Chapter 6

SRF Diagnosis using Attention-based Sensor Fusion and Transformer Models

6.1 Introduction

Machinery fault diagnosis using vibration analysis with the sequential learning models is on the rise in FDPM. The attention mechanism has been introduced mainly in sequence to sequence models to concentrate on certain parts of the sequence selectively. The attention scheme which quantifies the interdependence between the input and output elements is called the general attention, and that within the input elements is called the self-attention. The AM has been successfully implemented in natural language processing (NLP) applications and became the driving force of the recent breakthrough sequential model called transformer [228], [229]. It can identify both long-term and short-term dependencies of the sensor signals in complex systems like rotating machines. Despite the success of AM in diverse domains, their application in the fault diagnosis field is limited. The existing literature of complex RM systems with multiple sensors poses specific challenges in applying AM and transformer networks. Typically, vibration data presented by the sensors is a continuous long raw data sequence that is not acceptable for most sequential learning models.

Similarly, a multi-sensor system has more than one such sequence, demanding different levels of consideration for each fault. For example, the sensor placed in the axial direction gives more sensible data for angular misalignment than the sensor placed in the radial direction. In contrast, the sensor placed in the radial direction gives more sensible data in parallel misalignment. Hence, different sensor signals must be treated separately by assigning sufficient weightage by considering its fault discriminative capability. Similarly, it is also observed that the data values at different positions in the signal contribute differently in decision making. For example, the unbalance is a fault that shows cyclic nature, and therefore, the signal data points at certain rotational intervals are more sensitive than the others [10]. Hence, we can conclude that relatively weighted sensor fusion and temporal dependency considerations are the key factors in SRF diagnosis. Moreover, existing AM-related models, including transformers, prefer a content-rich short input representation of data called embedded representation rather than the raw vibration sequence. It is evident from the literature that compared to the commonly used feature extraction with ML or automatic feature extraction DL, the use of symptomatic fault features with sequential learning models brings about greater accuracy in SRF classification [10], [183]. Considering these facts, the following observations are made regarding the inadaptability issues of AM models in SRF diagnosis: i) lack of proper fault-specific embedding representation for a long sequence of vibration data, ii) inability to provide adaptive weightage to sensor segments based on its fault sensitivity in the sensor fusion, and iii) failure in incorporating symptomatic fault features in fault decision-making.

Meanwhile, the popular sequential models like simple RNN, LSTM, or GRU have also not been well explored in the literature. A few attempts have been made to capture the long-term dependencies of the signals to make a decision in SRF diagnosis. Transformers are the recently developed architecture that employs an attention mechanism to find the dependencies of input data. The recurring patterns with long-term dependencies are ascertained by transformers, which are exceedingly used in NLP problems. Transformers are able to access any part of the past data, unlike the RNN models. Simultaneously, the architecture is more suitable for a parallel processing environment than the other sequential models. Surprisingly, there have been no works reported on RFD with transformer or even with simple AM as per our knowledge. But, as the transformers are primarily designed to deal with NLP problems, it faces some practical difficulties when dealing with other TS data, especially the raw vibration signals. i) Compared to the word count of a typical sentence used for NLP classification or prediction tasks, the fault diagnosis sequence are more lengthy. ii) Rather than finding the individual attention to the data points, it requires finding attention between complete informative segments. iii) It is challenging to find proper embedding (informative vector) for the segments since the nature and properties of faults are subjective. Thus it is necessary to address the challenges in providing input to the transformers from a multi-sensor environment which additionally requires splitting the signal into a fixed number of segments without fragmenting the sensitive information.

It is observed from the literature of AM with multi-sensor data that attention is used primarily for selecting features from a set of features that are chosen using CNNs or other FD operations without considering the symptomatic fault features. Most of them deal with single sensor data assumption and do not attempt to give relative weightage to the sensor segments based on the fault sensitivity of each sensor. The literature lacks a proper embedded feature representation to the attention modules, including sufficient fault-specific information from multi-sensor vibration data. Hence, we propose a framework that generates a more domain-specific embedding representation to the AM networks, including transformers, to adapt to the fault diagnosis domain. This incorporates the DFC, that can act as the symptomatic fault parameter for fault decision-making. Along with the TD features, DFC enhance the fault information content of each embedding representation. We have addressed the multi-sensor fusion with fault pattern-based ranking in order to ensure the relative importance of fused sensor vectors and their fault sensitivity. The attention score among the fused sensor segments is assisted with fault pattern-based ranking in the sensor fusion to generate embedding representation. The combined feature representation generated by bagging these two categories of features endorses the most discriminative capability within fewer dimensions. We have used the encoder part of a typical NLP transformer with multi-head attention for final classification, providing the proposed embedding representations as tokens. With this reduced dimension-embedding, the transformers capture the different aspects of dependencies even from short-length sequences, thereby lessening the execution time. Three transformer models were designed, where the first one followed the parallel design philosophy of transformers with positional encoding and multi-head attention layers. The two other models used the recurrent design of transformers by replacing the positional encoding with RNN layers to address the local structure among the embeddings.

6.2 Theoretical Background

6.2.1 Transformers for TS classification

A typical transformer consists of an encoder and decoder parts for most of the NLP tasks. But in order to classify the faulty pattern instead of producing a new sequence, only the encoder part of the transformer is required. The term transformer block defines the multi-head attention followed by a position-wise feed-forward layer with residual connections and layer normalization in this regard. The TS adoption of the typical encoder operations of a transformer is described as follows:

The embedded vector representation is first fed into a positional encoding layer

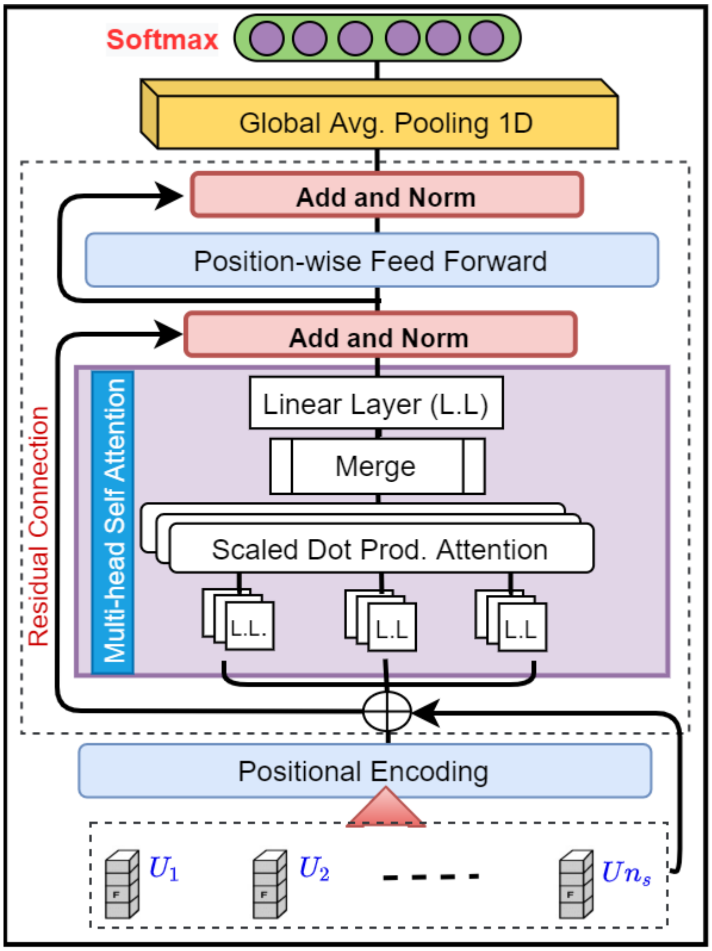


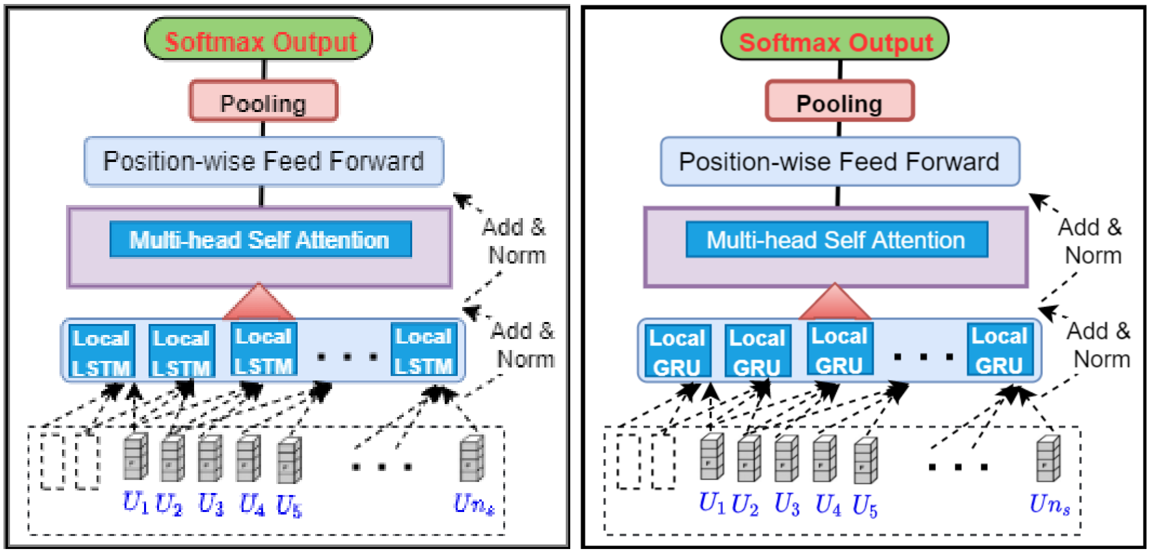
Figure 6.1: General transformer model ($\mathcal{M}_1$)

which adds the relative positional information to the segments. The positional encoding assists in estimating the long-term dependency between the segments. Let $U \in \mathbb{R}^{n_s \times d}$ represent the embedding representation with d being the dimension of an embedding for a sample sequence of n_s segments length. This particular layer creates positional encoding $P \in \mathbb{R}^{n_s \times d}$ and outputs $P + U$, where P can be defined as $P_{i,2j} = sin(i/10000^{2j/d}), P_{i,2j+1} = cos(i/10000^{2j/d})$ for $i = 0, \ldots, n_s - 1$ and $j = 0, \ldots, \lfloor (d-1)/2 \rfloor$. The input embedding and position encoding layers create a matrix of dimension $n_s \times d \times N$, where N is the number of samples. This is fed into the query, key, and value of the transformer that primarily consists of a multi-head attention layer. This layer enables the transformer to encode multiple relationships among the embeddings. Inside this module, three linear layers are provided to transform the query (Q_u), key (K_u), and value (V_u) matrices of each sample using transformation matrices W_q, W_k and W_v (with size $d \times d$), respectively. Both the data matrices and transformation matrices are logically split into separate sections for each attention head where the size of one section (s_s) is given by: $s\_s = d/\#heads$. Now, each head shares the same linear layer but works on its own logical part of the matrices so that the computations of all the heads are attained by a single matrix operation, maintaining reduced model complexity.

Now the model is provided with the transformed query, key, and value matrices (Q_U, K_U, V_U) that are split across multiple heads. These are further used to compute the attention score by the scaled dot product attention. One head takes an input of dimension $n_s \times s\_s$, and every head repeats the identical operation performed for all the samples in a batch. The dot product of the query with all keys will provide the weights that have to be assigned to each value. The product of Q_U and K_U is rescaled by a factor $\sqrt{s\_s}$ to keep the weights in a range since the average size of the dot product grows with the increased dimensionality of the input. The softmax function regularizes each row of the rescaled product, and finally, the weighted sum of 'values' gives the output of each head. This operation is given as follows:

$$Att(Q_U, K_U, V_U) = softmax\left(\frac{Q_U K_U^T}{\sqrt{s\_s}}\right) V_U \tag{6.1}$$

Here these separate attention outputs for each head are combined by the merge operation and passed through one linear layer to get the output of this module. Multiple heads learn different aspects of the segments so that the transformer captures richer interpretations from the sequence. The multi-head attention layer is followed by a position-wise feed-forward layer consisting of two fully connected layers which are applied to each position in the sequence, separately and identically. They use different parameters from layer to layer while the linear transformations are the same across different positions with the ReLU activation function. The output of this layer



(a) LSTM model ($\mathcal{M}_2$)

(b) GRU model ($\mathcal{M}_3$)

Figure 6.2: Recurrent transformer models

is represented

$$FFN(U) = max(0, UW_1 + b_1)W_2 + b_2 \tag{6.2}$$

where W_1 and W_2 are the weights, and b_1 and b_2 are the biases of the layers. A block containing layer normalization and the residual connection are placed around the multi-head attention and position-wise feed-forward layers. The residual connection allows the gradients to flow through a network, directly skipping the non-linear activation function, thereby controlling the vanishing/exploding gradient problem. Similarly, normalization helps with the internal covariate shift problem and ensures that the scale of features is not too different. In layer normalization, the calculation is performed across each feature instead of computing statistics across batch dimensions. As the transformer block is a sequence-to-sequence layer, we have applied a global average pooling to average the output sequence in order to produce a single vector. This is further projected down to a vector with one element per class with output probabilities performed by the softmax operation.

6.2.2 Recurrent transformers

The recurrent transformer models [230] capture the local structural dependency among the embeddings as well as the global long-term dependencies without using the position embeddings. The position embedding in a typical transformer is replaced by a local LSTM or GRU layer to create the local RNN (LRNN) design. This captures the local dependencies and produces a latent representation by operating on a local window. Fig. 6.2 shows the implementations of these models. Such a design has its practical significance because the multi-head attention mechanism solely depends on position embedding to capture the sequential property of data, but it is not effective [231] and requires considerable design effort. Here, LRNN considers local short sequences of length M, and processes them sequentially to produce the M hidden states. The state corresponding to position t is represented by: $h_t = LRNN(U_{t-M-1}, U_{t-M-2}, ..., U_t)$. In this way, by padding $M-1$ positions at the beginning of the sequence, the hidden representations for the whole sequence are generated as: $h_1, h_2, ..., h_{n_s} = LRNN(U_1, U_2, ..., U_{n_s})$. The upper layers of both these models follow the same structure of a general transformer shown in the Fig. 6.1.

6.3 Proposed Method

The most crucial task in making AM and transformers adaptable to the RM fault diagnosis domain is generating embedding representation from multi-sensor data. It requires proper segmentation of raw data, extraction of most sensitive features, and multi-sensor segment fusion. As we deal with the vibration data from varying speed and load industrial conditions, it is necessary to extract the DFC sensibly from the inconsistent raw data. The first phase of the proposed framework, called the embedding representation generation phase, deals with these issues and provides the tokens for classification. The transformer-based classification phase addresses both long-term and local dependencies among these embedded representations with the generic transformer as well as the recurrent transformer models.

6.3.1 Generation of embedded representation

In NLP tasks, a word is the primary information unit and is represented as an informative vector representation called word embedding before applying it to the AM or transformer models. But, in the fault diagnosis frameworks, one sensor segment forms a unit of information where there are multiple sensor segments with different sensitivity to each fault. Hence a compact embedded representation is generated in the proposed framework by means of two modules; namely, i) combined-feature representation module and ii) sensor fusion module. This phase corresponds to the word embedding layer of a typical NLP task.

6.3.1.1 Combined-feature representation module

As we have explained in section 3.4, the K number of sensors mounted at different positions, produce signals $V = [V_1, V_2, ..., V_K]$ with each $V_i \in \mathbb{R}^L$ where L is the length of one sample sequence. Provided the sampling points per rotation (S_r), the segment length (S_l), and the overlapping points in segmentation (S_o), then according to the condition $S_o \leq S_r \leq S_l$, the framework produces n_s number of segments, such that $n_s = (L - S_l)/(S_l - S_o) + 1$. Thus segments from sensor S_i are created as: $V_i = [V_{i,1}, V_{i,2}, ..., V_{i,n_s}]$.

The DFC extraction follow the same process explained in Section 3.4. Both amplitude and phase values are extracted to generate the ten DFC components of harmonic frequencies 1x to 5x. Similarly, the TD representatives of each segment, such as mean, standard deviation, variance, root mean square, absolute maximum, and kurtosis [232], are generated from the data bins of length S_r. Combining both will generate a 16 dimension feature vector for representing a particular sensor segment. Thus, the t^{th}

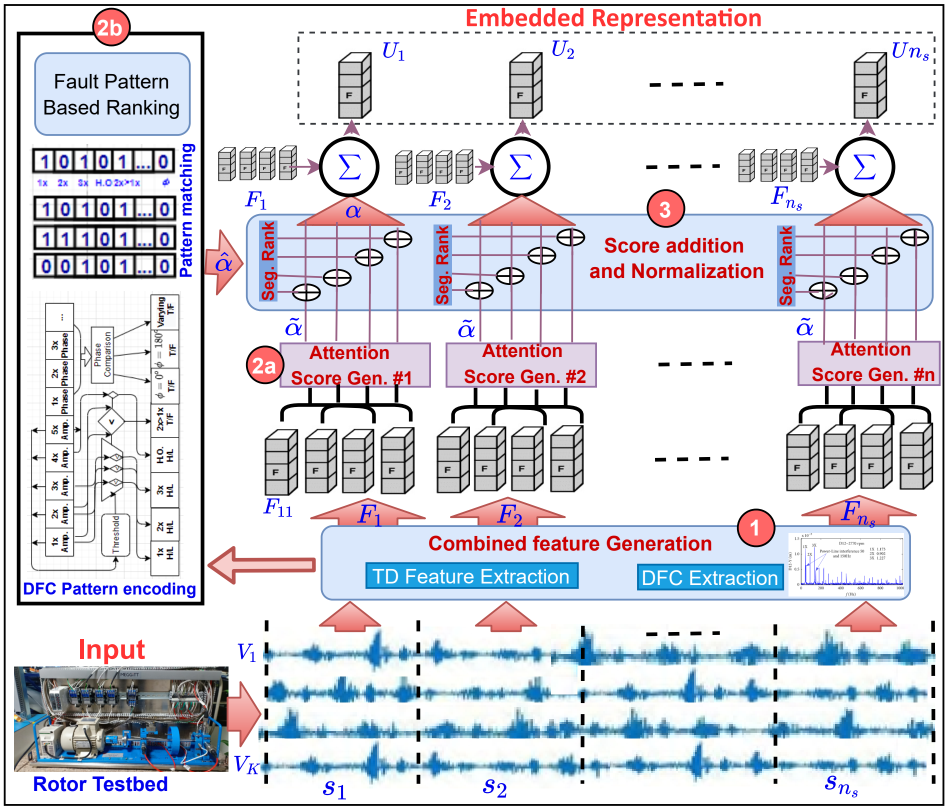


Figure 6.3: Generating embedded representation

segment of sensor i with length S_l can now be represented as $F_{it} \in \mathbb{R}^d$, where d is the dimension of combined-feature vector (16 in our case) such that $d \ll S_l$. This process is shown in Fig. 6.3 with number 1 given within a red circle.

6.3.1.2 Sensor fusion module

The SRF is sensitive to rotational frequency components depending on the mount position of the sensors as well as the different sensor modalities. The most sensitive mount positions of sensors for SRF are given in the Table 6.1 as the affected plane (A.P). In our experiments, accelerometers and proximity sensors are used in a contact or contactless manner to measure casing vibrations and shaft vibration. Hence, it is necessary to give relative weightage for the combined feature vector of each sensor to generate a single embedded vector for a segment. We use two-way scoring to ensure relative importance in segment-wise sensor fusion. The first score is based on attention

Table 6.1: SRF and DFC correlation and decoding

SRF	**DFC symptomatic information**			**Amplitude Decoding**					**Phase Decoding**		
	A.P	Symptomatic Frequency	Phase	1x	2x	3x	>3x	2x>1x	0°	180°	Var.
S_UB	R	Higher 1x with less than 15.0% amplitude harmonics	A 0° P.S. in Radial direction.	1	0	0	0	0	1(R)	0	0
C_UB	R	Higher 1x with less than 15.0% amplitude harmonics	A P.S. of 180° in Radial direction.	1	0	0	0	0	0	1(A)	0
D_UB	R	Higher 1x with less than 15.0% amplitude harmonics	A P.S. of 0° to 180°.	1	0	0	0	0	0	0	1
V_MA	R	2x >1x. Stronger 1x (Radial and torsional responses)	A P.S. of 180° in axial direction.	1	1	0	0	1	0	1(R)	0
H_MA	R	2x >1x. Stronger 3x in severe cases.	180° P.S. in the axial direction.	1	1	1	0	1	0	1(A)	0
LS	A,R	Higher 2x. Harmonics of 1x to 10x & multiples of its subharmonics.	Unstable reading.	1	1	0	1	0	0	0	1
HL/NM	N.A			0	0	0	0	0	1	0	0

A.P: Affected Plane, P.S.: Phase Shift, R: Radial, A: Axial, Var.: Varying

value among sensor segments, and the second score uses SRF fault similarity pattern matching.

Attention score generation: The attention mechanism helps to provide selective concentration to certain parts of the input and ignores the irrelevant parts. We utilize this property in emphasizing those sensor segments which has got the most sensitive patterns of SRF. The basic attention mechanism [228], follows three sets of elements, namely the 'query,' 'key,' and 'value.' In the proposed sensor feature fusion phase, these three elements are decided as follows:

Query: The input vector corresponding to the current output that matches against every other input vector is called a query. Our model uses a sensor-level context vector as the query vector, which is denoted by q_1.

Key: The vector that the query is matched against is called the key vector. We use the hidden representation generated from a single-layer neural network to achieve this. It is represented as $\lambda_{it} = g(WF_{it} + b)$, where $g(\cdot)$ is the activation function ($tanh$ in our

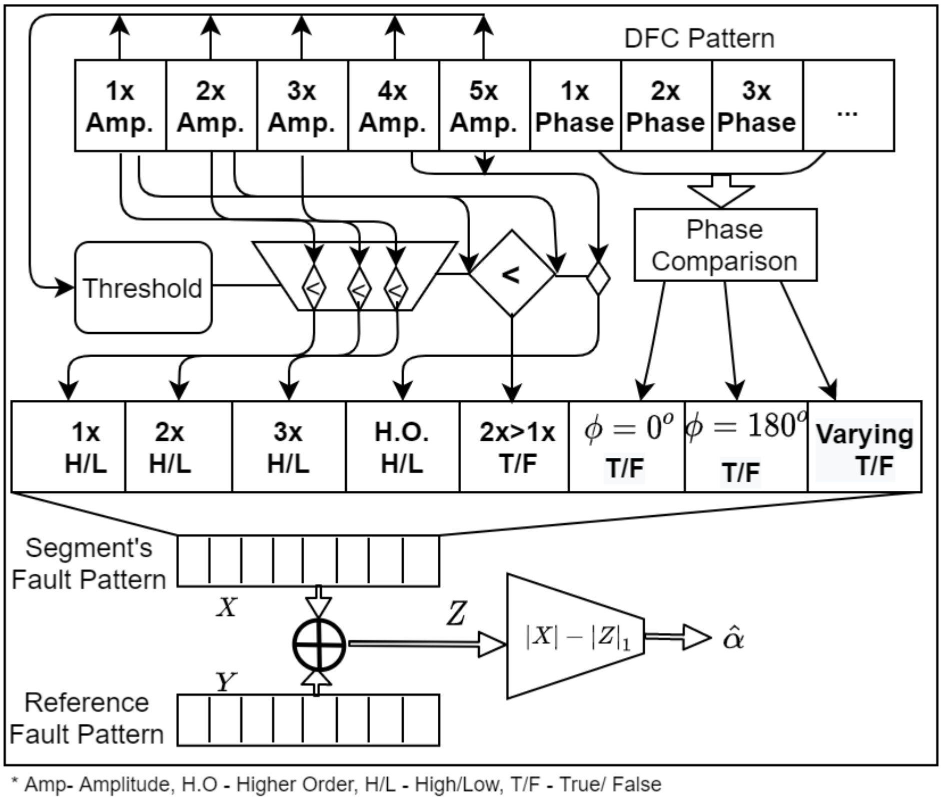


Figure 6.4: Generating fault pattern matching score

case) and W and b are the weights and bias of the neural network, respectively.

Value: The vectors used to provide the output of attention by weighted sum is called the value vector, which is F_{it} itself in our case. The attention weights that give relative similarity between the sensor segments are calculated by finding the dot product of the query and the key. This operation is given as :

$$\tilde{\alpha}_{it} = \frac{e^{(\lambda_{it})^T q_1}}{\sum_i e^{(\lambda_{it})^T q_1}} \tag{6.3}$$

The exponential operation always sets out the positive similarity, and its monotonically increasing property ensures the ordering of weights. The normalization process keeps the value between zero and one, which sums up to one.

Fault pattern matching score generation: This is the process of estimating the explicit similarity of DFC patterns of segments with the predefined fault patterns. For this purpose, the DFC of subsampled feature vectors is decoded to a binary pattern representing the most critical fault characteristics of SRF. Table 6.1 shows the amplitude and phase characteristics of the symptomatic frequency and the encoded fault pattern of SRF. The fault encoding process is summarized in Fig. 6.4. An 8-bit fault template is generated from each combined-feature vector corresponding to the sensor segments. The first three bits of the template signify the status of lower band rotational frequencies, i.e., 1x to 3x, as SRF is more sensitive in these frequency bands. A high or low value of 1x to 3x amplitudes is denoted by a one or zero value in the bit positions. The comparative status between 1x and 2x is given at bit position 5, where value one indicates that 2x is greater than 1x. The remaining higher-order harmonics are decoded to bit 4. The phase values are represented in the last three-bit positions. Bit 6 represents the phase shift of 0° and bit 7 indicates 180° phase shift. The final bit position holds a value one when the phase varies between 0° and 180°. This encoding scheme is subject to slight changes depending on the sensor mounting position (radial, axial or tangential). The fault pattern of each sensor segment (X) is compared with the reference pattern of faults (Y) shown in Table 6.1 for assigning the fault similarity score ($\hat{\alpha}$). The matching score is calculated by taking the hamming distance (Z) and subtracting it from the length of the pattern (refer Fig. 6.4). The reference fault pattern for similarity score calculation is set as the segment's closest fault pattern. Subsequently, the softmax function is used to get the normalized score for every t^{th} segment for every i^{th} sensor. Once the two-level scores are evaluated, their sum is normalized to decide the final score, i.e., $\alpha = \mathcal{N}(\tilde{\alpha} + \hat{\alpha})$. Here, $\mathcal{N}$ is the softmax normalization operation that sets each score between zero and one, such that the scores sum up to one. Finally, the uniform representation U_t that aggregates all sensor segments is computed by:

$$U_t = \sum_{i=1}^{K} \alpha_{it} F_{it} \tag{6.4}$$

The weighted sum of the final score (α) and the sensor feature vectors produce the embedding representation U_t for a segment for further transformer processing. The sub-functions of sensor fusion module such as attention score generation, fault-pattern matching score generation, and final embedded vector generation are shown in Fig. 6.3 with numbers $2a$, $2b$ and 3, respectively.

6.3.2 Transformer based classification

Since our objective is to classify the faulty pattern instead of producing a new sequence, we adopt only the encoder part of the transformer. We have used three transformer architectures, including one general transformer model ($\mathcal{M}_1$) and two models based on recurrent transformer design ($\mathcal{M}_2$, $\mathcal{M}_3$), which enjoy the advantages of both multi-head attention mechanism and RNNs.

The general transformer model contains the operational modules of the encoder part of a typical transformer as referred in Fig. 6.1 and explained in section 6.2.1. The embedded vector representation is first fed into a positional encoding layer which adds the relative positional information to the segments. In the proposed segmentation scheme, each segment is affirmed to contain one or more complete rotational information. The padding mask function is skipped in the proposed model since all the embedded vectors are of the same dimension. The transformer block primarily consists of a multi-head attention layer that mainly performs three operations. First, it linearly transforms the query, key, and value vectors and splits them among the number of heads. Since the proposed embedding contains both TD and DFC features, we have used two heads to learn from these embedding dimensions, which correspond to the time and frequency domains. Then it performs the scaled dot product attention. The the results from all the heads are merged and passed through another linear layer. The feed-forward network has two fully connected layers, with ReLU as the activation function. The residual connection and layer normalization are maintained in the architecture. As the transformer block is a sequence-to-sequence layer, we apply a global average pooling to average the output sequence to produce a single vector. This vector is further processed by an FC layer before being projected down to a vector with one element per class and then a softmax operation to produce the probabilities.

In the recurrent transformer models the position embedding is replaced by a local LSTM or GRU layers to captures the local dependencies. They operate on a local window of size three as shown in Fig. 6.2. The upper layers of both these models follow the same structure of $\mathcal{M}_1$ as shown in the diagrams.

The hyperparameter settings of the classification models are as follows: The transformer models were implemented on Python, and the same set of hyperparameters was used for all three models. A decent trade-off between accuracy and model complexity was observed for the recurrent layers and feed-forward layers of the models with 64 hidden nodes (HN) of two layers (L) and a learning rate (η) of 0.001. The initial hyperparameter search was performed over $L \in \{1, 2, 3\}$, $HN \in \{16, 32, 64, 128\}$ and $\eta \in \{0.1, 0.01, 0.001\}$ for 300 epoch. The experiments found that the transformer model with one multi-head attention layer having two heads provided decent performance with a batch size of 84 within the first 50 epochs. The dimension for embedding representation is given as 16, and the sequence length of a sample was selected as 120. The models $\mathcal{M}_2$ and $\mathcal{M}_3$ require an additional parameter called 'local window size,' which is set to three for all the experiments.

6.4 Results and Discussions

In this section, experimental results are presented and discussed. The different sensor data are segmented into 512 data points satisfying the conditions discussed in section 3.4. The same subsampling and feature extraction operations are performed on both datasets considering the closest matching speeds under all load conditions. We selected 70.0% of data consisting of all categories for training and the rest for validation/testing.

Table 6.2: Performance of embedded representation

		Simple sensor fusion		Proposed sensor fusion	
		DS-1 (%)	DS-2 (%)	DS-1 (%)	DS-2 (%)
$\mathcal{M}_1$	FD	94.57	94.84	95.36	95.45
	DFC	95.79	96.01	97.15	97.61
	TD+FD	96.11	96.19	96.96	97.73
	TD+DFC	97.05	97.54	**98.80**	**98.30**
$\mathcal{M}_2$	FD	95.01	95.34	96.22	96.04
	DFC	96.13	96.74	97.87	98.08
	TD+FD	96.92	97.19	98.23	98.30
	TD+DFC	97.24	97.96	**99.62**	**99.11**
$\mathcal{M}_3$	FD	94.79	95.26	96.40	96.33
	DFC	96.62	97.14	98.13	97.84
	TD+FD	97.01	97.05	98.40	98.21
	TD+DFC	97.76	97.75	**99.77**	**99.25**

6.4.1 Ablation study on embedded representation module

The embedded representation module is one of the key contributing modules in the framework. The compact embedded representation is generated in the proposed model using this module by means of two submodules; namely, i) combined-feature representation module and ii) sensor fusion module. The ablation study on the combined-feature representation module is conducted by creating a 16 dimension feature space with FD and DFC independently and their combinations with TD features. The TD features are clubbed since they represent the sequential component and provide direct dependency information among the segments. The independent FD and DFC features show the impact created in the absence of a combined-feature representation module. Similarly, the significance of the sensor fusion module is substantiated by replacing it with simple attention sensor fusion. In the same way, it shows the importance of the fault sensitivity score generation module in sensor fusion of SRF. Thus this section shows the overall ablation study in the embedded representation module in three ways.

It is observed from Table 6.2 that both the datasets demonstrate decent performance with the embedded vectors with the proposed fusion scheme as well as the simple fusion scheme where DFC is involved in decision-making. In feature-wise analysis, the

DFC alone performed better than FD features in both datasets and gave an accuracy close to that provided by TD + FD features. It shows that DFC is most suitable for the embedded representation in the frequency transform domain, which can still be improved by adding TD features, as shown in the results. Moreover, increasing the dimension of frequency transformed features (FD or DFC) has minimal impact on the model performance in terms of accuracy. But interestingly, adding TD features along with it enhances the performance of the three transformer models, especially the recurrent models. The average error rate reduction using the simple sensor fusion method is around 33.0% for both datasets with combined features. But, for the recurrent model, it is approximately 36.0% and 34.0%, respectively, for DS-1 and DS-2. Similarly, the average error rate reduction in the proposed sensor fusion is around 62.0% and 51.0%, respectively, for DS-1 and DS-2 with combined features. Besides, an error rate reduction of roughly 70.0% for DS-1 and 57.0% for DS-2 is noted for the recurrent models. Also, simple attention fusion gives higher accuracy for DS-2 compared to DS-1, while with the proposed fusion, this trend is not visible. Among the three models considered, the recurrent transformer models, particularly $\mathcal{M}_3$ shows better performance with the four types of feature set investigated. The feature set evaluation reveals that the most informative embedded vector representation can be generated from the TD features combined with DFC, and it has been employed as the embedded representation for further experiments.

6.4.2 Individual sensor performance

The ablation study replacing the sensor fusion module with individual sensors highlights certain reflections. The individual sensor performance given in Fig. 6.5 shows the sensor-wise accuracy, which is an essential factor in determining the overall effectiveness of the framework. We take the first four sensors from both datasets for analysing the extent to the multi-head attention deal with proposed embedding vectors on single

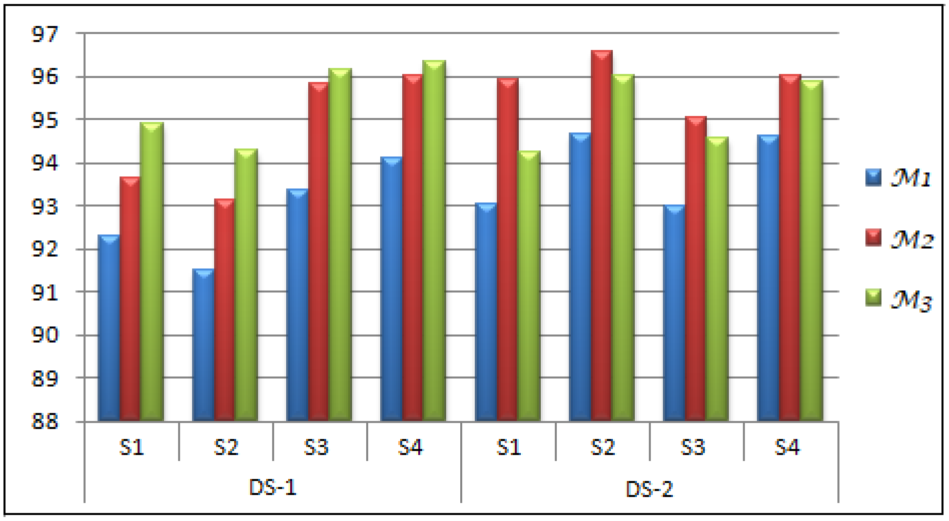
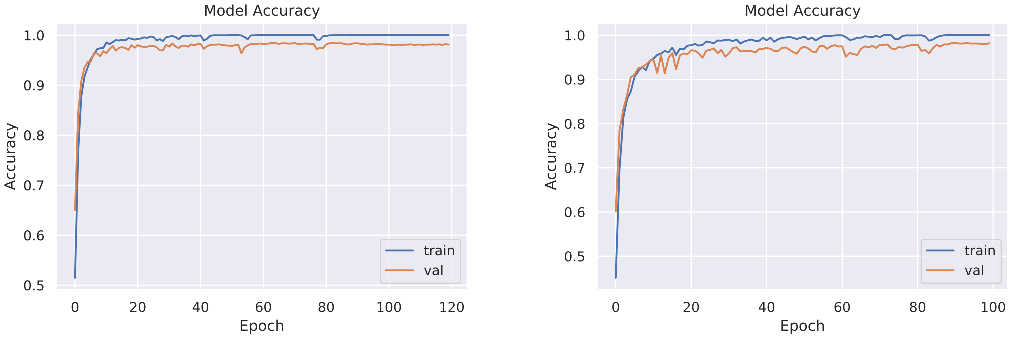


Figure 6.5: Individual sensor performance (Accuracy %)

sensor data without sensor fusion. It is observed that in DS-1, the sensors S3 and S4 give higher accuracy compared to S1 and S2 because of the impact of shaft vibration signals compared to the casing vibration signal. Meanwhile, in DS-2, the sensors placed in radial position (S2, S4) give better accuracy than those mounted in axial position (S1, S3). Thus, it is evident that the nature of sensors and their relative positioning in the acquisition system varies the accuracy to some extent (1.0% to 2.0% in our example), which signifies the importance of providing relative weighting in sensor fusion. Besides, the recurrent transformer models bring about an enhanced accuracy of around 2.0% to 3.0% compared to the basic transformer model. But it is less than 2.0% in the fused sensor data, which indicates that the impact of sequential information is more evident in a single sensor signal compared to the fused sensor signal. Comparing the overall performance, DS-2 gives slightly better results than DS-1 due to the smooth nature of sensor data. Though the recurrent transformer models are performing almost the same way, $\mathcal{M}_3$ turned out to be the best performing model with DS-1, while $\mathcal{M}_2$ gives the best results with DS-2. Besides, the individual sensor performance ranges from 91.0% to 96.0% for all the models on both datasets, and simple sensor fusion delivers performance in the span of 95.0% to 98.0%. Also, the proposed sensor fusion approach assists in achieving accuracy in the range of 97.0% to 100.0%, reducing the average error by more than 50.0%.

Table 6.3: Training performance

	$\mathcal{M}_1$ (Accuracy %)		$\mathcal{M}_2$ (Accuracy %)		$\mathcal{M}_3$ (Accuracy %)	
	Train.	Val.	Train.	Val.	Train.	Val.
DS-1	99.91	98.81	100.0	99.6	100.0	99.8
DS-2	99.83	98.29	100.0	99.15	100.0	99.3



(a) DS-1 ($\mathcal{M}_1$)

(b) DS-2 ($\mathcal{M}_1$)

Figure 6.6: Training performance (Accuracy Vs Epochs)

6.4.3 Transformer model performance

The training performance of the proposed framework is given in Table 6.3, and the accuracy curves of model $\mathcal{M}_1$, on both datasets are shown in Fig. 6.6. It is observed that both datasets demonstrate decent performance with the proposed embedded representation in terms of accuracy. It is interpreted as the efficacy of the embedded representation and the proposed transformer models. The models $\mathcal{M}_2$ and $\mathcal{M}_3$ provide higher training and validation accuracies compared to the model $\mathcal{M}_1$, but with increased training time due to recurrent structure incorporated with the transformer. It is observed that the parallelizable architecture of $\mathcal{M}_1$ reduced the training time around 33%-50% compared to the other two models. Among the two recurrent transformer models, $\mathcal{M}_3$ is offering the best validation accuracy for both datasets. The increased training accuracy and a close matching validation accuracy specify that there is no overfitting in the training process with both datasets. Besides, the accuracy curve indicates that the accuracy of DS-1 reaches a maximum value of around 20 epochs, where DS-2 requires around 50 epochs for the same due to the increased diversity of the dataset.

Table 6.4: Class-wise performance of transformers

	DS-1 (Accuracy %)							**DS-2 (Accuracy %)**				
	HL	S_UB	C_UB	D_UB	MA	LS	All	NM	UB	H_MA	V_MA	All
$\mathcal{M}_1$	98.80	100.0	95.60	100.0	99.60	98.80	**98.80**	98.12	99.89	99.01	96.17	**98.30**
$\mathcal{M}_2$	99.64	99.90	98.72	100.0	99.56	99.89	**99.62**	98.88	99.35	99.42	98.79	**99.11**
$\mathcal{M}_3$	100.0	99.87	99.63	100.0	99.67	99.45	**99.77**	98.98	99.63	99.23	99.15	**99.25**

The fault-wise performance of the proposed transformer model using the combined TD-DFC features is given in Table 6.4. The results demonstrate that the models $\mathcal{M}_2$ and $\mathcal{M}_3$ give higher accuracy compared to $\mathcal{M}_1$. However, $\mathcal{M}_3$ is the best-performing model in comparison to the other alternatives. But the rate of improvement in accuracy between the models is not as much as observed in individual sensor accuracies (refer Fig. 6.5). In DS-1, the highest overall accuracy is registered as 99.77% by model $\mathcal{M}_3$. Among the classes, D_UB class gives higher accuracy compared to the other classes. The overall accuracy of DS-2 is comparatively lesser than DS-1, producing 99.25% with the best performing model $\mathcal{M}_3$. The V_MA class and the NM class mainly give 1.0% to 2.0% reduced accuracy compared to the other classes. It is observed that the ability shown by $\mathcal{M}_2$ and $\mathcal{M}_3$ in exploiting the sequential dependency between the embedded representations is the key in accuracy enhancement.

Fig. 6.7 shows comparison of evaluation metrics using precision, recall, and F1-score. Since DS-1 contains more imbalanced class data compared to DS-2, these metrics are calculated for DS-1 alone. The higher values of these metrics signify that the models show stable performance even when the dataset is imbalanced. The graph clearly shows that model $\mathcal{M}_3$ is the best performer, while model $\mathcal{M}_1$ shows comparatively lower performance on C_UB and HL classes.

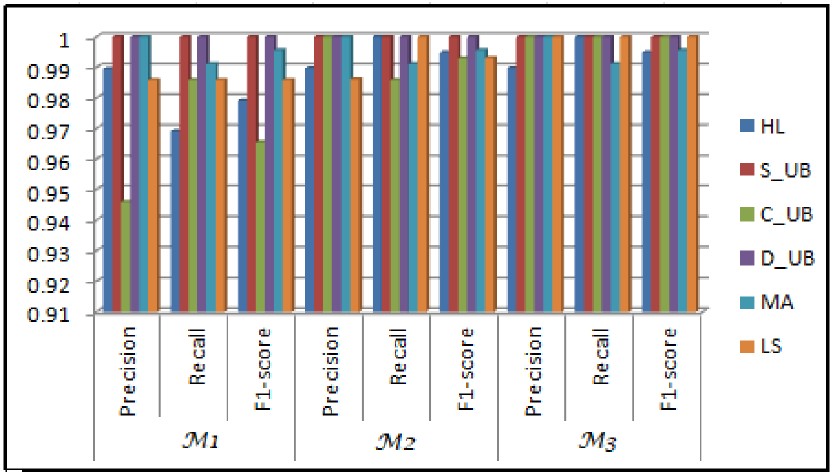
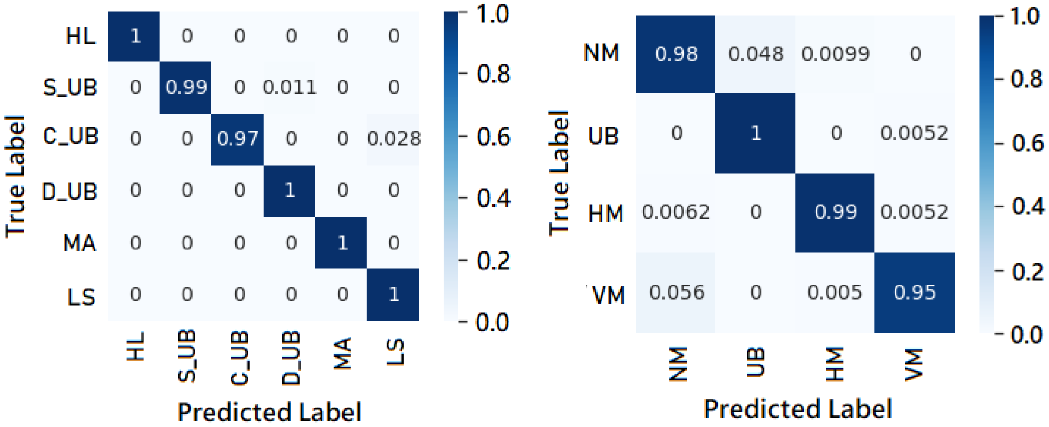


Figure 6.7: Precision, Recall and F1-Score of DS-1



(a) DS-1

(b) DS-2

Figure 6.8: Confusion matrix of $\mathcal{M}_3$

Further, confusion matrices of model $\mathcal{M}_3$ for both the datasets are shown in Fig. 6.8. The proposed transformer architecture demonstrates a decent performance for healthy as well as for faulty conditions except for C_UB in DS-1. A few instances of C_UB are misclassified as LS as both faults are highly correlated, creating adversity in classification. DS-2 suffers from misclassifying V_MA as the NM class, and very few samples of NM class are categorized as V_MA or H_MA. In short, the three proposed variations of the transformer architecture demonstrated exemplary performance with multi-sensor fusion producing accuracy above 99.0% for both datasets.

Table 6.5: Effect of synthesized data

	Synthesized Data		Synthesized+Original	
	DS-1 (%)	DS-2 (%)	DS-1 (%)	DS-2 (%)
$\mathcal{M}_1$	98.23	98.29	98.99	98.54
$\mathcal{M}_2$	98.98	98.53	99.41	99.33
$\mathcal{M}_3$	99.17	99.31	99.80	99.19

6.4.4 Performance with synthesized data

To analyze the performance of the proposed framework with synthesized data of DS-1 and DS-2, we followed the soft-DTW-based augmentation scheme proposed in [183].

Thus, a more diverse form of data is generated and tested with the proposed model to evaluate the generic nature of the framework. The augmentation process generates the consensus sequences of a set of original sequences smoother than the original data. In Table 6.5, the accuracy of synthesized data alone and its performance when added with original data are demonstrated. It is noticed that the synthesized data alone produces almost the same accuracy among the models on both datasets. It is worth noting that the sequential property exploited by models $\mathcal{M}_2$ and $\mathcal{M}_3$ from the original data, is not worked well with synthesized data. The model $\mathcal{M}_2$ is the best performer for dataset DS-2 with original data added to synthesized data, but with synthesized data alone, $\mathcal{M}_3$ performed best with the DS-2 dataset. Meantime, when the synthesized data has been added with original data precipitates minor enhancement in accuracy for both the models. This specifies that even without adding additional diverse training data, the model can capture discriminative features from the available training set.

6.5 Summary

This work presents a framework for SRF diagnosis using the transformer to bridge the gap between fault diagnosis domain and attention-based architectures, using a domain-specific embedding layer. The issues in applying the advanced sequential learning strategies in the SRF domain with raw vibration data are addressed in this chapter. The attention mechanism is effectively used with transformers to identify both long-term and short-term dependencies of the sensor signals in the complex RM system with multiple sensors. This is achieved by proposing an embedding representation that ensured symptomatic fault features by incorporating DFC and endorsed discriminative capability by combining the DFC with TD features. An attention-based multi-sensor fusion is presented in this work. This considered attention weights along with the fault pattern-based ranking and succeeded in ensuring the relative importance of the feature vectors of the fused sensors. Thus a more realistic and domain-specific embedding representation has been proposed for equipping the transformers to overcome the challenges in handling multi-sensor TS data, integrating fault-specific information. The basic transformer and two recurrent transformer models utilize the long-term dependency and the local dependency from the embedded tokens. The framework's running complexity is reduced since it can capture sufficient dependency information even from short-length sequences, with a fewer dimension embedding.

The experiments showed the impact of DFC in the embedded representation by analyzing its performance with traditional TD and FD features. The DFC with TD features turned out to be the best feature combination for embedded representation. The multi-head attention with two heads in the transformer effectively captured the dependency information from these two kinds of features in different aspects. The individual sensor accuracy showed the dominance of certain sensors in specific faults, which established the importance of relatively weighted sensor fusion in SRF diagnosis. It is also observed that the recurrent transformer models performed better than the general transformer model among which, the GRU-based models shown slightly better accuracy. Comparing the class-wise performance, a few instances of C_UB are misclassified as LS in DS-1, and in DS-2, V_MA fault is misclassified as the NM class on a very few occasions. In short, the three proposed variations of the transformer architecture demonstrated excellent performance with multi-sensor fusion. One important observation is that the proposed transformer architectures showed minor accuracy enhancements with the addition of synthesized data, which specifies that the model can capture discriminative features from the available training set without necessitating additional diverse training data. The overall experimental results signify the effectiveness of the proposed framework in utilizing transformers with multi-sensor data, incorporating fault information content, and excelling in various industrial working environments.