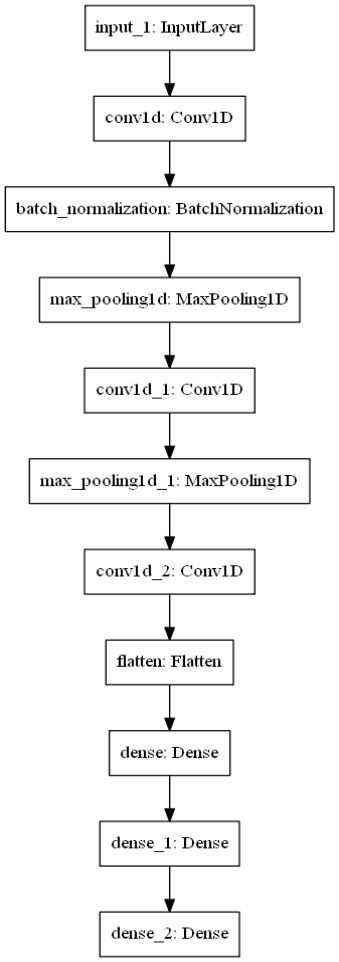


Figure 12 Model Plot of Simplified 1D Convolutional Neural Netwok

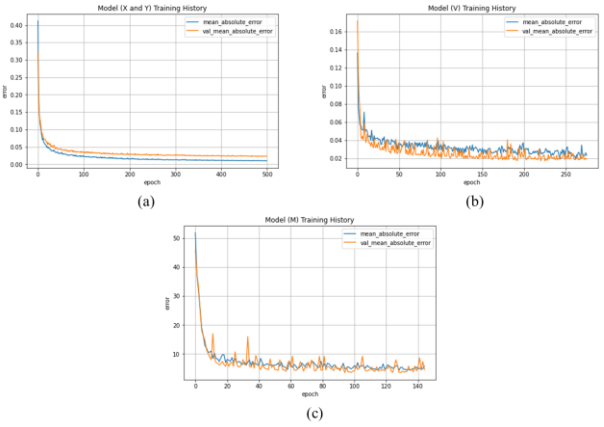


Figure 13 Training History of Model (a) X and Y, (b) V, and (c) M

From Figure 13, it can be noticed that the neural network training history of parameter X and Y showed a good and healthy trendline with its training mean absolute value slightly lower than that of validation set. Besides, Model (X and Y) also made its way through 500 epochs without hitting 50 non-progressive epochs and with significantly less fluctuations as compared to the previous approach. Since, there is relatively high progressiveness and low fluctuation magnitude in the training curve of Model (X and Y), it is believed that the mean absolute error can further be converged and reduced if the training continues. Although there is still an insignificant sign of unknown fit, Model (V) and Model (M) has considerably small gap between the training and validation mean absolute error around lower-error region. As expected, all models perform well on test set and produce a very good overall result with Model (X and Y) at percentage mean absolute error 1.1147%, Model (M) at 1.9967%, and Model (V) at 1.4943%. The overall mean absolute error for this multi-model solution is 1.5352%.

4.3 Summary

The proposed Patch-Based 1DCNN is stated to have its limitation in dealing with unknown and overfitting and this may be the result of excessively deep layers in the model architecture. FFT waveform data is also found to be unfruitful in bringing out distinctive information to characterize each data instance in both cases of single and multi-model approach. The data in the frequency domain might have been subject to distortion by reflections of wave after the collision which is not appropriately treated. For multi-model approach, the lack of real-time incremental and interactive learning in ML algorithms is the main cause of inefficient hyperparameter tuning. In this experiment, the need of ML algorithms for narrow and specific training necessitates further feature engineering to transform the raw wave data from four sensors into an 1D array, which might not promise a good result. On the other hand, a simplified 1DCNN with a self-defined loss function and the most basic architecture of 3 1D convolution blocks and 2 hidden layers yields the best result with the suitable use of target data scaler and activation functions in the final output layer of each model. It can be summarized that a very deep and complex neural network architecture does not necessarily pledge a promising performance especially in regression problems. The overall experimental results of each model from both approaches are tabulated as shown in Table 4.

Table 4 Overall Performance of Experimental Models

			Mean Absolute Error				Overall
			X	Y	M	V	
Patch-Based 1DCNN			11.6219%				11.6219%
Multi-model Approach	ML Algorithms	CART	25.5580%		24.9399%	11.0310%	20.5096%
		CART+GBM			15.1316%	6.8796%	15.8564%
	Simplified 1DCNN		1.1147%		1.9967%	1.4943%	1.5353%

5.0 Conclusion

In conclusion, the main objectives of this study, namely the data processing method development, model training, selection, tuning, evaluation and selection are achieved. Among the 3 experimented models, the best performing model on mean absolute error (percentage) metric is the simplified 1D convolutional (1.5353%), followed by the proposed Patch-Based 1D Convlutional Neural Network (11.6219%), CART+GBM (15.8564%) and CART (20.5096%). While being the only many-to-many (multi-output) model in this experiment, Patch-Based 1DCNN still performed better than all ML algorithms running on one-to-one (single-output) modelling, which has once again proven the advantages of deep learning in terms of tuneability, flexibility. With a switch to multi-model approach and the simplification of proposed model in counter of its overfitting problem, the model has shown massive improvements in its stability, generalizability and predictability which hits the main goal of the experiment that is to achieve a mean absolute error of less than 1.6% benchmarked to the previous similar FDD experiment done by Yu et al. (2019) on damage identification.

Acknowledgment

This study is guided by Prof. Ir. Dr Nahrul Khair Bin Alang Md Rashid who provides encouragement and constructive critics in improving this study.

References

[1] Chao L, Jiafei L, Liming Z, Aicheng G, Yipeng F, Jiangpeng Y, et al. Nuclear Power Plants with Artificial Intelligence in Industry 4.0 Era: Top-level Design and Current Applications—A Systemic Review. IEEE Access. 2020.

[2] Korean Atomic Energy Research Institute (KAERI). (2020). Collision Detection using Vibration Data (Version 1).

[3] Nwankpa L, Jiafei L, Liming Z, Aicheng G, Yipeng F, Jiangpeng Y, et al. Nuclear Power Plants with Artificial Intelligence in Industry 4.0 Era: Top-level Design and Current Applications—A Systemic Review. IEEE Access. 2020.

[4] Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems. 1997;39(1):43-62.

[5] Datta L. A survey on activation functions and their relation with xavier and he normal initialization. arXiv preprint arXiv:200406632. 2020.

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097-105.

[7] Li J, Cheng J-h, Shi J-y, Huang F. Brief introduction of back propagation (BP) neural network algorithm and its improvement. Advances in computer science and information engineering: Springer; 2012. p. 553-8

[8] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.

[9] Glorot X, Bengio Y, editors. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics; 2010: JMLR Workshop and Conference Proceedings.

[10] He K, Zhang X, Ren S, Sun J, editors. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; 2015.

[11] Ioffe S, Szegedy C, editors. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning; 2015: PMLR.

[12] Avci O, Abdeljaber O, Kiranyaz S, Hussein M, Gabbouj M, Inman DJ. A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. Mechanical Systems and Signal Processing. 2021;147:107077.

[13] Abdeljaber O, Avci O, Kiranyaz MS, Boashash B, Sodano H, Inman DJ. 1-D CNNs for structural damage detection: Verification on a structural health monitoring benchmark data. Neurocomputing. 2018;275:1308-17.

[14] Yu Y, Wang C, Gu X, Li J. A novel deep learning-based method for damage identification of smart building structures. Structural Health Monitoring. 2019;18(1):143-63.

[15] Chae, Y. H., Kim, S. G., Kim, H., Kim, J. T., & Seong, P. H. (2020). A methodology for diagnosing FAC induced pipe thinning using accelerometers and deep learning models. Annals of Nuclear Energy, 143, 107501. doi:10.1016/j.anucene.2020.107501.

[16] Chen, F.-C., & Jahanshahi, M. R. (2018). NB-CNN: Deep Learning-Based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion. IEEE Transactions on Industrial Electronics, 65(5), 4392–4400. doi:10.1109/tie.2017.2764844.

[17] Chen C-C, Liu Z, Yang G, Wu C-C, Ye Q. An Improved Fault Diagnosis Using 1D-Convolutional Neural Network Model. Electronics. 2021;10(1):59.