Deep transfer learning for fault diagnosis of roller bearings under scarce data conditions

A Project

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of

Master of Engineering (MEng)

in

Industrial Systems Engineering

University of Regina

By

Mrigank Kumar

Regina, Saskatchewan

August, 2020

Abstract

The tracking of health state of machine has become very significant in the industry in the recent years. Avoiding the equipment failure by predicting and detecting the failure types in advance has significant impacts on decreasing machine downtime and providing large cost savings. In this project we have presented a data driven deep learning method based on transfer learning using ImageNet data for predicting faults under real world conditions has been presented. The availability of fault data is small and the experiments uses the MFPT (machine failure prevention technology) dataset. A convolutional neural network (CNN) model trained on the ImageNet dataset is used for transfer learning purpose by freezing all the weights and removing the last fully connected layer for solving the specified problem with then updating the network. All the methods employed are data driven, hence the requirement of field expert and background equipment operating conditions and related information is not necessary. In this project we have successfully classified three fault types. While the accuracy of neural network from scratch gave accuracy of 68% of the data, while the same dataset trained on VGG16 classifier using transfer learning achieved accuracy of around 90% with significant reduction in loss rate. As shown, this method worked great under conditions where the availability of data is scarce and with various different operating conditions.

Keywords: deep learning, machine fault diagnosis, transfer learning, Vibration signal, Roller bearing

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Predictive maintenance involve sets of techniques utilizing condition monitoring tools to detect and categorize faults prior to a machine or system failing. The significant impact is on cost savings and efficient use of the machine or related operations to eliminate unnecessary or just-in-time maintenance that if neglected would have huge damaging consequences.

Bearings are very critical components of machinery and should be taken care of properly. Unexpected failure of roller bearings can cause machine breakdown followed by high repair and downtime costs. Therefore, tracking health condition of the bearing is of great significance. With the growing trends of Internet of things (IoT) based sensor devices and their low cost availability have led to a huge surge in data generation which can be used to build deep learning models for data driven decision makings. This project seeks to use deep learning to predict failure of bearings for allowing for predictive maintenance practices.

However, industrial applications rarely have readily available historical data for each type of bearing, or sufficient data for developing deep learning models. This project faces challenge of scarce real world industrial historical data on bearing by incorporating a transfer learning approach. For classifying faults in the roller bearings a CNN that was pertained was employed, experiments performed were accessed form the Machine Failure Prevention Technology dataset (MFPT). Also, the approach being investigated was based on data rather than relying on domain experts to manually extract or select features.

1.1 Introduction to roller element bearings

Roller element bearing is integral in rotating machinery and provides support for different loads as the machines undergoes its operation. Bearing allow for free rotation of components by reducing the rotational friction. Unexpected failure of roller bearings could severely damage the machine and adversely affect operation with unplanned or extended downtime. Replacement and repair costs tend to be high. According to a study published by Bradley William Harris 40% of provisional industrial motor failures are caused due to the bearing failure [1].

Figure shows the different component composed in a roller bearing:

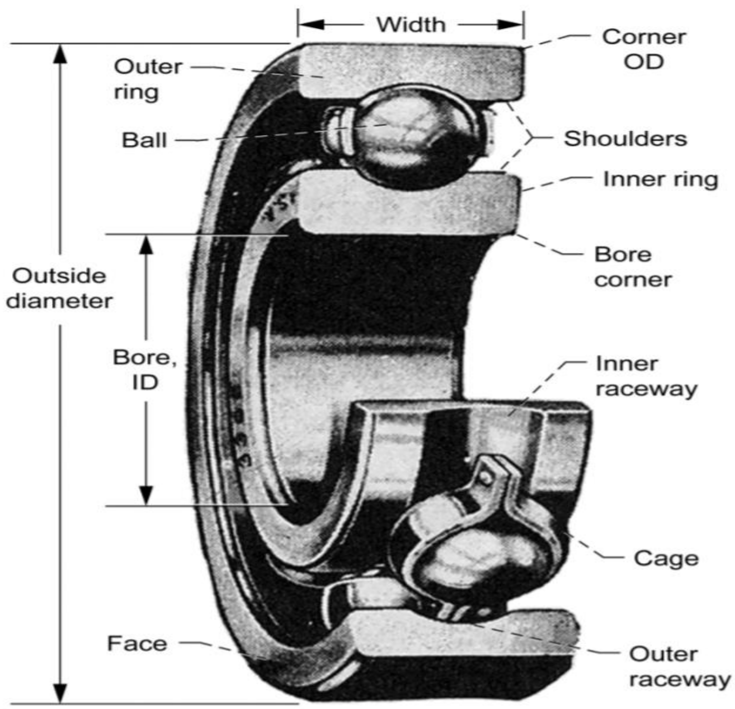

Figure 1 Roller element bearing component. Image reproduced from [2]

Generally, the operation of rolling bearings is unaffected by the working temperatures. They have low static friction. The prime cause of the bearing failure is metal fatigue. This happens under extreme operating conditions of excessive, rapid loading and unloading, corrosion, misalignment, overheating, having unlubricated parts and/or contamination. [3]

1.2 Identification of failures

Vibration is critical in diagnosing roller bearing failures; a characteristic squeal is heard when they fail. Predicting failure or the degree of bearing deterioration is possible by examining vibration peaks in the frequency spectrum of the bearing component. By analyzing the spectral amplitudes, failure identification becomes fast and easy. Bearing failures typically occur at the following frequencies [4]:

a) Ball spin frequency:

   Which is defined by how many spins the roller makes with one full rotation the shaft.

   Ball spin frequency = $\frac{f_{r\,.\,D}}{2d}\ (1 - (\frac{d}{D}\cos\alpha)^2)$

b) Fundamental Train frequency(cage speed):

   which is defined as the number of spins the cage makes with full rotation of shaft.

   FTF = $\frac{f_r}{2}\ (1 - \frac{d}{D}\cos\alpha)$

c) Ball pass frequency of inner race frequency:

   Which is defined as with one rotation of shaft number of balls passes through an inner ring.

   BPFI= $\frac{f_{r\,.N_E}}{2}\ (1 + \frac{d}{D}\cos\alpha)$

d) Ball pass frequency, Outer race:

   Which is defined as with one rotation of shaft number of ball that passes through outer ring.

BPFO = $\frac{f_r . N_E}{2}\left(1 - \frac{d}{D}\cos\alpha\right)$

where f_r is the shaft speed, N_E is the number of rolling elements, α is the angle of load from radial plane and D =$\frac{D_1+D_2}{2}$. The angle α varies with the position of the each roller element in the bearing. Figure 2 illustrates these values.

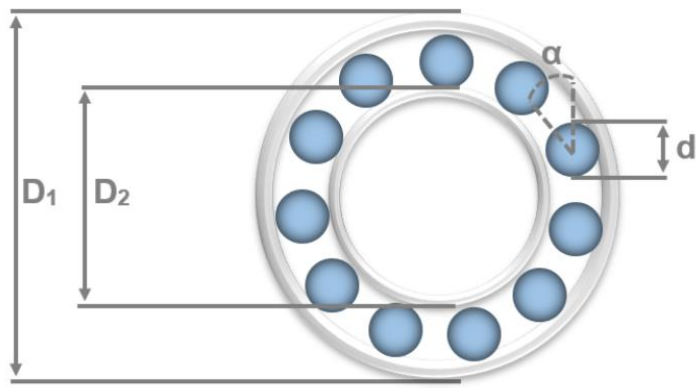

Figure 2 Failing frequencies of bearing element. Image reproduced from [23]

Vibration analysis for roller bearing failure is shown in Figure 3. The peaks in the characteristic frequency of the bearing represents the failure frequency of the component; these peaks appear in the frequency spectrum are above the pre-defined threshold amplitude for a given bearing. Upon detected, a warning alarm should be deployed. If the peaks crosses the second threshold the critical alarm should be deployed which warns, of imminent failure.

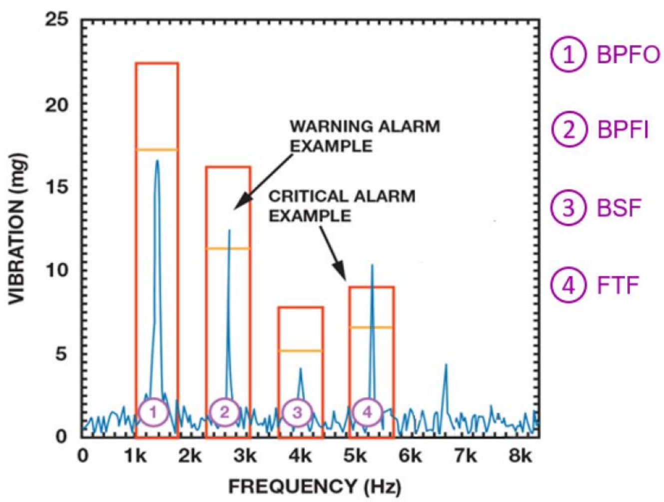

Figure 3 Vibration Analysis to Detect Bearing Failure. Image reproduced from [23]

1.3 Levels of Bearing Damage

Severity of bearing failure depends on multiple factors, but the most common failure is to the outer ring. The other kinds are the inner ring, rolling element and cage failures. The expected life of bearings with outer race fault is longer than bearings with inner race fault. When comparing the characteristic amplitudes of the outer and inner race fault, the amplitudes of the outer race was found to be higher. This maybe because the placement of the sensor devices is situated near to the outer race, therefore the signal from the inner race has to make its journey all the way from rolling element through the cage and the outer race. The distance travelled by the signal teds to attenuates and lower the amplitudes. Categorizing of failures can be done in three stages [3]:

1st stage: the peaks in the amplitude appears to be low, caused by the degradation in the visible frequency bands.

2nd stage: the magnitude of the high frequency harmonic bands gradually increases.

3rd stage: the acceleration at high frequency appears with background noise; in this final stage the newer types of failure might develop implying greater level of fault severity.

1.4 Description of problem

This project aims to classify and identify the condition of the bearing and determine if there appears to be any kind of failure, and if so identify the type. The requirements are as follows:

a) For classification of bearing failure the model based approach is sufficient in the cases where the coefficient of bearing failure is given and shaft velocity is known. Therefore the project focuses on the classification tasks when shaft velocity is not given or maybe variable. Thus model must work well for different rotating velocities.

b) Since the bearing coefficient are not known the model must not make use of such features.

c) As actual conditions may not have access to lots of historical data, the model must work well with limited or small training datasets.

1.5 Data Structure

This section discusses the structure in which the data is used for building the model.

Machine failure prevention technology (MFPT) dataset, prepared by Dr. Eric Bechhoefer, chief engineer of NRG Systems is a publically available and served as the dataset for the project. The dataset contains a snapshot of vibration signals for given time intervals. An accelerometer was installed on the housing of the bearing to record the signals. A vibration signal waveform is shown in figure 4.

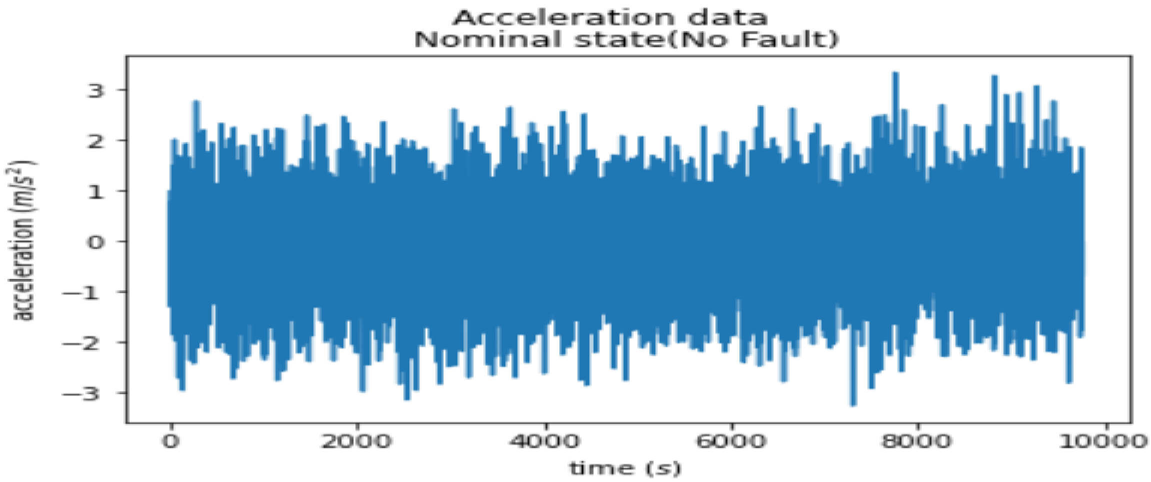

Figure 4 Acceleration Waveform from MFPT dataset.

Chapter 2: Literature review

Various techniques employed to classify faults in the roller bearing are presented. Model based and data based fault classification methods are currently used. To determine the underlying physical processes and predicting the behavior of the system, the model based mathematical model based approaches are used. These fault classification methods requires expert knowledge and a significant amount of detailed system information. Prior understanding of shaft speeds and the coefficient of bearing failures are required to implement the spectrum analysis of fault types. Analyzing the vibration data using the frequency spectrum of bearing failure is one the examples for the model based approach mentioned in section 1.3. For data driven approaches there is no requirement of field expertise to understand the underlying physical behavior of the system. With the recent development of low cost IoT sensors and the large amounts of data generated these methods are gaining greater attention for advancing signal processing techniques for fault detection. The following summarize state of art key advances in this field of study and practice.

Saufi, S. R., Ahmad, Z. A. B., Leong, M. S., & Lim, M. H. (2019). Provides the general overview of how the deep learning is being employed in the era of Industry 4.0. The operational time of the machine components will increase and consequently the risks of breakdowns increases and the need to perform predictive maintenance by deep learning based model increases. Also, the requirement of expert knowledge for conducting fault diagnosis and detection which is labor intensive and costly is discussed. Implementation of automated feature learning processes the potential for solving these problems have improved.

Verstraete, D., Ferrada, A., Droguett, E. L., Meruane, V., & Modarres, M. (2017). Presented a deep learning based automatic feature extraction technique using time frequency representation of raw data and converting them to images for input into CNN using time-frequency analysis, for fault classification. This removes the labor intensive process of requiring the field expert to manually extract the features. This method deals very well with the experimental noise and produces good results with less learnable parameters.

Zhang, R., Tao, H., Wu, L., & Guan, Y. (2017). Proposed the fault classification using transfer learning, since, real world scenarios and working conditions are never ideal, but it changes with respect to time. Also the data is difficult to obtain and also limited. In this approach the CNN is firstly trained with massive sources of data with adjustments to parameters as the network learns, next the learned parameters are transferred to the target task in some different working condition. Finally the training of the model is done in some other working condition with less amount of target data.

Yuan, L., Lian, D., Kang, X., Chen, Y., & Zhai, K. (2020). Reported on an approach to using support vector machine for fault classification. The training of datasets of large size are difficult to attain, so the first step uses time frequency analysis images by continues wavelet transform of vibration signals. Then these generated images are feed as input to the CNN for developing the model used to analyze the fault location and their level.

Hon, M., & Khan, N. M. (2017, November) have used of computer vision using transfer learning for detecting the Alzheimer's disease (AD) form neuroimaging data using deep learning models. The state of art architectures inception and VGG was used for building their detection model, using their pretrained weights. The fully connected layer is retrained with only a small number of MRI images. Through experimentation better results were reached even though the size of training sample was very smaller than state of art architecture.

Liu, H., Yao, D., Yang, J., & Li, X. (2019). Proposed a method for roller bearing fault diagnosis under the conditions where the load is constantly changing. It has made use of ShuffleNet V2 with batch normalization and L2 regularization. STFT (short time Fourier transform) is used to convert one dimensional time signal into 2D time frequency graph. This comprised of one unit for spatial down sampling and the other for feature extraction. The model is constructed with building units that are repeatedly placed on top one another. This method has produced a very good results in diagnosing bearing faults.

Wang, R., Meng, X., Xiong, B., & Wang, Z. (2021). In this experiment they have made use of multi-view time frequency analysis of STFT images in fault diagnosis. In this experiment the pretrined network is used for training the model using same segment of the vibration signal that is generated by the different analysis method for time frequency domain. This method have better results in generalization by substituting global max pooling layer in place of flatten layer.

Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2018). Proposes a method for fault diagnosis of bearings under conditions where data is not labelled, and also for the conditions where massive labeled data is available but it fails to classify the unlabeled data form different source. This method had made for use of transfer learning framework for domain adaptation of target task from the source task by deploying deep convolutional transfer learning network (DCTLN). For recognition of the condition of the bearings by automatically learning the features the 1D CNN is constructed for condition recognition. Then in the next step maximization of domain recognition errors and minimizing of the probability distribution distance is done with facilitating the 1D domain adaptation module, this produces very high accuracy in classifying faults in the conditions where data is unlabeled.

Yang, Z., Baraldi, P., & Zio, E. (2016, October). In this paper they have compared the experiment machine learning technique and artificial neural network for predicting the remaining useful life of the machine component. These two methods are compared on the turbofan engine where prognosticα metrics that were used are RMSE, accuracy index, α-λ metric, and the required time for carrying out training of the model. It showed that this models can be employed under conditions where model is updated periodically with constantly changing working conditions.

Niyonambaza, I., Zennaro, M., & Uwitonze, A. (2020). It this paper it was proposed how predictive maintenance in the hospital of Rwanda has made significant social and economic impact, it helped in reducing the unnecessary down time of the equipment by predicting the failure of the equipment, hence reducing the unplanned crisis and also helped to reduce the wait time of patients in queues. I t has made use of IoT based sensor devices mounted on the machine equipment collecting real time data then building long short term memory (LSTM) neural network model for predicting of failure with 90% accuracy.

Chapter 3: Background

3.1 Convolutional deep learning neural network

Neural networks is the archetype for data preprocessing inspired by the biological nervous system [15], it consists of large network of interconnected neurons called nodes processes the data. The deeper the layer more the number of neurons is constituted in the network. The two types of deep learning paradigm include supervised and unsupervised learning. Supervised learning is when the data is labelled prior to inputting it to the algorithm. Predictions can be made about the unlabeled data by running the classification or regression against these given labels.

a. CNN (convolutional neural networks)

   Amongst all the types of available neural network CNN is strongest in the computer vision community as it surpasses other networks for generalize purposes tasks. There are multiple stages in the construction of CNN, and its architecture is built of multiple layers, pooling, weights shared across and local connections. CNN specializes in processing the high dimensional array data types such as time series data or images i.e. one and two dimensional data types. In CNN operation a filter matrix (kernel) is passed though the input matrix to generate the features for the next layer in the convolution operation. The filter slides over the input matrix, this mathematical operation is called convolution. Feature map is created by summing all the element multiplying at all the locations in the matrix. The mathematical representation of convolution operation is given as follows [16]:

$$S(i,j) = \sum_m \sum_n I(m,n)k(i-m,j-n)$$

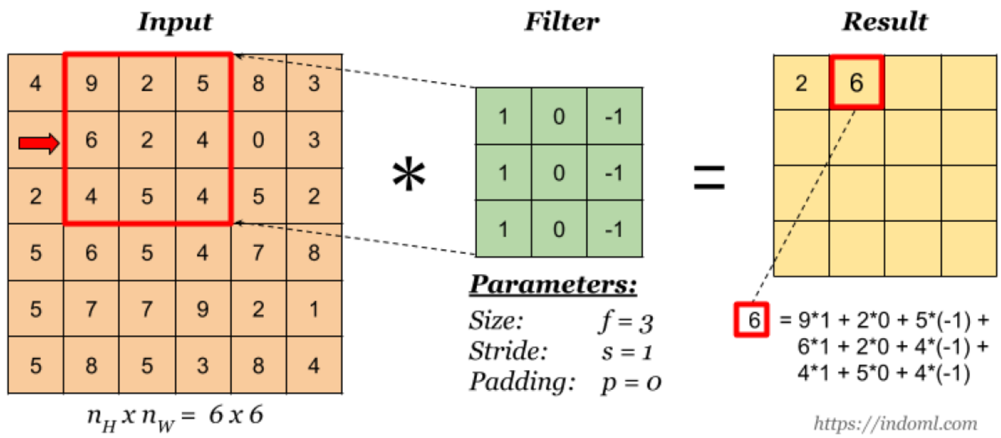

Figure 5 Element wise multiplication and sum for feature mapping in convolution layers. Image reproduced form [16]

b. Non-linear activations (ReLU)

To perform nonlinear activation of the input layer, an activation functions is applied to the nodes of the succeeding convolution layer. The ReLU (rectified linear unit) activation gives zero output for all the other inputs except the positive ones.

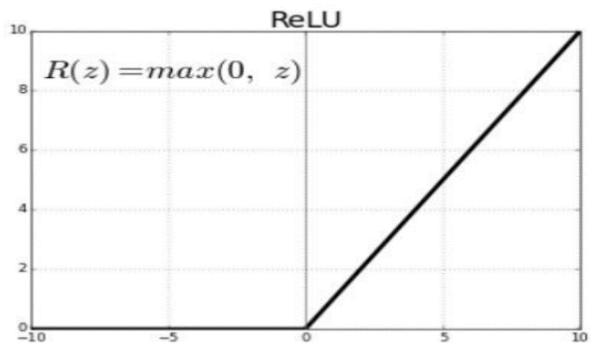

Figure 6 ReLU activation function. Image reproduced from [17]

b. Pooling layer

The precise locating of the input features is the setback of the feature map output form the convolution layer. This layer yields different outputs every change such as when cropping or rotation is done on the input images. Down sampling is one the remedies for dealing with these problems, this is resolved by associating pooling layer after the nonlinearity layer. Minor translations of the inputs results in representation that are generally invariant, i.e. the majority of the pooled outputs do not change significantly when a small changes in the input translation occurs.

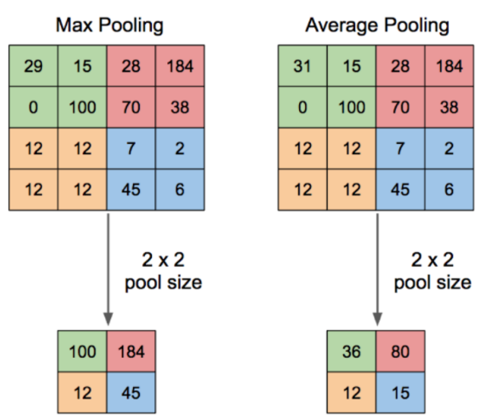

Figure 7 Pooling operation types. Image reproduced from [17]

d. Fully Connected Layer

The output of the pooling layer that is situated at the end of the neural network acts as the input to the fully connected layers. Many layers associated with this can exist the nodes of the first layer are fully connected to the nodes in the second layer as shown in figure 3.4

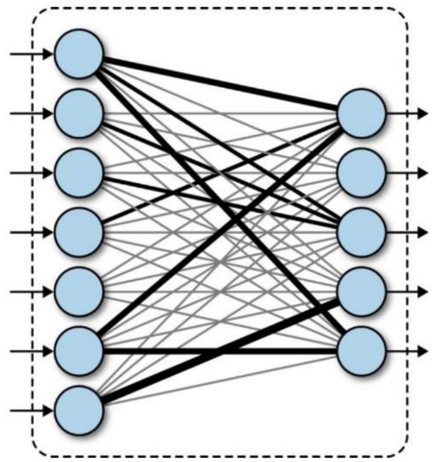

Figure 8 Fully Connected Layer. Image reproduced from [17]

3.2 Time Frequency Analysis Methods

Representation of the raw signals form is done in both the time and frequency domain. This is typically done using spectrogram and scalograms. Frequency resolutions of the spectrograms remain constant, which lean on to the window size; spectrogram is the visualization of time and frequency signals using a short-time Fourier transform. The axis demonstrates the signals in the time, and frequency domain and the color scale of the images illustrates the amplitudes of the frequency. STFT is the most forthright frequency analysis. It is a depiction of a series of sinusoidal waves. Figure 3.5 represents the spectrogram for raw signal:

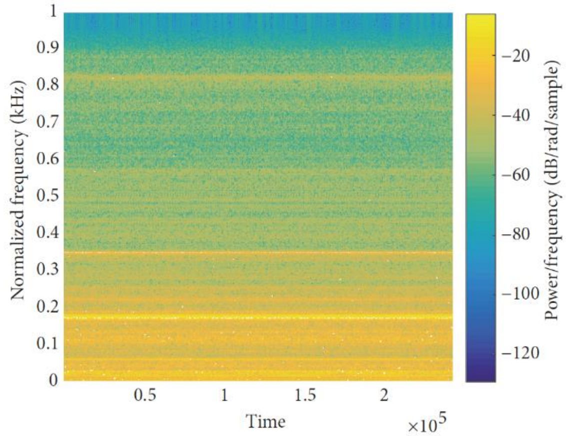

Figure 9 Raw signal baseline spectrogram (STFT). Image reproduced from [18]

3.3 Transfer Learning with ImageNet Pretrained Models

The domain transfer of the learned characteristics of one activity to another is the basis of transfer learning. In other words, a kind of learning influences other similar kinds of learning. The various methods of transfer learning are parameter based, relationship-based, case-based, and feature-based [17]. Transfer learning basically tries to apply the knowledge from previously learned model D (source) to model T (Target). The parameters of the pretranined model with weightings that were trained previously.

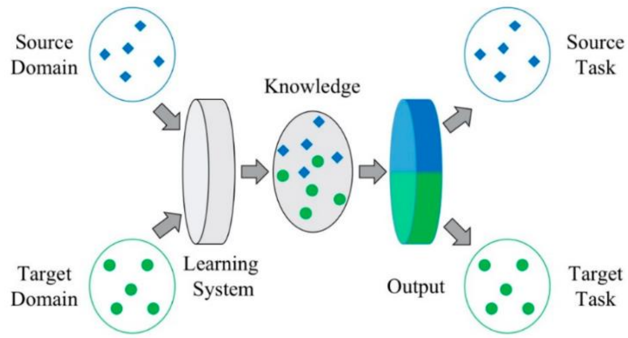

Figure 10 Transfer Learning Architecture. Image reproduced from [19]

The parameters that were transferred to the target domain will now be used for retaining and fine-tuning to identify the sample labels from some other anonymous dataset. The architecture on CNN consists of weights that are randomly initialized and updated constantly according to the given sample dataset and the loss function. This leads to substantial computational complexity, and the process becomes very tedious. Transfer learning effectively resolves these problems using pretrained neural networks that used anonymous dataset of similar representation. Since the CNN is the learning framework based on learning features from image representations in a series of hierarchical order, weights of the knowledge from the pretrained models can be used to classify new tasks.

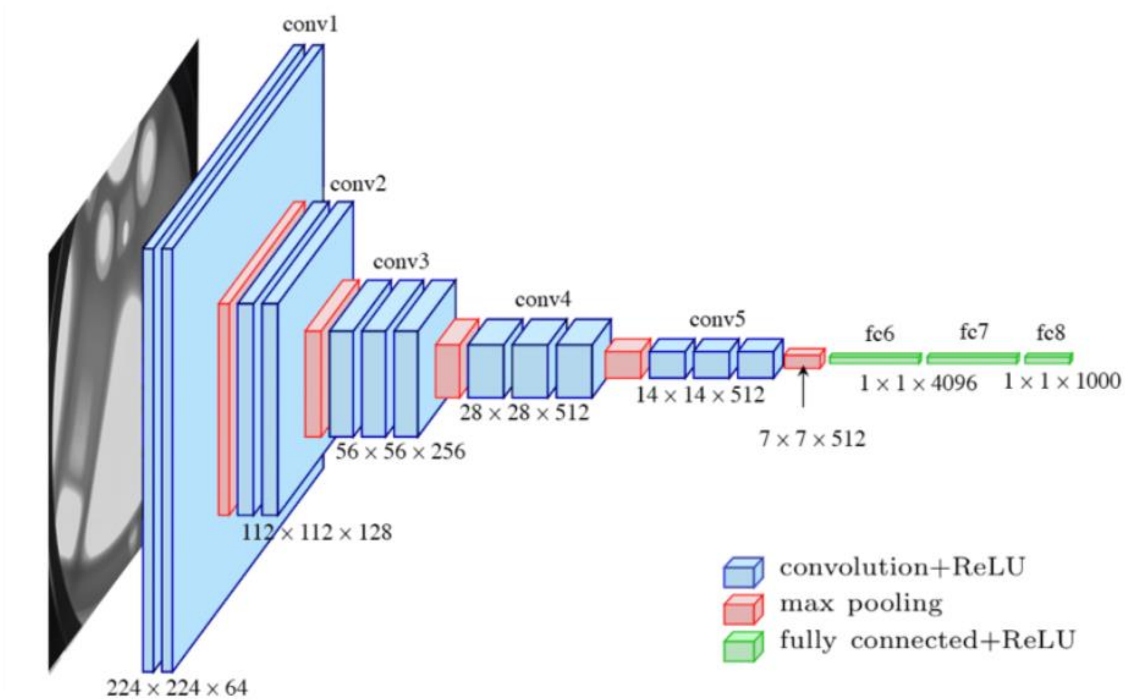

Figure 11 Architecture of VGG-16 model. Image reproduced from [20]

ImageNet was a project developed on a vast image database for classification purposes of various categories for research. Different pertained models trained on the millions of images form the ImageNet database are InceptionV1, VGG-16, and InceptionV2; they are prepared to classify thousands of distinct categories of images. These models are developed from scratch and trained in a high GPU environment for months. Knowledge transfer is done by sharing low-level features such as spatial, edge, rotation, and shapes trained on millions of images. This acts as a feature extractor for computer vision experiments with totally different images than the dataset.

3.4 Performance Accuracy Measures (confusion matrix)

Confusion matrix [24] is used for classifying the performance accuracies of the model, the matrix correlates the vales that were predicted by the model and the actual values. This provides a comprehensive view of the kinds of errors the model is making when classifying the types. The figure below illustrates the 2x2 confusion matrix:

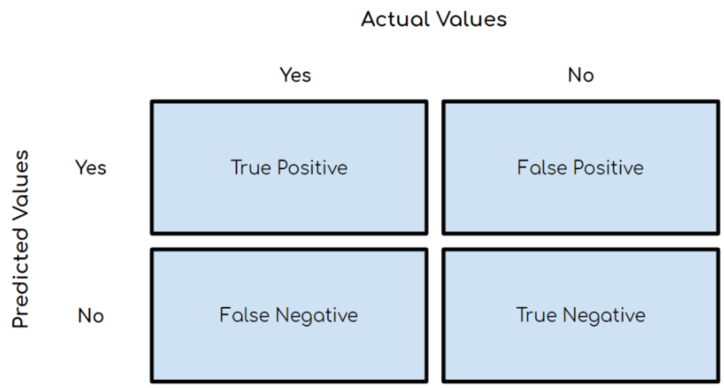

Figure 12 2x2 confusion matrix. Image reproduced from [24]

a) **True positive**: when the model correctly identifies the actual positive values.

b) **True negative**: when the model correctly identifies the actual negative values.

c) **False positive**: when the model falsely identifies the actual positive values.

d) **False negative**: when the model falsely identifies the actual negative values.

e) **Positive values**: the correct classification of a particular fault type (N, OR, IR).

f) **Negative values**: the correct classification that it's not a particular fault type.

Accuracy:

The equation that represents that if the model classification task was done successfully:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

Precision is the ratio of how many positive values were actually true; it is also called "positive predicted value".

$$precision = \frac{TP}{TP + FP}$$

Recall:

Recall is the rate of true positive i.e. the ratio of the instances that were actually recovered to the total amount of the instances.

$$recall = \frac{TP}{TP + FN}$$

Specificity:

Specificity is the opposite of recall, it gives the ratio of all the negative values that were rightly identified (rate of true negative).

$$specificity = \frac{TN}{TN + FP}$$

F1 score:

This score indicates the preciseness and robustness of the model performance; the maximum value is 1(perfect score) and the minimum value is 0.

$$F1\ score = \frac{2TP}{2TP + FP + FN}$$

Chapter 4: Details of implementation

All of the models used and the dataset both the details and the preprocessing tasks are provided for the transfer learning project.

4.1 Data set (MFPT)

The machine failure prevention technology (MFPT) dataset is used to test our bearing fault classification system using transfer learning. This data set is provided by NRG systems by Dr. E.B., chief engineer [21]. The dataset comprises of bearing testing under various conditions of baseline, inner race fault and outer race faults, with 270lbs load and shaft rate of 25Hz. The data sampling rate was 97,656sps for 6 seconds for baseline condition; and 48828sps for 3 seconds for inner and outer race faults with loads varying at 25, 50, 100, 150, 200, 250 and 300 pounds. The data is comprised of 3 base line conditions with 1,757,808 data points, 7 inner race fault condition with 1,025,388 data points and outer race fault conditions with 1,025,388 data points.

The Table 1 shows the parameters of the bearings used in the dataset.

Table 1 Parameter of test bearing. Table from [21].

Parameters	Value
Roller diameter	0.235
Pitch diameter	1.245
Number of elements	8
Angle of contact	0

4.2 Preprocessing Techniques

Data needs to be prepared according to the real-world simulations to feed into the CNN model to achieve good performance. The preprocessing for this project involved adding a sliding window for cropping the signals into a set number of timestamps (or intervals). This step ensures that input is of fixed size. The timestamps that are generated were comprised of a minimum two periods of the rotating frequency of the machine signal; if the inputs are massive, the training of the CNN could be very complex and difficult. The data was then normalized and rescaled in the range of [0, 1] to aid in model convergence. This also limits the model learning just from the amplitude of the waveform.

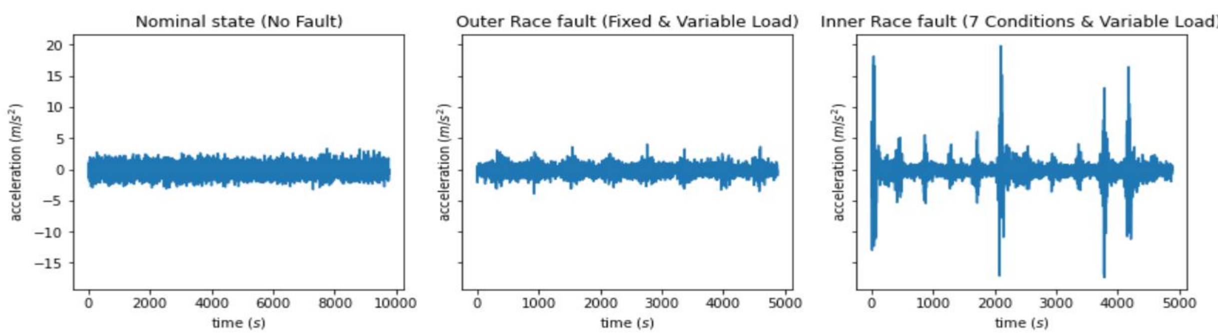

Figure 13 Interpretation of waveform signal data

For generating the spectrogram image, the Scipy library in Python was used. A larger number of data points for the Outer race fault than the other conditions exists; hence the number of spectrogram images generated for the outer race failure exceeds those for the inner race and nominal conditions. The STFT (Short Time Fourier Transform) spectrogram images generated for feeding into the CNN is illustrated in Figure 14.

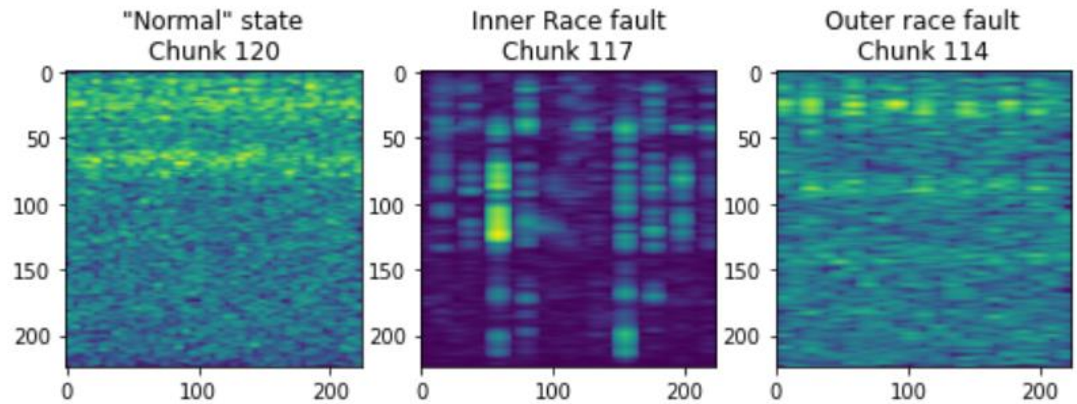

Figure 14 Short Time Fourier Transformed image of training data (224x224) (spectrogram)

4.3 CNN Trained form the Beginning (no transfer)

Applications of the convolutional neural network (CNN) include recognition of images, object detection, and video processing. The vibration signals received from the accelerometer are recorded and then stored in the form of images to feed it into the CNN for the classification tasks. The signal is preprocessed and stored as spectrogram images (STFT); in this application, the data was divided into chunks, i.e. subsets of data points. This method is also called tumbling time windowing. The method is more efficient method because the dataset consists of a huge number of data points, and uploading the entire set would drain available computational resources and processing would be inefficient.

The created training images were divided with 30% used for training and 70% used for testing. The architecture of the CNN was borrowed from [22]. After training the CNN model, the feature extraction was done through Softmax activation to classify the image into inner race failure, outer race failure and nominal bearing conditions. The CNN was trained for 50 epochs because after a certain number of epochs, the network's learning appeared saturated. All the training and testing was done in TensorFlow.

Table 2 structure of neural network trained form scratch

Layer	**Type**	**Output shape**	**Parameters**
Conv2D_1	Conv2D	96*96*3	505
Leaky_relu_1	LeakyReLU	96*96*3	0
Conv2d_2	Conv2D	96*96*10	1260
Leaky_relu_2	Leaky_ReLU	96*96*10	0
Max_pooling2d_1	Maxpooling	48*48*10	0
Conv2D_3	Conv2D	48*48*10	2510
Leaky_relu_3	Leaky_ReLU	48*48*10	0
Max_pooling2d_2	Maxpooling	24*24*10	0
Flatten_1	Flatten	5760	0
Dense_1	Dense	10	57610
Leaky_relu_4	Leaky_ReLU	10	0
Dense_2	Dense	5	55
Leaky_relu_5	Leaky_ReLU	5	0
Dropout_1	Dropout	3	0
Dense_3	Dense	3	18
Activation_1	Activation	3	0

4.4 CNN with transfer learning (VGG16)

The VGG16 architecture was trained on images of a particular specifications. The size of the images was 224*224 in RGB channel. Prior to presenting the spectrogram images into the CNN, the image was channeled down to 224*224 resolution using ImageDataGenerator. The weights were imported from the pretrained VGG16 model. The weights for the input as well as the output layer are discarded to enhance the suitability of the architecture suitable to the bearing classification problem. The images were split into 30% for training and 70% for testing. The last layer consisted of only three outputs for classifying the three bearing faults using 'Softmax' as the activation layer function. This is a significant reduction from the original 000 classes used by the VGG16 classifier. All other parameters were unchanged.

Table 3 Proposed modified transfer learning structure using VGG16

Layer	***Type***	***Size (channels)***	***Output***	***Parameters***	***Stage***
Input	Data	-	224*224*3	0	Transferred
Block1-conv2	convolution	3*3-64	224*224*64	1792	Transferred
Block1-conv2	convolution	3*3-64	224*224*64	36928	Transferred
Block1-pool	Maxpooling	2*2	112*112*64	0	Transferred

Block2-conv2	Convolution	3*3-128	112*112*128	73856	Transferred
Block2-conv2	Convolution	3*3-128	112*112*128	147584	Transferred
Block2-pool	Maxpooling	2*2	56*56*128	0	Transferred
Block3-conv1	Convolution	3*3-256	56*56*256	295168	Transferred
Block3-conv2	convolution	3*3-256	56*56*256	590080	Transferred
Block3-conv3	convolution	3*3-256	56*56*256	590080	Transferred
Block3-pool	Maxpooling	2*2	28*28*256	0	Transferred
Block4-conv1	convolution	3*3-512	28*28*512	1180160	Transferred
Block4-conv2	Convolution	3*3-512	28*28*512	2359808	Transferred
Block4-conv3	Convolution	3*3-512	28*28*512	2359808	Transferred
Block4-pool	Maxpooling	2*2	14*14*512	0	Transferred

Block5-conv1	convolution	3*3-512	14*14*512	2359808	Transferred
Block5-conv2	convolution	3*3-512	14*14*512	2359808	Transferred
Block5-conv3	convolution	3*3-512	14*14*512	2359808	Transferred
Block5-pool	Maxpooling	2*2	7*7*512	0	Transferred
Fully connected layer 1	Fully-connected	1*1*4096	4096	102764544	To be trained
Fully connected layer 2	Fully-connected	1*1*4096	4096	16791312	To be trained
Output	Fully connected	1*1*class	C	4096*c+c	To be trained

Chapter 5: Results of the experiment

A CNN model trained form scratch displayed in the figure below was used to train and the vibration signal images to classify faults, the images were of resolution 64x64 spectrograms. The data was divided into 30% train and 70% test, the model was trained for 200 epochs. The accuracy of the model was around 68% and model loss was also high which indicates that model was bad in predicting the fault classes for small dataset.

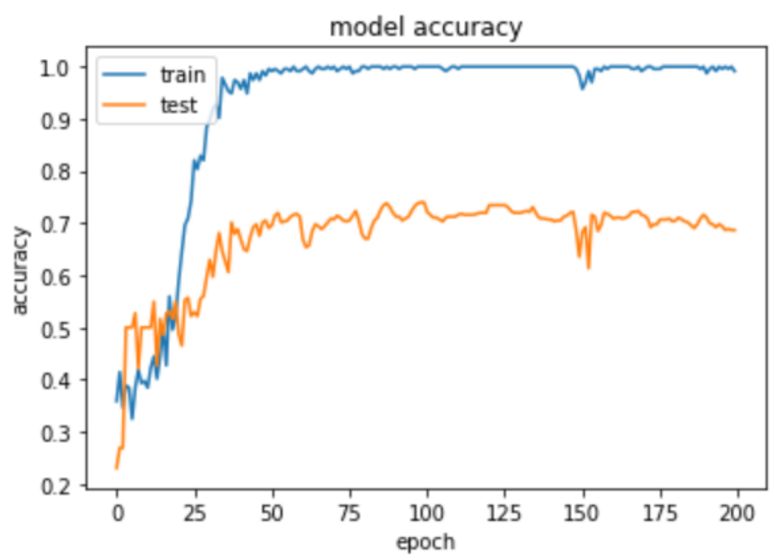

Figure 15 Accuracy performance statistics of CNN

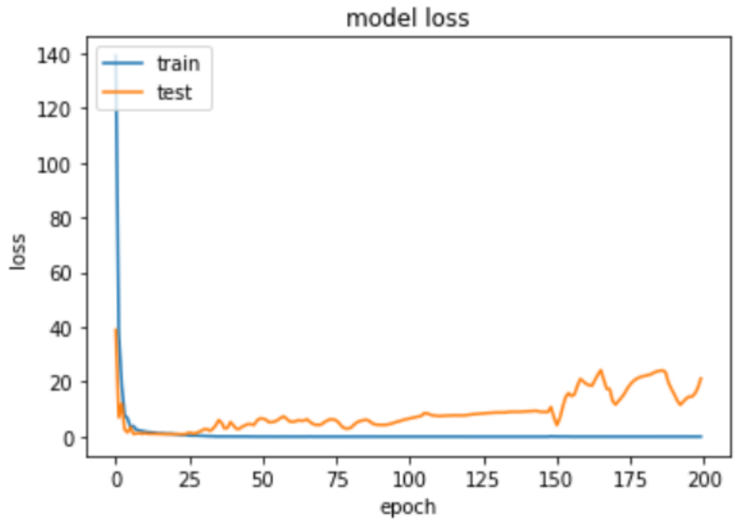

Figure 16 loss rate of CNN

Training the network with ImageNet weights using vgg16 architecture lead to huge improvement in the accuracy of the model the accuracy was around 90% on validation data, which was trained for 200 epochs. Only 234 images of nominal, outer race and inner race was used for training the data and validation was done for 546 images, which confirms that the transfer learning framework works well in classifying fault types for small datasets. The figures 5.3 demonstrates the accuracy metrics and loss curves of the model training and testing.

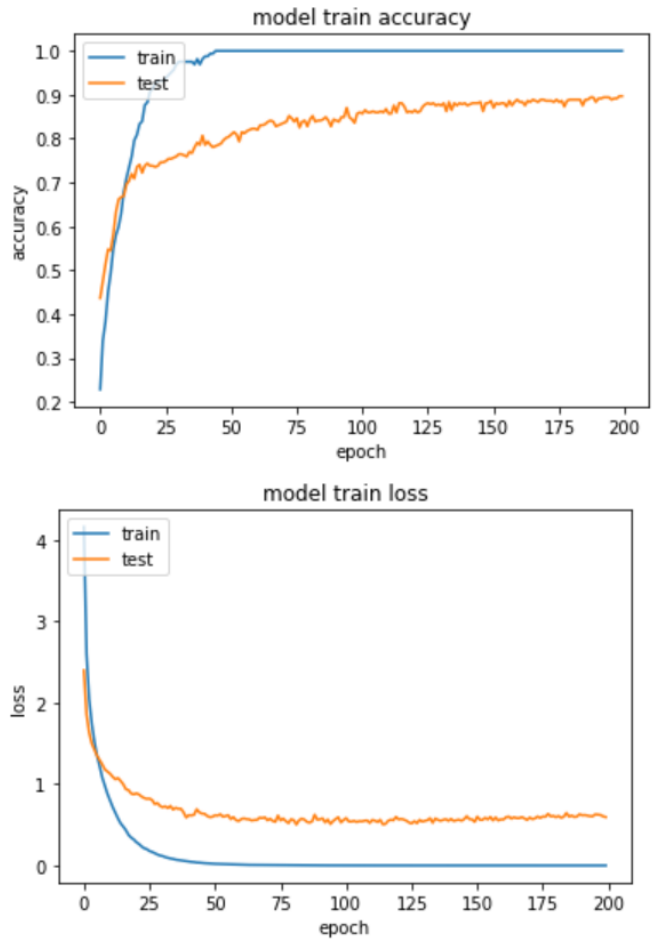

Figure 17 Accuracy performance and the loss performance with transfer learning.

As we can observe the loss curve reduces with increasing number of epochs, this shows that model is well suited to make predictions. CNN architecture has lower accuracy than one trained with ImageNet weights using transfer learning. The table below demonstrates the confusion matrix of the fault classification for fault types where OR = outer race, IR= inner race and N= nominal conditions of the roller bearing.

Table 4 Confusion matrix of predicted classes with transfer learning on ImageNet pretrained model

	N	IR	OR
N	126	0	0
IR	0	123	24
OR	19	28	226

Table 5 Confusion matrix of predicted classes of CNN trained from scratch

	N	IR	OR
N	126	0	0
IR	3	117	27
OR	126	15	132

Table 6 performance of the models

	Accuracy
CNN from scratch	68%
Transfer learning using vgg16	90%

Chapter 6: Conclusions

Unexpected failure of machinery is costly in terms of repair costs, downtime and other environmental impacts. Rolling bearing elements is a significant part of machinery that affects machinery operation. Monitoring bearing health is one of the main research concerns in recent times as it tends to be the root causes of many equipment failures.

Predictive maintenance is detecting the potential defects of the rolling elements before planned maintenance activity. Vibration analysis is one the most common method employed in the industry for fault detection. This method is model-based. The method is simple, but it requires background information of rotating velocities and failing coefficients, which can vary based on operating conditions. Therefore, a new approach is required when condition where data availability is scarce or attributing background information is difficult to attain.

For solving this challenge, deep learning-based approach that is data-driven and does not require past information of the system was proposed. The model was designed to be robust and perform well under different operating conditions with varying rotation velocities. Furthermore, the deep learning model for classifying faults operates on when available characterizing data is scarce (real-world industrial applications). Vibration signals data generated from accelerometers in the form of a waveform was used as training data.

This project used the time frequency-based image analysis method for training the CNN with transfer learning to deal with real-world conditions of small data availability. Firstly, the short time Fourier transformed images were generated from raw vibration data signals which were feed to the neural network trained from initial set of random weights (scratch condition). A second CNN was constructed with transfer learning using ImageNet weights with VGG16 architecture. These methods were compared. The transfer learning framework using VGG16 produced very good results in classifying faults in small training data with an accuracy of 90% and low model loss; whereas, the CNN trained from scratch had an accuracy of 68% with significantly high model loss. The training and testing of the models were done for a small batch of 32 for 200 epochs and was executed in Python programming language with Keras and Tensorflow.

Future research needs to enhance the model's accuracy under noisy data sources that parallels actual industrial situations. Moreover, combining other fault data sources generated using generative adversarial networks (GAN) and computer simulations software using transfer learning could make the model more practical and robust. More datasets with other fault databases should also be considered.

References

Harris, B. W. (2013). *Anomaly detection in rolling element bearings via two-dimensional Symbolic Aggregate Approximation* (Doctoral dissertation, Virginia Tech).

https://www.researchgate.net/figure/Deep-groove-ball-bearing-details-a-Sketch-b-Cross-section-showing-internal_fig1_254336288

Alfonso Fernandez. Vibration analysis learning. 2017. URL: https://www.powermi.com/content/rolling-element-bearings.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical systems and signal processing*, *25*(2), 485-520.

Hon, M., & Khan, N. M. (2017, November). Towards Alzheimer's disease classification through transfer learning. In *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)* (pp. 1166-1169). IEEE.

Lightweight Convolutional Neural Network and Its Application in Rolling Bearing Fault Diagnosis under Variable Working Conditions

Verstraete, D., Ferrada, A., Droguett, E. L., Meruane, V., & Modarres, M. (2017). Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. *Shock and Vibration*, *2017*.

Zhang, R., Tao, H., Wu, L., & Guan, Y. (2017). Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. *IEEE Access*, *5*, 14347-14357.

Yuan, L., Lian, D., Kang, X., Chen, Y., & Zhai, K. (2020). Rolling bearing fault diagnosis based on convolutional neural network and support vector machine. *IEEE Access*, *8*, 137395-137406.

Wang, R., Meng, X., Xiong, B., & Wang, Z. (2021). Multi-view Bearing Fault Diagnosis Method Based on Deep Learning. In *Journal of Physics: Conference Series* (Vol. 1757, No. 1, p. 012093). IOP Publishing.

Saufi, S. R., Ahmad, Z. A. B., Leong, M. S., & Lim, M. H. (2019). Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review. *Ieee Access*, *7*, 122644-122662.

Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2018). Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics*, *66*(9), 7316-7325.

Yang, Z., Baraldi, P., & Zio, E. (2016, October). A comparison between extreme learning machine and artificial neural network for remaining useful life prediction. In *2016 Prognostics and System Health Management Conference (PHM-Chengdu)* (pp. 1-7). IEEE.

Niyonambaza, I., Zennaro, M., & Uwitonze, A. (2020). Predictive Maintenance (PdM) Structure Using Internet of Things (IoT) for Mechanical Equipment Used into Hospitals in Rwanda. *Future Internet*, *12*(12), 224.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). San Francisco, CA: Determination press.

Glinserer, A. (2021). *Autopruning with Intel Distiller and Evaluation on a Jetson Xavier AGX* (Doctoral dissertation, Wien).

Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(10), 143-150.

Verstraete, D., Ferrada, A., Droguett, E. L., Meruane, V., & Modarres, M. (2017). Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. *Shock and Vibration*, *2017*.

Li, X., Hu, Y., Li, M., & Zheng, J. (2020). Fault diagnostics between different type of components: A transfer learning approach. *Applied Soft Computing*, *86*, 105950.

Team, T. and Team, T., 2021. *The Architecture and Implementation of VGG-16*. [online] Towards AI — The Best of Tech, Science, and Engineering. Available at: <https://towardsai.net/p/machine-learning/the-architecture-and-implementation-of-vgg-16> [Accessed 4 August 2021].

Bechhoefer, E. (2013). Condition based maintenance fault database for testing diagnostics and prognostic algorithms. *MFPT Data*.

Guo, T., Dong, J., Li, H., & Gao, Y. (2017, March). Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (pp. 721-724). IEEE.

Domingo Colomer, L. (2020). Deep learning for predictive maintenance of rolling bearings.

Medium. 2021. *Understanding the Confusion Matrix and How to Implement it in Python*. [online] Available at: <https://towardsdatascience.com/understanding-the-confusion-matrix-and-how-to-implement-it-in-python-319202e0fe4d>.

Appendix A

- Short time Fourier spectrogram images of "normal conditions". Figure 18 represents the one hot encoding label for the images of different chunks.

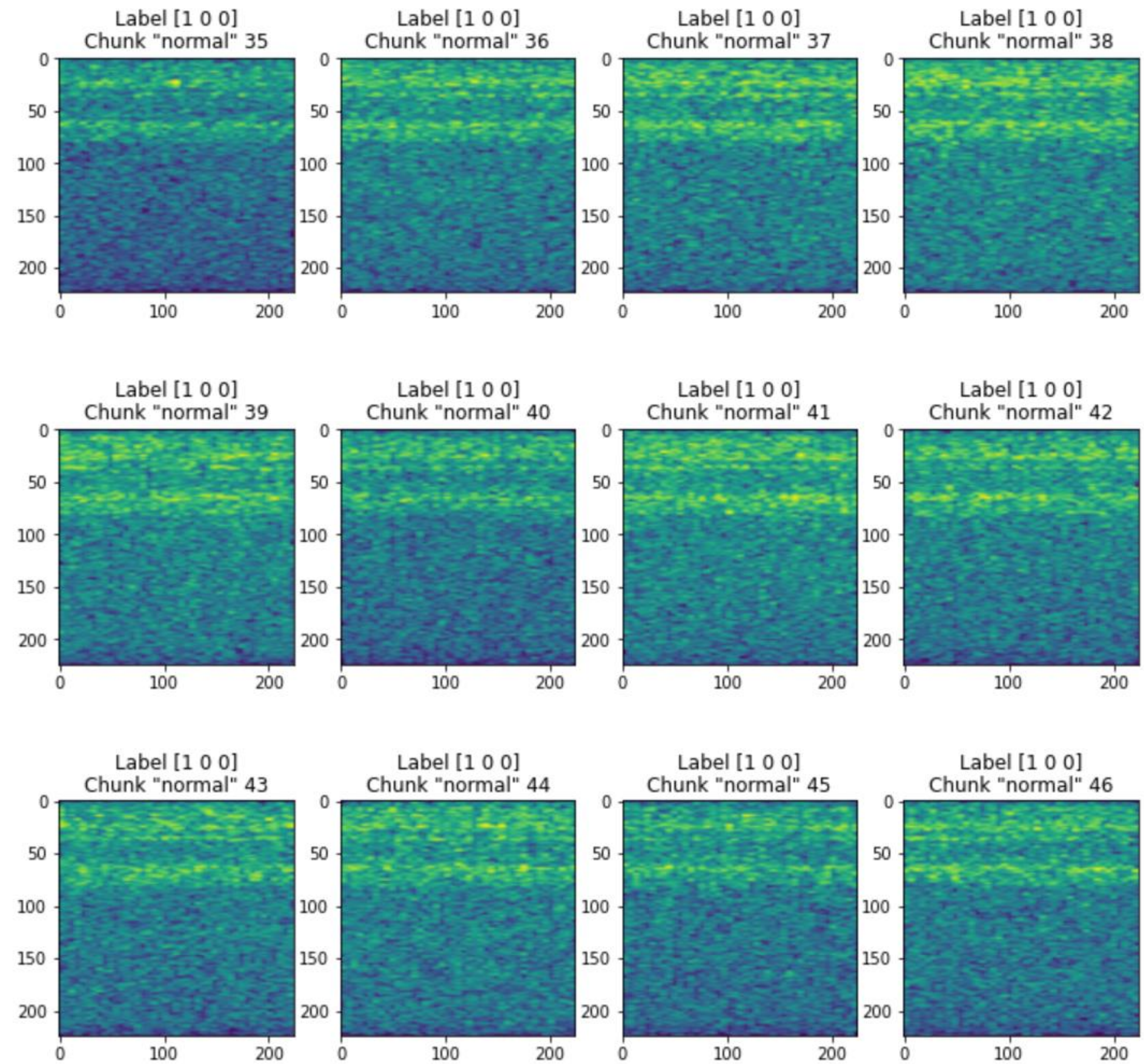

Figure 18 Sample Spectrogram Images for Normal Conditions.

- Short time Fourier spectrogram images of “inner race fault conditions”. Figure 19 represents the one hot encoding label for the images of different chunks.

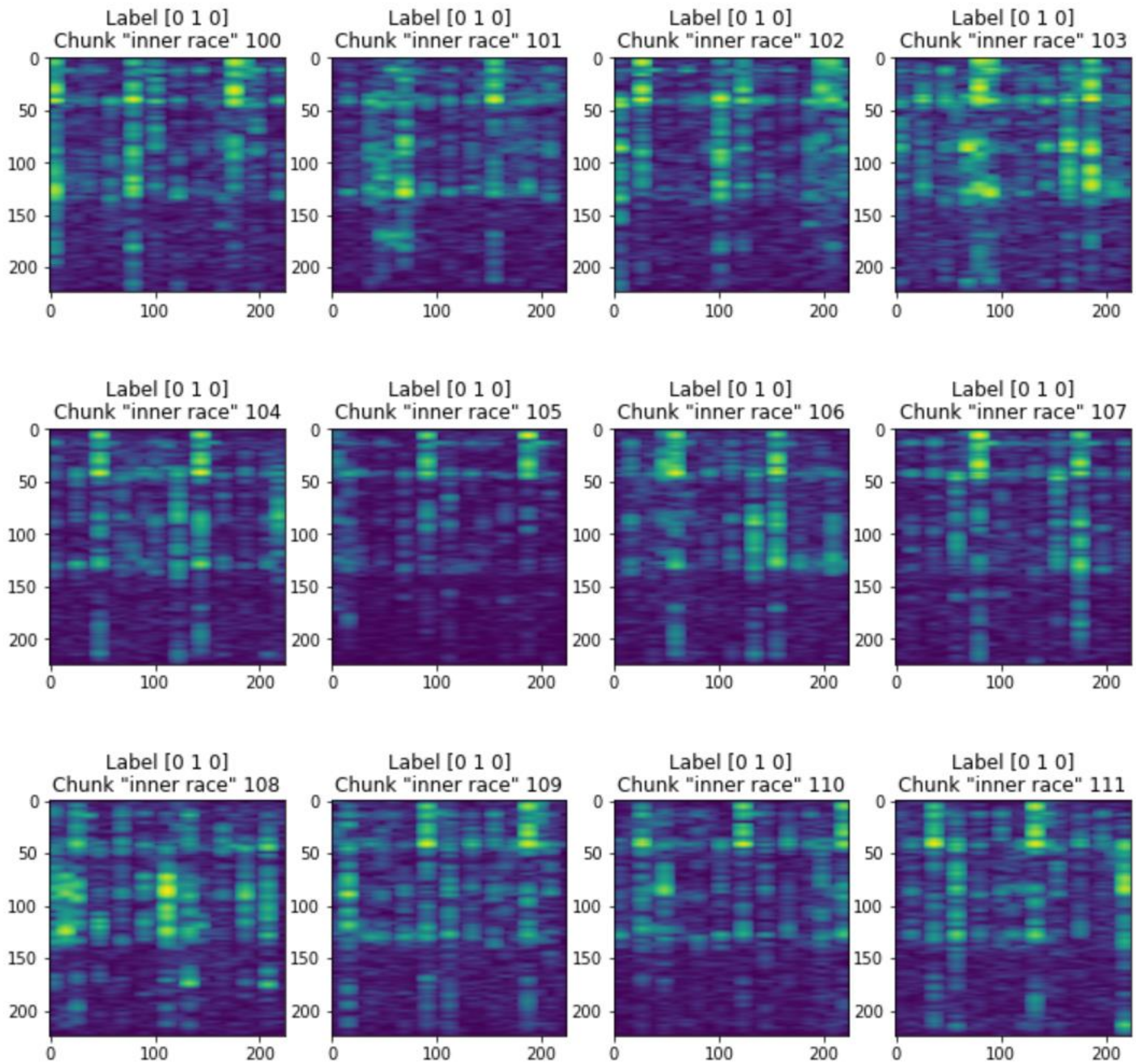

Figure 19 Sample Spectrogram Images for Inner Race Faults.

- Short time Fourier spectrogram images of “outer race fault conditions”. Figure 20 represents the one hot encoding label for the images of different chunks.

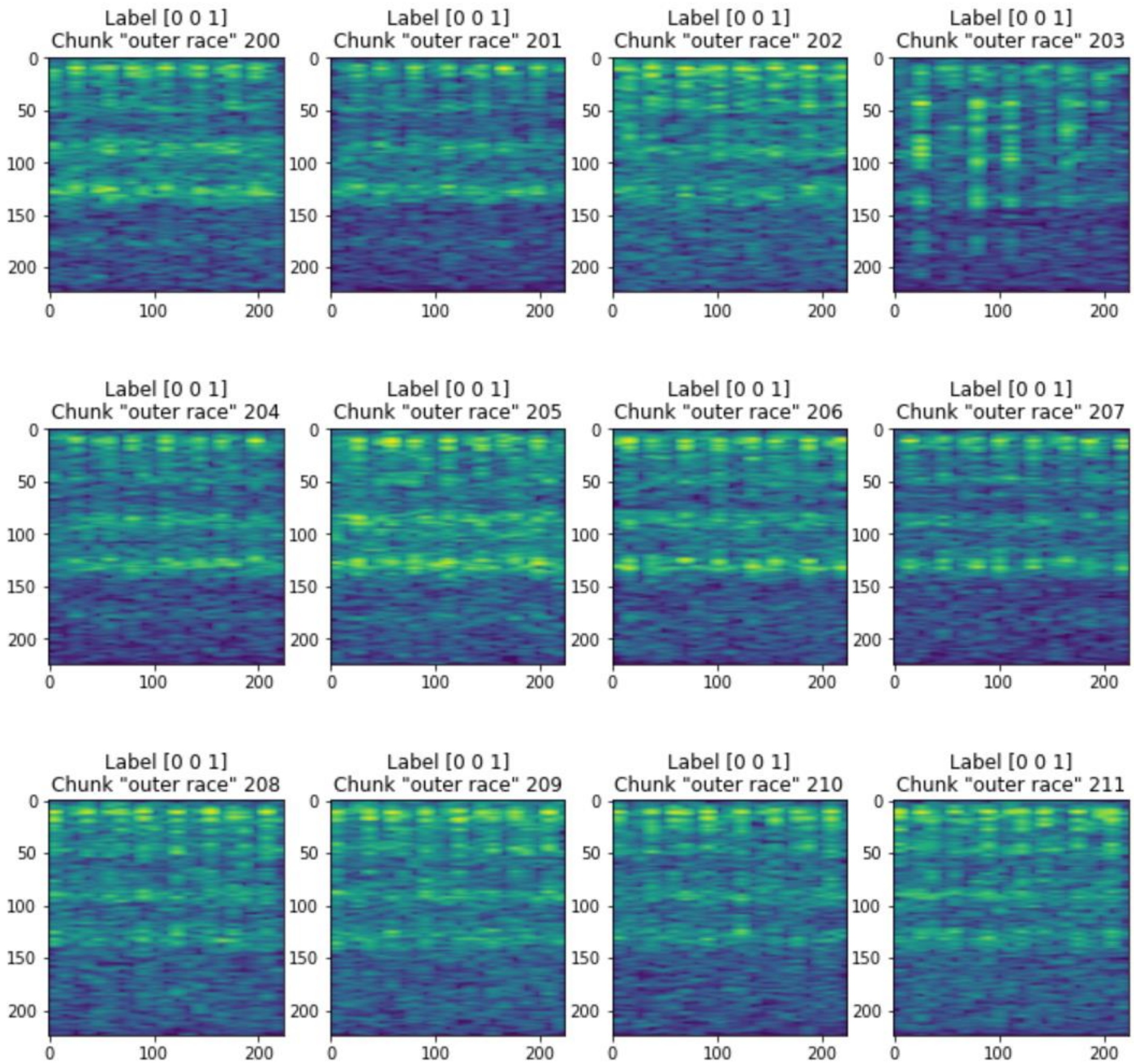

Figure 20 Sample Spectrogram Images for Outer Race Faults.