Chapter 2

Literature Review

2.1 Introduction

The rise of Industry 4.0 precipitated the development of PHM, which gave way to an industrial system that is more complicated with higher automation and increased precision. Thus, the health state monitoring of such equipment became a challenging responsibility to improve the availability of the system. A significant amount of real-time heterogeneous data generated by various sensors of CPS from several RM needs to be effectively analyzed using AI-based algorithms to perform intelligent digital monitoring in a collaborative and distributed environment of the industry. It supports identifying the root cause of numerous faults, investigates their development, and proactively forecasting the maintenance activities to overcome cataclysmic accidents, unplanned shutdowns, and substantial loss.

In the beginning, the RFD research was split into two categories: i) model-based diagnosis and ii) signal-processing-based diagnosis. In the model-based diagnosis, the features of the monitored system and physical quantities like mass, stiffness, damping-matrices, etc., are simulated using a mathematical model along with the facts of physics [48]. The modeling can be performed at two levels: (i) system-level or (ii) component level. That said, the researchers observed some drawbacks with this method. To begin

with, it has proven to be impracticable as the system turns more complex with more advanced machinery and high-tech technologies. Furthermore, this method has proven inadequate for updating real-time processes with newly read data. More research studies on the model-based identification of rotor faults can be found in the literature review done by Bachschmid et al. [49]. In the meantime, signal processing-based techniques [50] depend on drawing fault-specific characteristic features from the acquired signal and yoking them with the suspected fault's statistical model. Similar to the model-based methods, using methods based on pure signal processing for fault diagnosis also has some restrictions. Notwithstanding, for a specific period, the basis of fault diagnosis and the decision-making algorithms was the characteristic features recognized by the signal processing methods [12, 51–53].

Recently, in the context of Industry 4.0, the motto of 'right information and data at the right time for decision making' emerged as the primary force of AI-based data-driven FDPM [4,54]. Under the AI umbrella, there are two main categorical divisions: shallow learning (SL) – also known popularly as machine learning (ML) and deep learning (DL). The decision-making philosophy driven by data has been enhanced due to low-cost sensors and big data availability. The far-flung utilization of ML and the fast outgrowth of DL algorithms have established their popularity in the field of RM fault diagnosis [55, 56]. What is more, these algorithms are highly regarded among the research community due to their flexible and robust nature [55]. Regrettably, most of the research related to AI-based fault diagnosis of RM focuses solely on the bearing or gear faults [12, 55, 57], while giving little emphasis to rotor faults. Hence in this literature study, we proposed a more clear RM fault classification from the vibrational characterizes perspective to affirm the position of rotor faults. The theoretical background of rotor faults was then provided to give an explicit idea that makes the researchers capable of coming up with new fault-specific enhancements. The research fields of ML- and DL-based RFD are examined in this section with other advanced methods.

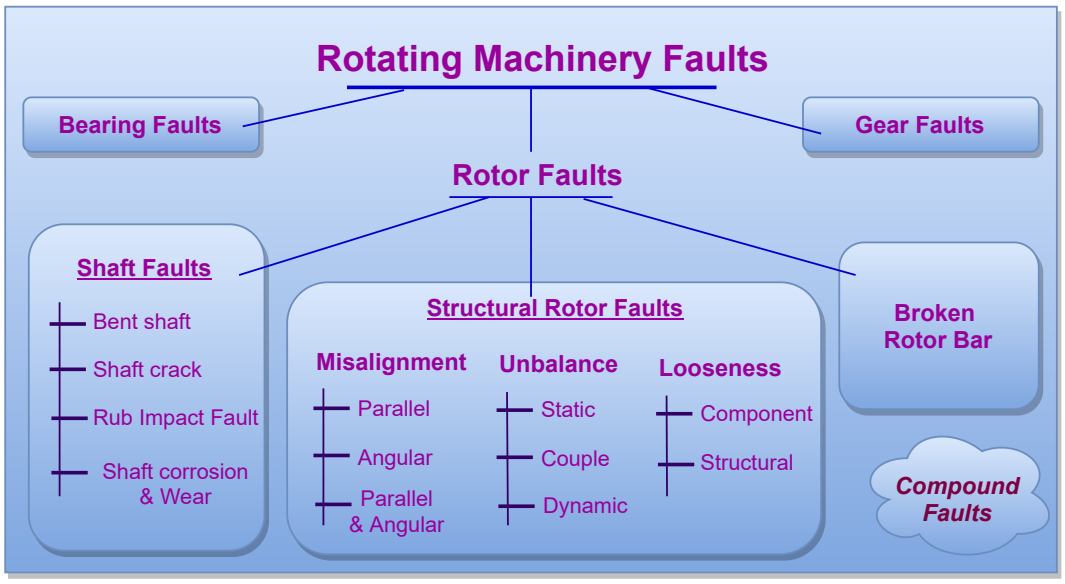

Figure 2.1: Rotating machinery fault categorization

2.2 Rotating Machinery Fault Categorization

As seen before, the right categorization of rotating machinery faults is significant for moving past component-wise analysis and enhancing the fault-characteristics-based analysis for RFD. An example of fault categorization is shown in figure 2.1. In this example, before the rotor faults are categorized from a 'cause of vibration' perspective, the rotor faults are isolated from the bearing and gear faults in the first step. Based on this, the primary cause of vibration is SRF or 1x fault, and the secondary cause is shaft-related faults. Together they make up the first two categories. In the meantime, the broken rotor bar (BRB) fault has also been added to the fault category list. The most common and vital rotor faults like unbalance (UB), misalignment (MA), and looseness (LS) are placed under SRF. Shafts can be affected by faults such as a bent shaft (BS), a rub impact fault (RIF), a shaft crack (SC), and corrosion and wear (Cr&Wr). These are believed to be part of the category of shaft fault and are often associated with SRF. In a practical scenario, it is impossible to assume that a single fault occurs in the rotor at any given time. Therefore, to represent the combination of multiple faults, we

incorporated compound fault (CF) in the fault grouping process.

The uneven distribution of mass in the motor stimulates the rotor to unbalance, which causes the inertia axis of the rotor to become imperfectly aligned with the geometric axis [58]. The asymmetry can cause misalignment in the applied load, the improper alignment of the couplings, shafts, bearings, and the thermal distortion of the bearing-housing supports. The bearings end up holding a heavier load than they were particularly designed for [59] due to this fault. The present literature is filled to the brim with research on bearing pedestal looseness. Improper assembly or the prolonged running of machinery can cause looseness [60]. Looseness related to the bearing can reproduce the effect similar to that of unbalance, whereas secondary damage and detachments are brought on by component-based looseness. It is worth noting that around 40.0% of rotor-related problems are associated with unbalance, 30.0% with misalignment, 20.0% to resonance, and the leftover 10.0% to other reasons [61]. The rub fault is brought on by the contact between the rotor and the stationary parts of the machine under tighter clearances. The shaft is no longer able to withstand forces released during normal operation because of shaft crack, and it is created due to extreme thermal and mechanical stresses [62]. A broken rotor bar fault [63] generally takes place in induction motor (IM) rotors, inducing an uneven current flow through the rotor and creating both thermal and bending issues with the rotor. The corrosion-based faults on the shaft's surface are worsened due to the electrochemical reaction from environmental factors [64].

2.3 Background of Rotor Faults

The nature and behavior of the vibration are affected by rotor-related faults; that is why vibration sensing has become the most popularly used signal sensing method for RFD [65]. The RFD researchers prefer the subharmonic and superharmonic frequency components of non-linear and complicated vibration motion, which characterize differ-

ent rotor faults precisely. A rotor testbed used for acquiring the vibration data with the provisions for rotor fault simulation is shown in Fig. 3.1 with description in section 3.2.

As seen before, this work displays a grouping of an array of generic fault categories that affect the rotor under the category of rotor faults. These include SRFs, shaft faults, and broken rotor bar faults. According to what was drafted in the introduction, these categories are further split into subcategories depending on the nature and cause of faults [66–68].

2.3.1 Structural rotor faults

Structural rotor faults are the main reason for abnormal vibration in the rotor and are called 1x faults. These can be further split into three: misalignment, unbalance, and looseness.

2.3.1.1 Misalignment

Misalignment is a scenario that takes place when bearings, shafts, and couplings are not properly aligned along their centerlines [68]. With prolonged machine operation, component expansion and cold alignment are respectively caused by heating and cooling. The net result is the misalignment of the machinery components. Misalignment can also be caused by continually operating with an unsteady foundation, a movement in the foundation, or the improper alignment brought on by imparted forces from other components [67]. Additionally, the risk of installation misalignment to the system increases if the couplings are not installed properly. In the meantime, faults due improper bearing seats and bent shaft causes a specific type of misalignments similar symptoms of normal misalignment [68]. When the bearings have to bear a higher dynamic load than they are designed for specifically, it leads to misalignment and unbalances, which ultimately causes failure due to premature fatigue. Misalignment causes excess heat and

friction and subsequent damaging of the component in the couplings, while in horizontal shafts, vibration on both the vertical and axial planes is brought on by misalignment. In the end, excessive vibration in the horizontal and axial planes characterizes any misalignment in the overhung horizontal shaft. In both these cases, axial vibration is a salient component. Inordinate horizontal and axial vibration points to misalignment in the case of the vertically aligned shaft [67].

Parallel misalignment, angular misalignment, and parallel and angular misalignment are the three categories of misalignments [69].

(i) Parallel misalignment : The centerlines of shafts joined by the coupling are parallel in the case of parallel misalignment. However, they will be at an offset, and it is characterized by solid radial vibration.

(ii) Angular misalignment: Angular misalignment takes place when a bending force is induced on the shaft by the joint at the coupling. In cases like these, a strong vibration in the axial direction is caused when joining shafts' centerlines are crossed at an angle between them.

(iii) Parallel and Angular misalignments: When parallel misalignment and angular misalignment simultaneously occur in a rotor system, it is called parallel and angular misalignment. This gives rise to vibration in both axial and radial directions.

2.3.1.2 Unbalance

Unbalance is the defective state of a machine that takes place when the centerline of the mass of the rotor (inertia axis) and the center of rotation (geometric axis) are non-coinciding [58]. The primary reasons for an unbalanced rotor are the adding of new fittings to the rotor before proper counterbalancing, rotor mass eccentricity produced by the uneven build-up of debris on the rotor, and assembly errors (unidentical blades of the wind turbine, windings of the generator rotor, etc.) [67].

Apart from these aspects, several other rotor faults can also cause unbalance. For

instance, unbalance can be brought on by factors like the falling of damaged rotor parts, a bent shaft or loose parts, or corrosion and abrasion. An unbalance fault can damage the critical components of the machine, including the gears, bearings, and couplings. As a matter of fact, the bearings are particularly affected by unbalance in that. As mentioned previously, the bearings experience ultimate breakdown due to premature fatigue as they are made to carry a higher dynamic load than what they are designed for specific. A wobbling movement that rotating structures face during operation characterizes the vibration that results from the unbalance. A radial vibration that is partly vertical and partly horizontal is produced by unbalance. Excessive vibration is a fine way to point out unbalance as the machine experiences more flexibility in the horizontal plane. Most forces are generated perpendicular to the shaft, which is why in an ideal scenario, axial measurements should point out weak vibration [67]. The dominant frequency component will be 1x in vertically placed shafts, as the unbalance is caused by the mass effects of radial plane vibration. A phase shift of 90° takes place as the sensor moves from the horizontal to the vertical position [58].

Generally, unbalance fault can be split into three types:

i) Static unbalance: It is the unbalance that is noted when at rest, in which the balance is affected by only one force. In the said scenario, there is a displacement in the inertial axis of a rotor, and it lies parallel to the axis of rotation. Because of a critical chance of uneven mass distribution that gives rise to a parallel shift between the axes, this problem is experienced widely in disk-shaped rotors. The risk of static unbalance is higher in simple systems than the risk of couple unbalance.

ii) Couple unbalance: In couple unbalance, the rotor looks balanced statically when two equal weights or forces are placed 180° apart. It is impossible to observe this issue when the rotor is at rest. The elongated cylindrical type rotors commonly face this phenomenon. The supposed 'wobble effect' is caused by a couple unbalance producing a 180° out-of-phase reading from opposite ends of the shaft. In general, this type of

fault is mostly observed in systems with more than one coupling or complex systems with unbalanced weights at multiple locations in the rotor.

iii) Dynamic unbalance: Dynamic unbalance is an unbalance condition that takes place in real systems. This unbalance condition presents a combination of static unbalance and couple unbalance. This type of unbalance is observed to be present in almost every rotor, and to address this, weights must be applied on a minimum of two planes.

2.3.1.3 Looseness

Based on whether the looseness affects by a structural component or mechanical part, this fault is categorized into two types: component looseness and structural looseness. The Component looseness occurs when the mechanical components are inappropriately fitted, whereas structural looseness occurs due to the relative movement among the surfaces of the fundamental structures. Excessive horizontal, vertical, and structural vibrations in the horizontal and overhung horizontal shafts are caused by this type of looseness. The issue is extended to the vertical shafts through excessive horizontal and structural vibrations. The looseness issue accounts for more vertical vibrations than horizontal vibrations [70].

i) Component looseness: Component looseness can be observed in rotating components and/or non-rotating connections that constrain the shaft to its rotating axis (such as the bearing base [pedestal], bearing mounts, and bearing caps) because of improper fittings, wear and tear, and thermal expansion. Secondary damage is induced in this issue as it leads to the components becoming damaged or separated from the assembly. For instance, relatively small residual misalignment can be caused by component looseness, thus producing increased vibrations affecting both the radial and axial planes. Unbalance may occur if the loosened components are rotor mounted.

ii) Structural looseness: The fundamental structures of an RM are not supposed to move freely. The structural looseness occurs when a little movement occurs between

the surfaces of structures like bedplates, a disintegrated concrete foundation, loose or distorted machine mountings, etc. It is usually produced between one vibrating component, which is generally the foot of the machine, and one stationary component, the foundation. Vibration in the radial plane can be caused by both structural looseness and soft foot looseness.

2.3.2 Shaft faults

Shaft faults directly affect the shaft like bending, cracking, rub impacts, and corrosion and wear. They can be brought on either as a result of SRF or because of other external reasons. Most of the shaft faults are considered as the secondary cause of abnormal vibration. Shaft faults which are known to be the most important, are laid out below.

2.3.2.1 Bent shaft

It is the most frequently observed RM fault and usually develops because of thermal distortion, creep, or a large unbalance force [71]. In contrast, another cause of thermal rotor bowing is rotor rub. Gravity can make a rotor can go through a cold bow during a resting position, particularly in shafts with a high length to width ratio. The static unbalance faulty condition's reemergence can be brought on by the rotor bow on a rotating machine. A bent shaft can also result from improper handling and high torque pressure. Effects that are similar to the misalignment are brought on by bent shafts, which causes the shaft to bear a higher dynamic load than they are specifically designed for, thus finally resulting in failure because of premature fatigue [67].

2.3.2.2 Shaft crack

Shaft crack is observed when weak spots in the shaft are developed because of severe thermal and mechanical stress or manufacturing defects, limiting its ability to hold up the forces produced during normal operation. The causative stress can be cyclical in

nature, wherein the initial crack converts into a fatigue fracture which results in an abrupt breakage of the shaft. The reduction in bending stiffness in the direction of the crack is the primary effect of shaft crack, resulting in inducing excessive 2x vibration in the shaft [71]. The rotor bow comes as the second effect, wherein the bending causes the natural axis shift corresponding to the direction of the crack. This effect gives rise to 1x components that will progressively add to the pre-existing residual unbalance frequency component [72].

2.3.2.3 Rub impact

A rub impact is created by the contact that emerged between the rotating components and the stationary components. Its effect will stay passive in terms of overall vibration. Highly non-linear vibrations are caused when a stator–rotor rub fault takes place in the system due to misalignment, excessive unbalance, self-excited instability, and resonance [73]. The stator–rotor rub is caused by the static forces working on the rotor in the radial direction and the thermal distortions in the casing. A thermal bow in the rotor can be the result of the heat induced by the asymmetric friction generated by the rub impact in a one-per-revolution fashion. This gives rise to both a phase difference and an unbalance effect. The degree of non-linearity grows with the increase in the rub impact, thus creating higher amplitude harmonics of rotating frequency.

2.3.2.4 Shaft corrosion and wear

Corrosion and wear fall under the category of non-fracture-type shaft failure. Corrosion results from the electrochemical reaction due to environmental factors, which ultimately leads to the metal being worn away. Moreover, this increases the stress and ultimately culminates in fatigue cracks. By and large, shaft failure will not be caused by these faults. That said, they may give rise to fatigue failures which leave clear evidence when they are in conjunction with the other faults. Here, cracks expand from the place

where the metal part is extracted due to the debris created by oxidation in a direction perpendicular to the applied stress on the area. In the meantime, a corrosion issue is presented by pitting causes short-term failure, resulting in a small amount of material loss from the shaft periphery [64].

2.4 Fault Diagnosis Approaches

In industrial applications, most RM fault diagnosis systems were based on traditional signal processing methods such as the discrete Fourier transform (DFT), FFT, EMD, WPD, HT, HHT, Wigner–Ville distribution (WVD), WT, etc. But nowadays, fault diagnosis framework follows either ML-based SL algorithms or DL-based algorithms as shown in Fig. 1.1. Considering the data acquisition phase, both ML and DL pursue the same data collection process, though they demand data at different quantity and precision. For an acceptable level of generalization, DL cannot compromise with the amount of data customarily needed by ML, but can deal with noisy data environments. But in feature processing phase, DL exercises the automated feature extraction and selection by learning discriminative features in an end-to-end manner. As opposed to DL end-to-end learning, ML goes for manual feature extraction which requires sufficient domain expertise and time.

The conventional SL algorithms are constrained in their ability to learn the non-linear relation of features. The need for extensive computation, the required time (especially for the feature processing), and the demand for specialized expertise in the domain, are other limiting factors. By using multiple-layer deep architectures, DL methods could imbibe advanced levels of representation of input data as they go deeper, which enables them to identify more complex features on their own. The most widely used SL and DL approaches in the literature of RFD, especially in SRF are discussed in this section.

2.4.1 Machine learning-based approaches

RFD research has been successful in employing a variety of SL algorithms before and after the advent of DL methods in the research arena. Irrespective of the fact that the ML methods require a complex feature engineering ,it has certain advantages that favours the machine health monitoring research, like the provision for applying domain knowledge and the comparatively fewer data requirements. One of the earliest methods found in the existing literature involves using ANN for motor fault diagnosis. Thereafter, the research in this domain flourished with several algorithms adopted across the supervised and unsupervised categories.

2.4.1.1 Artificial neural network

In AI-based machine health monitoring, the history of the application of ANNs stretches over three decades [74], and it has been an indispensable part of the literature on SL of RFD right from its inception. The approaches using ANN in RFD are custom-made according to the variations of ANN architecture and signal processing techniques used in feature processing. The method of integrating finite element analysis with the neural networks is adopted in specific works of RFD [33, 75]. Several researchers have tried to find out the best suited ANN architecture and feature extraction method for specific tasks in RFD [30, 76–78]. A few works have registered to incorporate the TD features in ANN, whereas FD and TFD features can be seen in a fair amount of works lately [23, 79, 80]. Certain significant contributions have been observed in the preprocessing of input data for the smooth learning of neural networks in RFD [26, 81], [82–85]. Moreover, ANN is viewed as a replacement for conventional feature extraction methods. One such approach was proposed with a two-level learning stage by Lei et al. [86]. A feature fusion model based on information entropy and the probabilistic neural network was proposed by [87]. While summarizing the literature it is clear that the majority of the researchers have tried different feature processing methods to

improve the output, while a fair number of works follow TD approaches, despite the prevailing trend of adopting FD approaches with ANN. Similarly, the proportion of works that have modified the ANN structure and the number of works that consider DFC feature extraction very high with ANN. Moreover, the percentage of works that deal with RFD exclusively (without considering bearing or gear faults) demonstrates the fact that ANNs are highly suitable for RFD analysis.

2.4.1.2 Support vector machine

As one of the frontrunners in both classification and regression tasks, SVM has been proven to be a decent performer in machine health monitoring over the last two decades. There are numerous ways in which researchers modify SVMs for adapting to their research problems in RFD like wavelet-SVM [88], multi-layer SVM [89], multiclass wavelet kernel-SVM [90], fuzzy SVM [91], proximal SVM [92], etc. By fusing the advantages of the information entropy method and SVM, Fei et al. [93] proposed an information fusion based method known as process power spectrum entropy and SVM. Tang et al. [94] showed that the multiclass SVM trained with chaos particle swarm optimization outperforms ANN in identifying rotor faults. In data and feature representation, SVMs show their variegation with the thermal images [28], 2-D gray-scale texture [29], histogram features [95], etc. In some works, SVMs work with features extracted by CNN to enhance the fault diagnosis performance [96]. There are some works reported in the literature where SVM gave far-reaching attention to the feature extraction process. SVMs used central limit theory [80], compensation distance evaluation technique [97], characteristic frequency band energy entropy [98], Park's vector approach [99], GA [37], etc., for features engineering of RFD. Certain works demonstrated the dominance of SVM over state-of-the-art ML techniques [79, 100, 101], while some others provided fault-specific contributions to RFD [25, 32, 102, 103]. It is worth noting that the researchers mainly attempt to change the kernel and incorporate new data structures to

SVM to achieve a decent RFD accuracy. SVMs account for the maximum share in ML methods for RFD analysis. SVMs succeed in utilizing information entropy fusion and spectrum analysis in certain applications, demonstrating its ability to deal with a host of input representations, while it is even used to evaluate the performance of feature extraction. Due to the excellent potential for handling outliers, SVMs can be used with any form of sensing method. Similarly, a number of the significant contributions that made use of DFC in RFD involved using SVM for classification. In short, SVMs are the most versatile of the ML models from an RFD perspective.

2.4.1.3 k-Nearest neighbor

k-NN is an instance-based, non-parametric algorithm renowned for its interpretability and ease of implementation. It is the most popular algorithm used in RFD after SVMs and ANNs. Chen et al. [34] proposed a k-NN based fault classification method that used of the 'maximum cross-correlation sum operator' as a similarity measure. The k-NN with Euclidean and Mahalanobis distances was applied for rotor fault classification by Biet et al. [104] using rotor flux measurements and classical electrical measurements, which was a follow up work of [105]. Glowacz et al. [39] proposed a method for selecting essential frequency components as features and applied to k-NN. Recently, Gohari et al. [106] studied the identification of unbalance parameters of a rotating shaft having multi-discs with k-NN. The fault specific frequency selection used with k-NN for selecting essential frequency components as features in some works [39, 107]. In the works [22, 100,108,109], k-NN was employed for comparison purpose and proven as one of the best-performing algorithms specifically appreciated for its faster and simpler operation. Even though k-NN is simple and convenient, it is not widely adopted in RFD. In fact, barely 10.0% of the works utilize k-NN despite its ability to deal with decision boundaries of any form. A number of researchers developed different similarity measures derived from well-known distance functions, while others operated well-established algorithms such

as GA to boost the performance of k-NN. However, the inability to recognize important attributes, the overheads involved in deciding the parameter 'k,' and the interpretability issue due to its non-parametric nature have been identified as the main limitations of this algorithm, which make researchers reluctant to adopt it for RFD analysis.

2.4.1.4 Naïve-Bayes

Naïve-Bayes is the most popularly used Bayesian model in RFD, which works on conditional probability basis. Wang et al. [110] used envelope features of the motor current extracted using Hilbert transform. Bayesian modeling has a drawback that it is unable to model and learn from the TS level change of data. As an attempt to mitigate this disadvantage, Yusuf et al. [35] presented the NB classifier on the fault groups created by the hidden Markov model (HMM). Xu et al. [36] developed a Bayesian belief network with three layers, namely machine running conditions layer, machine faults layer and fault symptoms layer with two topological configurations of causality, and fault symptom. It was one of the attempts to incorporate human expertise in the field of RM fault diagnosis. The higher chance of correlated features in the RFD dataset results in certain restrictions to apply Bayesian family classifiers, as the independence assumption is its core aspect. The inability to model and learn from the TS data, as well as the issues created in the classifier due to the absence of data, are just some of the limiting factors.

2.4.1.5 Fuzzy logic and other methods

The ability of fuzzy logic to mimic human reasoning has been utilized in RFD by certain researchers. El-Shafei et al. [77] used the learning vector quantization (LVQ) neural network in parallel to a fuzzy inference engine for addressing the SRF. The fuzzy Sugeno model adaptive neuro-fuzzy inference system (ANFIS) [111] is a commonly used fuzzy model in RFD [112]. Lei et al. [113] combined multiple ANFIS with GAs

for a more reliable and intelligent RFD. Zhang et al. [91] introduced a fuzzy support vector machine by integrating fuzzy logic with SVM. Qiu et al. [114] summarized a significant analysis of unbalanced non-linear rotor systems and lately, fuzzy neural network technology application in equipment fault diagnosis was proposed by Xu et al. [115]. The decision making based on fuzzy logic has advantages over other methods since the rules derived from the fault state play the key role. As the fuzzy classification system does not have the capacity for pattern recognition as other ML models do, and since certain concerns exist in terms of its implementation, it is not widely accepted in RFD scenarios. However, individual attempts to utilize the harmonics of the rotating frequency as the input to the fuzzy system and the use of ANFIS have returned some significant result improvements.

Among the remaining ML algorithms, the random forest was used by Yang et al. [109] with multiple class feature selection. Quiroz et al. [38] used the same algorithm with TD statistical features. The linear descriminant analysis algorithm was explored for RFD with acoustic signal input by Glowacz et al. [39], and decision tree was used by Nguyen et al. [116]. A vibration image-based diagnosis was proposed by Yan et al. [117] in which three features such as histogram of vibration image, histogram of oriented vibration image, and 2-D FFT generated from vibration images were used. The AdaBoost proposed in this work overcome the overfitting problem, and it fuses multiple features. Martin-Diaz et al. [40] selected AdaBoost for comparative analysis to prove its significance in RFD analysis.

2.4.2 Deep learning-based approaches

The abundance of data and the evolution of new algorithms supported by the hardware expansions have accelerated the development of DL, which is essentially a nested hierarchical and more abstract representational framework. The advantage of the automatic feature extraction of DL models made them widely accepted within the field of machine

fault diagnosis since this allows for avoiding the requirement of high domain expertise in feature engineering.

2.4.2.1 Convolutional neural network

The CNN is established as a convolution-based hierarchical image pattern recognizer and is the most widely used DL model in RFD. The discrepancy caused by using 1-D TS signals in a 2-D convolution-based network is solved by different methods such as 1-D to 2-D conversion or by introducing 1-D CNNs. Several image representations of input data have been widely used with CNNs such as orbital images [118,119], infrared images [120,121], bi-polar images [122], etc. There are several ways by which vibration data has been applied to CNNs in 2-D form. Symmetrized dot pattern image [123,124], 2-D image based on the spectrogram [125], continuous wavelet transform scalogram (CWTS) [126] are some of the examples. Dislocated time-series CNN is proposed by Liu et al. [127] to handle disparity of TS industrial data (most often in 1-D form) with 2-D images. Ince et al. [128] introduced adaptive 1-D CNN, while [129] demonstrated another 1-D CNN with fusing multi-sensor data for multiple motor faults. Similarly, vector CNN developed by Xiaoxun et al. [130], and a 1-D CNN-GRU model was proposed by Liao et al. [47]. A recent contribution in 1-D CNN was presented by Zhao et al. [131] to identify shaft misalignment and crack in the rotor system. The variant CNN structures like physics-based CNN [132], WDCNN (Wide first-layer kernels deep CNN) [133] are also present in the literature. In order to deal with multi stream data, a multi-stream CNN was proposed by Yuan et al. [134], and a multi-channel CNN was developed by Sonkul et al. [135]. A novel multi-mode CNN was proposed in [136] with adjustable filter banks to decide appropriate convolution mode. Scarce labeled samples issue of data has been addressed with the help of SVM to deal with CNN [137,138], where Xu et al. [139] developed a small data-driven CNN for similar issue. Hardware associated approaches with CNN that directly deal with the industrial IoT based hardware platforms, as well

as with other embedded systems, have also been observed in RFD literature [140, 141]. The majority of the RFD-related DL works adopted CNN as the classifier. A fair number of the related works altered the structure of CNN to make it compactable with RFD issues, while the review substantiates that utilizing DFC with CNN is close to impossible. It is also clear that 1D CNNs began to gain in popularity, and adding the discriminative ability to CNN ensure the wider acceptance.

2.4.2.2 Deep belief networks

DBN is an unsupervised, probabilistic DL network with a stacked structure of RBMs or AEs, pre-trained using a greedy learning algorithm. Oh et al. [45] developed a method for converting raw vibration signals to the image by the omnidirectional regeneration method. Guan et al. [142] made a significant contribution to structural fault identification by combining EMD and DBN. Compressed sensing was effectively applied by Shao et al. [143] along with convolutional DBN. To handle heterogeneous data sources, Yan et al. [144] proposed a multi-DBN model with information fusion. DBN was observed as a fault feature extractor by Shao et al. [145] and used for the performance evaluation of layer-wise feature extraction. Li et al. [146] applied DBN by stacking five RBM layers with Bernoulli functions for RFD and mechanical degradation assessment. Most of the works demonstrated the scope and possibilities of using DBN in terms of different aspects, including -feature extraction, -fusing of DBNs for multi-sensor data, -using DBN with regenerated input, and convolutional DBNs. Meanwhile, layer-wise feature extraction is another interesting aspect of utilizing DBNs. In short, the possibility of using extracted features rather than raw data is the distinguishing feature of this method.

2.4.2.3 Autoencoders & DNN models

As an unsupervised method capable of learning features, AE is widely used in many applications, including RFD. The basic AE models are enhanced by stacking multiple AEs to form a stacked autoencoder, while, through the addition of denoising capability, stacked denoising autoencoders can be developed. Chen et al. [147] proposed a simple and straightforward implementation of AE. A hybrid AE model combining one denoising autoencoder and multiple contractive autoencoders, was proposed by Shao et al. [148] that enhanced the feature learning ability of standard AE. Zhao et al. [44] illustrated SDAE to detect the rotor faults with FFT transformed FD signal as input. Meng et al. [149] addressed irregularity of raw vibration, generalization issue, etc., when using AE for SRF. Narendiranath Babu et al. [150] presented two AEs and a softmax layer for the realization of deep neural network (DNN) for RFD. The same authors enhanced the study with TD, FD, and TFD features and increased number of hidden layers in [151]. An SDAE was employed by Kong et al. [152] for identifying rotor related faults of aero-engines. Recently, Tang et al. [153] introduced an SAE to identify dynamic unbalance, and static unbalance faults. To improve the robustness of feature representation of sparse AE, Sun et al. [154] proposed adding partial corruption into the input with the help of the denoising coding. Lei et al. [43] proposed a two-stage learning process utilizing sparse filtering in SAE.

Extreme learning machines (ELM): To reduce the overhead of tuning the parameters of the DL models an ELM based AE was proposed by Yang et al. [155]. Another approach with ELM was proposed by Sharma et al. [156], in which RMS value of 3-phase voltage and current signature was considered as input to the model. A hierarchical structure designed by concatenating the forwarding list of ELM was proposed in [155]. Another architecture modification in ELM was observed in ensemble ELM, proposed by Wang et al. [157] for compound-fault diagnosis of rotating machinery. Recently Zhao et al. [158] proposed a multi-manifold deep extreme learning machine algorithm to classify the SRF

and the rub fault. In addition, DNN structure of multiple feed-forward layers (neither with RBMs nor with AEs) are also available in the literature of RFD. For instance, Bo et al. [159] developed a simple DNN model, with DFC parameters as input and Li et al. [146] applied DNN with five hidden hyperbolic tangent layers.

The autoencoder-based models are the most widely used DL models after CNNs in the RFD-related literature. In fact, since RFD uses the features learned by AE, only nominal attempts have been made to utilize DFC in the literature, and there is an ongoing trend within the research fraternity to enhance the feature-learning capability of AE. Deep AE performs well in terms of multiple-fault or overlapped-fault situations. Generally, SDAE performance has been studied with varying numbers of hidden nodes and different deep architectures. Meanwhile, ELM-based AE is becoming increasingly popular since it shifts the emphasis from the overheads involved in parameter tuning.

2.4.2.4 Sequential DL Models

An RNN is capable of learning temporal information of sequential data by holding the past information and sharing weights using its recursive structure.

LSTM & GRU: These are two variants of RNN that are equipped with gates to avoid vanishing gradient problem. There were only a few attempts made to capture the long-term dependencies of input TS signal using the sequential analysis techniques of AI. One such work was manifested by Lei et al. [46], who introduced LSTM for categorizing SRFs, bearing faults, and other compound faults in a wind turbine test rig. Xiao et al. [160] proposed a more advanced approach of hybrid feature learning that combined statistical parameters, recurrence quantification analysis, and three-layer stacked LSTM. Liao et al. [47] proposed a 1-D CNN-GRU architecture that adaptively learns fault factors. We can conclude the following things about sequential learning DL approaches in RFD. Given the majority of sensing methods used in RFD produce TS data, only a few attempts were made to pinpoint the sequential analysis for capturing

the TS behavior of RFD data, resulting in a nominal share for RNN in RFD literature. LSTM and GRU models are being recognized in recent times, which provides a roadmap for RFD analysis to explore the long-term dependencies. These models are easily incorporated with the other models to exploit the effectiveness of both.

Attention Mechanism (AM): A few applications have used AM in RM fault diagnosis for selecting features adaptively by utilizing the dependency information. Interestingly, a feature attention mechanism has been developed for adaptive feature selection by Huang et al. [161] using shallow multi-scale CNN for the classification. Li et al. [162] used AM for assisting the deep networks in locating the informative data segments as well as in extracting discriminative features for LSTM. Similarly, Wang et al. [163] used AM to optimize the CNN structure and Hao et al. [164] facilitated AM for optimal feature selection from the original vibration signal. A GRU with attention has been used by Zhang et al. [165] to perform classification without being affected by the length of the data. A motor fault diagnosis framework has been proposed by Yan et al. [166] in which an attention mechanism is used to integrate the features of different time points adaptively. An RNN based on an encoder-decoder framework with attention mechanism has been developed by Chen et al. [167] for the remaining useful life prediction of bearing. A few more works have been found in the literature almost following the same approach for using attention in fault diagnosis for wind turbines [168], gears [169], wheelset bearing [170], roller bearing [171] etc., with different DL models.

2.4.2.5 Generative adversarial networks

Being an input reconstruction based semi-supervised learning method, GAN turned out to be a critical solution to data-related issues in RFD. In one of such attempts, the data imbalance issue was addressed by Lee et al. [172] by applying GANs to oversample the minority class. Data imbalance issue was solved using an adversarial network as discriminator and CNN as the classifier, by Han et al. [173]. In this method, the small

datasets are transferred to the model, and then trained with the adversarial strategy. We can infer the following observations about GANs from the literature. GANs are mainly used for oversampling data in RFD research. While dealing with small-sized datasets, GANs appeared to be a solution to many issues, including data imbalance. The philosophy of adversarial training is applicable in other DL models as well.

2.4.3 Classifier fusion

The conventional fault diagnosis method of using a single information source together with a unique decision method exhibits certain shortcomings. Classifier fusion is an extensively used method to combine multiple classifiers to generate better classification results than any single classifier [174–178]. Niu et al. [179] proposed a decision fusion by finding the optimal sequence of classifiers' for fusion. It is based on selected decision vectors using the correlation measure of classifiers and the sensor fusion method using relativity theory. SVM, LDA, k-NN, improved iterative scaling, Gaussian mixture model, and LVQ classifiers were used, and multi-agent classifiers fusion algorithm was engaged for fusing them. The comparison results with majority voting and Bayesian belief classifiers showed the superiority of multi-agent fusion. The same authors continued the fusion method in [41] using multi-level wavelet decomposition with transient current as the input. Santos et al. [180] combined bagging, AB, general boosting projection, and RF classifiers to obtain an ensemble classifier for performing unbalance and misalignment classification of wind turbines under various speed and load conditions using current and vibration signals as inputs. The results of the proposed method compared with C4.5 DTs, k-NN and NB, and found that the AB using J48 DTs as base classifiers achieved the highest accuracy. Tao et al. [181] proposed a novel classifier ensemble technique known as weighted majority voting with a different confidence level to ensemble NB, RF, and SVM classifiers. Based on the vote through the confidence diversity, they assembled multiple classifiers, and the results were compared with the

conventional normal weighted majority voting method. The AB ensemble classifier presented by Martin-Diaz et al. [182] addressed the issue of imbalance data in RM fault diagnosis. The fusion of classifiers in RFD literature can be summarized as follows: The works that performed classifier fusion constitute only less than 5.0% of overall ML literature of RFD. SVM, k-NN, and NB are often found among the fusing classifiers. The different fusing methods like multi-agent fusion, Bayesian belief fusion, decision level fusion, majority voting, and its variants, have been used in most of the works, and their performance comparison has been carried out. The literature proves that classifier fusion in RFD can address multiple drawbacks of individual classifiers, but still, the research in this direction has a long way to go.

2.5 Summary of Literature Review

This section summarizes the research progress in RFD, in line with the various phases of data-driven AI-based RFD framework. The overall statistics of the state-of-the-art research is demonstrated in the heatmap shown in Fig. 2.2. In data collection, the data source selection is very critical in the case of RFD. From the literature summarized in the heatmap, it is identified that around 42.0% of works depended on RTB method for data collection and about 20.0% bank on the other sources, including test rigs for bearing, gear, or wind turbine whereas a few works utilized open-source datasets as well. The 30.0% of works collected induction motor current and voltage as input. It is noticeable that the ML had to depend hardly 13.0% on other open sources, while DL heavily draws around 34.0% for other sources. This indicates that testbed kind of data collection methods are unable to mitigate the data requirement issues of DL. The facts about the data source of RFD summarized so far draws in two important conclusions. Firstly, testbed, which is recognized as the primary data source of RFD, often fails to provide sufficient data for DL so that DL methods choose some open-source datasets. Secondly, these datasets lack rotor specific data though they have a

		SVM	ANN	Others	Multiple Classifiers	CNN	AEs & DNNs	Others	Overall
		Machine Learning				Deep Learning			
Faults Considered	UB	74.19	86.36	66.67	68.75	75	81.82	63.64	74.45
	MA	38.71	45.45	16.67	31.25	50	54.55	45.45	40.15
	LS	22.58	31.82	22.22	6.25	35.71	36.36	36.36	27.01
	BS	29.03	22.73	22.22	43.75	14.29	9.09	18.18	23.36
	SC	6.45	9.09	5.56	0	0	0	0	3.65
	RIF	12.9	22.73	11.11	6.25	25	18.18	18.18	16.79
	BRB	32.26	9.09	38.89	87.5	10.71	18.18	27.27	29.93
	CF	19.35	27.27	5.56	0	28.57	9.09	18.18	17.52
	OF	35.48	27.27	72.22	56.25	53.57	81.82	63.64	51.09
Signal Processing	TFD	25.81	31.82	16.67	56.25	21.43	9.09	18.18	26.28
	FD	51.61	45.45	38.89	12.5	21.43	36.36	27.27	35.04
	TD	12.9	18.18	27.78	18.75	0	0	0	11.68
	Comb.	9.68	4.55	16.67	12.5	14.29	9.09	45.45	13.87
	Raw Data	0	0	0	0	42.86	45.45	9.09	13.14
Monitoring Technique	Vibration	51.61	81.82	44.44	18.75	67.86	100	90.91	62.04
	V & C	29.03	13.64	22.22	56.25	7.14	0	9.09	20.44
	AE	3.23	4.55	11.11	0	3.57	0	0	3.65
	Temp.	6.45	0	5.56	0	3.57	0	0	2.92
	Comb. /Others	9.68	0	16.67	25	17.86	0	0	10.95
Data Source	RTB	51.61	72.73	27.78	12.5	42.86	36.36	27.27	42.34
	IM	38.71	9.09	44.44	68.75	14.29	9.09	27.27	29.93
	WT	3.23	9.09	0	6.25	10.71	9.09	18.18	7.3
	Other Source	6.45	9.09	27.78	12.5	32.14	45.45	27.27	20.44

Figure 2.2: Summary of literature review

large amount of bearing or gear fault data. The other challenges identified from the literature in acquiring faulty data are: i) In real situations, the machine runs in faulty conditions are very rare compared to its normal running conditions. So it is difficult to get faulty data keeping the balanced sampling. ii) Though we can simulate faulty conditions in testbeds and acquire data, it will be a challenging and difficult task to run the testbeds for a long time in a faulty environment. iii) Commonly available testbeds find it difficult to simulate the frequently changing RPMs, load, and other environmental conditions like noise, which makes the data unrealistic most of the time. This scenario opens a new research direction of data generation or augmentation. Among the signal

sensing methods in the data acquisition phase, the most commonly adopted approach is vibration analysis since SRF is the root cause that influences the characteristics and behavior of vibration.

In feature processing phase, DL exercises the automated feature extraction and selection by learning discriminative features in an end-to-end manner, and therefore, this phase is skipped in DL. But certain signal processing techniques have been utilized in DL models for data preprocessing and proper input representation. As opposed to DL end-to-end learning, ML goes for manual feature extraction, which is summarized from RFD perspective, as follows: i) The symptom parameters of SRF are frequency harmonics. Hence, the TD methods are unable to capture and utilize these parameters, resulting in poor performance for RFD. ii) The non-linearity and non-stationary nature of the rotor vibration signals complicates the accurate fault diagnosis for the FD methods. Because the Fourier transform is suitable only for stationary signal processing, and it fails to reveal the inherent information of non-stationary signals. In the case of the wavelet transform, the selection of a wavelet basis and its lack of ability to adapt to changes are the key factors adversely affecting the performance. iii) While the TFD processing techniques, particularly STFT, WVD, etc., gained immense popularity in RFD by its exceptional ability to handle non-stationary signals, these are exhibiting certain deficiencies. For instance, STFT fails to produce an ideal resolution to frequency and time simultaneously. Similarly, WVD produces inference terms in its decomposition process. The literature review statistics indicate that 18.0%, 40.0%, 31.0% of works still depend on conventional TD, FD, and TFD features, respectively. Irrespective of the fact that TFD processing is best-suited for extracting SRF specific features from the non-linear and non-stationary vibration signals, a more prominent share is enjoyed by FD operations.

In classification phase, around 80.0% of ML models operated in the RFD literature is constituted by SVM (41.0%), ANN (29.0%), and k-NN. The NB and all the

other classifiers provide less than 5.0% each in the literature. Similarly, CNN seized a predominant position among the DL methods with a 56.0% share in the literature followed by AE based models, while DBN also registered some significant contributions. In terms of accuracy and classification speed, SVM and ANN are top-notch, which is evident from their wide acceptance among the ML models in RFD. But, architecture versatility of ANN is higher compared to SVM, which is evident from the fact that a higher fraction of ANNs in the literature altered their basic structure in accordance with the RFD problem. NB and k-NN are worthy of their ability to deal with overfitting and have good interpretability while SVM and ANN still have to advance a lot in this regard. Since the raw signals acquired from sensors are affected by noises, it is observed that the researchers depended on preprocessing steps to deal with noisy raw data, rather than going for NB like ML models which project staunch robustness to noise. Among the DL methods, CNN is most popular in RFD but fails to incorporate fault specific features. But as we mentioned earlier, DBN, AE-based models, and DNNs give some scope in this direction since they prefer processed data to raw data. One of the remarkable features in the existing literature is that the fault specific discriminative feature extraction reduces the size of the DL model and eventually leads to better performance. Some other works indicated significant effort to change the 'black box' nature of DL by using proper visualization tools like t-SNE. Almost all DL algorithms are robust to noise, but CNNs present a bit more denoising capability compared to others, especially better than AEs. SDAE like models overcome these limitations. The performance generalization issues of DL related to the size and diversity of data, class imbalance, etc., have to be addressed in the literature.

2.6 Research Gap

The present research can be enhanced in the following ways to inspire research on RFD, with special emphasis on SRFs in order to provide a more generalized, industry

conforming, and realistic solution,

1. Synthetic data generation: To address the data scarcity issues, make use of synthesized data by engaging data augmentation or GAN-based data generation. Research can be carried in TS data augmentation, emphasizing the TS properties of the data, maintaining the correlation between different columns of multivariate TS, and finally confirming the labels of synthesized data. It results in the emergence of a challenging research problem in TS domain and augmentation literature.
2. Challenging datasets: Another research exists in bridging the gap between testbed data and real-world industrial data. As previously mentioned, developing a complex dataset by applying varying RPMs and load conditions in testbed simulations can give rise to novel issues in research and make the solution ready for industry.
3. SRF specific symptom parameters: Novel research challenges are posed by extraction and usage of fault specific symptom parameters in both refining classifier architecture and feature engineering process.
4. More domain-specific data representations: The imaging techniques used in the SRF literature are either stacking the raw data or arranging the data in a transformed domain, which loses the properties of the original signal. In short, the literature lacks the TS imaging technique in input data representation. Similarly, it is challenging to present proper embedding representations to transformer architectures from raw vibration data.
5. Sensor fusion: The complex RM systems deal with multi-sensor data, where each sensor signal acts upon the fault at different levels. No works reported in the literature of RFD considering the relative weightage sensor fusion.
6. More learning strategies: In RFD, there is the least application of classifier fusion and transfer learning kind of advanced learning strategies. Many opportunities can be opened up by merging new classifiers at different levels (data-level, feature-

level, decision level, etc.) in an attempt to improve accuracy.

7. Sequential deep learning: As indicated by the literature of RFD, no substantial attempts have been made so far to tap into the sequential nature of the sensed signals utilizing RNN based deep sequential models like LSTM and GRU. Moreover, no works were found in the literature to exploit the long-term or short-term dependency of the input data segments, which can be captured by an attention mechanism to produce better results.