Technology Agnostic Anomaly Detection using Multi-modal Sensory Data in Industrial IOT

Wesley O'Quinn and Shiwen Mao
Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
Email: weo0004@auburn.edu, smao@ieee.org

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437317

***Abstract*—Large scale deployment of the Internet of Things (IOT) technology has produced a disruptive effect in many fields in the recent past. Continuous connectivity combined with relatively low physical implementation cost has produced a Big Data paradigm shift, especially in industrial contexts. Unfortunately, the ability to process and adequately make use of this data has not kept pace with deployment. Specifically, models in use today lack the ability to perform well with data from a variety of sources. For instance, many models are trained using only one type of data. Even models trained on multi-modal data lack the ability to predict on different combinations of this data. Sensor deployment on identical machines is often different depending on context, leading to the need for multiple models created for the same machine. The data in question has the ability to radically shift how equipment failure is predicted and when maintenance is completed; when processed correctly. The cost savings on large industrial machines and potential saved downtime could be enormous. This research proposes investigation of a new unsupervised, technology agnostic anomaly detection framework that can be utilized on any combination of data modes for a given machine. This framework is then tested on a real-world anomaly dataset, with results achieved that are significantly better than prior approaches.**

***Index Terms*—Industrial Internet of Things (IIoT), Anomaly Detection, multi-modal sensory data, Technology agnostic approach.**

I. Introduction

In the world of industrial machinery, unplanned downtime represents a costly liability. According to the Wall Street Journal [1], "Unplanned downtime costs industrial manufacturers an estimated 50 billion US dollars annually." Equipment failures accounted for 42 percent of this unplanned downtime. Not only do unplanned equipment failures result in large business costs, so too does their method of prevention. Often in an attempt to prevent unplanned failures, companies will perform preventative maintenance on perfectly healthy pieces of machinery. Rather than performing maintenance driven by data, the company takes a one-size-fits-all approach. Just like unplanned failures, this leads to unneeded, albeit planned, down-time. Additional costs are incurred due to the maintenance itself. In a cost driven environment, neither of these outcomes is desirable. Continuous monitoring through IOT technology paired with Deep Learning driven insights, provides the natural solution to these two issues.

Anomaly detection can be utilized to detect rapid degradation trends in machinery to inform diagnostics and timely maintenance. This allows a predictive maintenance scheme to be utilized. For the most part, companies are realizing the potential associated with the techniques described above. Retrofitting existing machinery with IOT sensing technology is embraced and recommended. However, since the ideology associated with Industry 4.0 is relatively new, the IOT technology installed often represents a patchwork of capabilities. For instance, any large pump or motor may have vibration, temperature, pressure, current, acoustic, thermal imaging, and error logs installed. But the matter is complicated by the fact that any given pump or motor may have all, one, or any combination of the data modes described above installed. Although both Anomaly Detection and remaining useful life (RUL) have been investigated with regard to industrial IOT. The concept of a generalized model that can utilize multi-modal data, including unique combinations of that data, has not. Therefore, a technology agnostic, multi-modal approach to equipment monitoring represents an open area of research, which has the potential to add value in a variety of industries.

In this paper, we propose an unsupervised, technology agnostic anomaly detection model. This model leverages a convolutional neural network (CNN) feature extractor paired with a deep auto-encoder (AE) to predict equipment anomalies. The model development is split into two phases. The first phase investigates different CNN structures with a standardized AE to determine the best possible feature extraction technique. The method chosen for feature extraction is a VGG16 model using the concatenated output of layer FC1. This phase also investigates the conversion of sensory data into the Mel Spectrogram format to facilitate meaningful feature extraction. Phase 2 investigates the effects of sensor loss with respect to both training and test sets. Finally, the models from both phases are compared to a variety of other anomaly detection methods on a benchmark repository. Our methodology outperforms all other models on most metrics. Critically, on the key metric (i.e., AUC), our model outperforms all prior techniques by 9 percentage points. Key contributions of this work include:

- The development of a highly performing multi-modal based, unsupervised anomaly detection framework that can be used on a variety of sensory data.
- Providing optimal pre-processing techniques for sensor fusion.
- Selection of the best CNN architecture for feature extraction purposes.
- Investigation into the effects of sensor loss on model

performance.

The remainder of this paper is organized as follows: In Section II, we introduce the background and motivation of this research. We then present the proposed methods in Section III and our experimental study in Section IV. Future research directions are discussed in Section V and Section VI concludes this paper.

II. Background and Significance

As mentioned in the introduction, the ability to perform diagnostic processing on a variety of data type combinations represents a high value objective in the field of industrial machinery. Through literature review, we find that this specific objective has not been well studied as represented by the dearth of papers in the area. Provided below are the literature review results attained and a discussion on each work.

A. Unsupervised Anomaly Detection in Industrial IOT

Ref. [2] is an excellent introduction to the field of industrial IOT anomaly detection. This work investigates state of the art, discusses significance and challenges associated with the field, and proposes a new architecture for anomaly detection. One comment that particularly stands out, is that in the industrial manufacturing sector, anomalies are correctly classified only 20 percent of the time. Unfortunately, this makes sense. Anomaly detection is difficult; a supervised approach cannot often be used effectively, due primarily to the definition of the field. Anomalies do not occur often. This leads to a dearth of available data and an imbalance in data distribution that is available. Therefore, creative techniques must be utilized to predict events. Ref. [3] is one of the most recent works in this field and also utilizes a multi-source approach. This work utilizes a custom built model containing statistical extraction, convolutional layers, two-stage LSTM auto-encoder, and dense layers. This model is utilized in both anomaly detection instances and for predicting RUL. This work does not expressly investigate handling multi-modal data and does not consider channel loss; it does represent an excellent work that advances the state of the art in this field.

B. Multi-modal Approaches to Equipment Monitoring

The authors in [4] propose a novel technique in relation to multi-modal fault detection. Their specific use case is industrial refrigeration. The tabular data utilized includes power consumption, temperature, and current. This tabular data is fused with features drawn from thermal images utilizing a CNN. The tabular data and image features are then entered into an Autoencoder for prediction purposes. Faults are then predicted based on a threshold value associated with reconstruction error. This work does not consider channel loss, but does represent one of the only current approaches to fusing multi-modal data. The reason this concept is significant is because most industrial anomaly detection problems are inherently multi-modal in nature. Fig. 1 visualizes one of the simplest industrial system, namely a pump motor combination, and possible deployed IOT monitoring sensors. Even on this simple

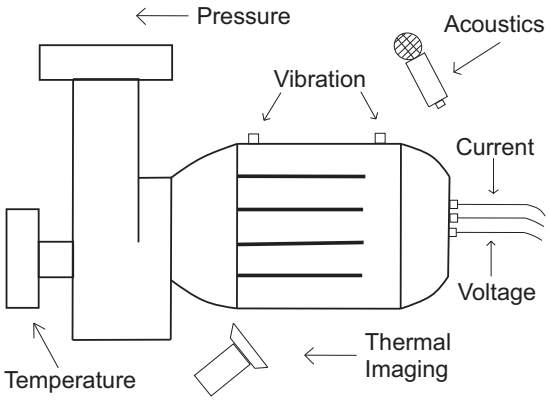

Fig. 1. Illustration of the multi-modal pump/motor sensors.

piece of equipment, there are multiple modes of data, including pressure, vibration, acoustics, current, voltage, thermal imaging, and temperature. As systems become increasingly complex, the number of monitoring data modes will increase as well. Leveraging this information in a meaningful way is critical.

C. Technology Agnostic Approaches

Technology agnostic methods constitute a deeply under-researched field. In [5], the authors provided one of the few examples of works in this field. The proposed TARF model utilizes a combination of Convolutional Neural Networks and a Domain Adversarial Deep Neural Network to predict body positions based upon phase data in RFID signals. The truly novel contribution is the combination of inputs through image encoding then extraction of features through a CNN. The idea is that the underlying features will be the same, no matter what type of IOT device is used to detecting them. This allows the same network to be used for a variety of data modes. This concept translates well to the work provided here, as our framework seeks to draw out anomalous features from multiple data sources and then fuse this information into a singular prediction. Many of the concepts from this paper were leveraged in the development of our techniques.

III. Research Design and Methods

A. Key Challenges

The main challenges associated with this work can be summarized in three main points:

1) Development of a standardized and modular sensor fusion method;
2) A technique which is robust enough to handle multiple channels of lost sensors;
3) An un-supervised scheme that can handle multiple data modes, while without requiring the more costly labeled data.

Item 1 is a difficult issue. Problems such as this do not often lend themselves to standardization. Even so, this piece is of the utmost importance. For a machine learning model to be useful, generalization is crucial. The goal behind this work is

to be agnostic to the hardware used for prediction, allowing the scheme to be rapidly deployed on a variety scenarios of different sensor types. A generalized feature processing technique is indispensable in this matter.

Item 2 is likely the most under researched portion of this work. The handling of sensor failure has been rarely, if ever, discussed in similar works. Usually, research is focused on initial deployment of networks and models. However, once these systems are deployed, sensors will fail. Such failure is not a question of if, but rather of when. The way the system functions following a sensor failure is of the utmost importance in terms of long-term viability.

Finally, the model must be unsupervised, full stop. This is because for most real world equipment anomalies, labelled data sets do not exist. Even where labeled data sets do exist, there is no guarantee that the failure mechanisms from the past will be the failure mechanisms of the future. This is inherently more difficult to develop when compared to a supervised approach, but in the long run, it will generally be more valuable.

B. Sensory Data Utilized

Model development and research is only as good as the data used to train the model. For this work, the goal is to utilize a two pronged approach to data development. For model development and initial testing, a repository is used that meets three basic criteria.

- Formatting and storage that allow rapid access and can be leveraged for expeditious proto-typing;
- Real-world data that is indicative of a difficult-to-solve anomaly detection problem;
- Prior publications having tested their methods on the repository to allow for comparison purposes.

We believe the dataset chosen and discussed below adequately meets all three of these criteria.

The baseline dataset detailed in [6] is used for bench-marking our proposed anomaly detection scheme. The data captures the operational sensor outputs of 15 high voltage converter modulators (HVCM) from the years of 2020 through 2022. The data collection process was undertaken at the the Spallation Neutron Source facility of Oak Ridge, Tennessee in the United States. It represents a world-class industrial anomaly detection benchmark, through providing a huge real-world data repository with quality labeled ground truth examples. Recent studies are also available, providing output metrics for comparison purposes [7]. Additionally, this dataset provides a variety of sensor waveforms, making it ideal for our multi-modal experiment. The sensor outputs available in this repository are summarized in Table I

TABLE I
SUMMARY OF THE AVAILABLE SENSORS (FOR ANOMALY DETECTION) [6]

Symbol	Description	Units
A+IGBT-I	Current through IGBT switch-phase A+	A
A+*IGBT-I	Current through IGBT switch-phase A+*	A
B+IGBT-I	Current through IGBT switch-phase B+	A
B+*IGBT-I	Current through IGBT switch-phase B+*	A
C+IGBT-I	Current through IGBT switch-phase C+	A
C+*IGBT-I	Current through IGBT switch-phase C+*	A
A-Flux	Magnetic flux density-phase A	–
B-Flux	Magnetic flux density-phase B	–
C-Flux	Magnetic flux density-phase C	–
Mod-V	Modulator voltage	V
Mod-I	Modulator current	A
CB-I	Cap bank current	–
CB-V	Cap bank voltage	V
DV/DT	Time derivative change: Mod-V voltage	–

C. Proposed Model Architecture for Anomaly Detection

The proposed scheme can be subdivided into three parts, which coalesce to address the key challenges enumerated before. The sections are: (i) Data pre-processing, (ii) Feature extraction/Sensor fusion, and finally (iii) Anomaly prediction. The proposed model architecture is shown in Fig. 2.

1) Pre-processing Techniques: To accomplish the goals discussed in Key Challenge 1, it is of the utmost importance that the sensory data be fused in such a way that allows the use of any combination of sensors. The proposition behind this methodology is that the underlying features will be the same for a given pump or motor, regardless of the sensor type used to detect the features. A CNN is an excellent method for extracting the underlying features. Additionally, the outputs of a CNN lend themselves well as input to an auto-encoder. The problem therefore lies in translating any given data type into a modular image format.

The method chosen for this was a Fast Fourier Transform (FFT) conversion into a spectrogram format. This method translates both a time and frequency component into a modular and easily accessible format that can be used by a CNN. These images were all shaped into the $224\times224\times3$ format to standardize the input window for use with pre-trained CNNs (excepting for the EfficientNetV2S model, which requires an input size of $384\times384\times3$). Critically, the method chosen for this translation utilized an MEL scaled format rather than a traditional spectrogram [8]. The Mel spectrogram is usually used for audio signals, which applies a frequency-domain filter bank to audio signals windowed in time. Although a detailed comparison of the two methods was not completed, initial testing was performed that showed the MEL format produced exceptionally better results than the traditional format. Additionally, this format is used almost exclusively in current state of the art audio classification techniques [9]. The better performance is likely due to enhanced resolution at lower frequencies, but a complete comparison represents a path of future work. A visual comparison of the two methods applied to the same sensor signal is provided in Fig. 2, where the left plot is the traditional Spectrogram and right plot is the MEL Spectrogram. The visual differences between the two methods are clearly distinguishable even to the human eye.

2) Feature extraction/Sensor Fusion: As mentioned in the previous section, a CNN was the obvious choice for feature extraction. Its ability to draw out high-level features in a variety of image types has been proven again and again [5], [10], [11].

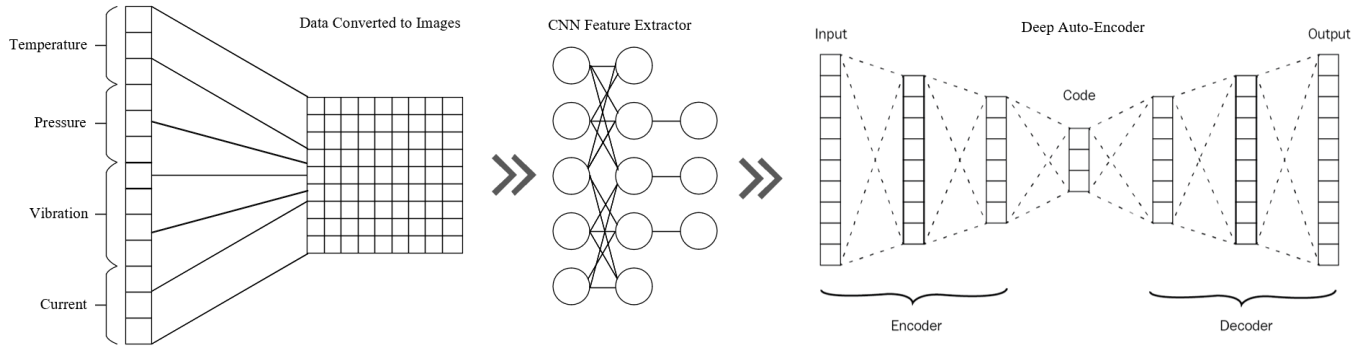

Fig. 2. Proposed anomaly detection model.

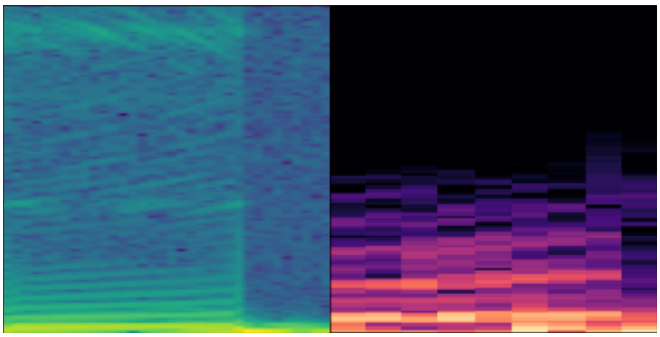

Fig. 3. Traditional Spectrogram (Left) vs. MEL Spectrogram (Right).

Additionally, a variety of different industrial data types such as thermal imaging, sensory data transformed to the frequency domain, vibration, acoustic signals, and video are all either native or close to native for this structure. Unlike pre-processing, the choice of CNN is a choice that can be heavily optimized, since there are so many architectures available. We utilized a phased approach to picking the optimal CNN, with the final choice being VGG16 Layer FC1. The sensor fusion occurs at the output of the CNN with a pre-defined one dimensional feature vector provided to the next section of the model. This vector is of the same size regardless of all the sensors being available or not. The size is defined by a concatenation of the maximum amount of sensors applicable in a given situation. If a sensor is failed or unavailable, then that portion of the feature vector is left as null.

3) Anomaly Prediction: This portion of the model is where the actual prediction of anomalies occurs. The use of Auto encoders for outlier detection has long been accepted as an optimal method for unsupervised approaches [12]. High dimensional data is entered into the encoder portion of the network, which is then compressed in a non-linear fashion. The network then attempts to try and generate the original high dimensional data with the decoder. The network is trained using reconstruction error as the target, hence learning what "normal" looks like. The idea is that anomalous data will be difficult for the network to generate hence leading to a high reconstruction error [13]. This attribute is leveraged to identify previously unseen failure mechanisms.

To be robust enough to handle channel loss, the autoencoder must be trained using data showing lost sensors. During phase one (CNN Selection), the autoencoder was standardized to provide a fair comparison with regard to feature extraction and sensor fusion. During phase two, the autoencoder structure was optimized to provide the best possible architecture for anomaly detection in the presence of lost sensors. The final structure chosen in this research is visualized in Fig. 2.

IV. Experiment Evaluation

As mentioned previously, the experimental and model development portion of the work was divided into two distinct phases. Phase one attempts to identify the optimal method of feature extraction/sensor fusion through testing a series of CNN techniques. During this time, the autoencoder portion of the model was frozen. Phase Two utilizes the optimal method chosen in Phase One, but seeking to optimize the autoencoder portion of the model while also training on a large dataset that includes sensor loss.

A. Phase 1: Feature Extraction/Sensor Fusion Selection

To allow rapid prototyping of models, this portion of the work solely used a subset of the HVCM data. Particularly, the data from the radio-frequency quadrupole (RFQ) subsystem is utilized. This includes all of the features discussed in Table I. The raw data is provided as a time series of three-dimensional array. Following conversion to image format, there are 9,660 normal images available, representing 690 images per sensor. An additional 2,548 anomalous images are available for testing purposes. For feature extraction purposes, the models listed in Table III were chosen and tested. The weights developed from Imagenet training were also used for each model, representing a degree of transfer learning. The output of each of these models was flattened, if required and concatenated for entry into the autoencoder. Following this sensor fusion, 500 samples were used as the normal data for model training. The test data was composed of 190 normal samples and 10 anomalous samples.

The autoencoder chosen for testing remains standardized throughout Phase 1. The goal here is not necessarily to

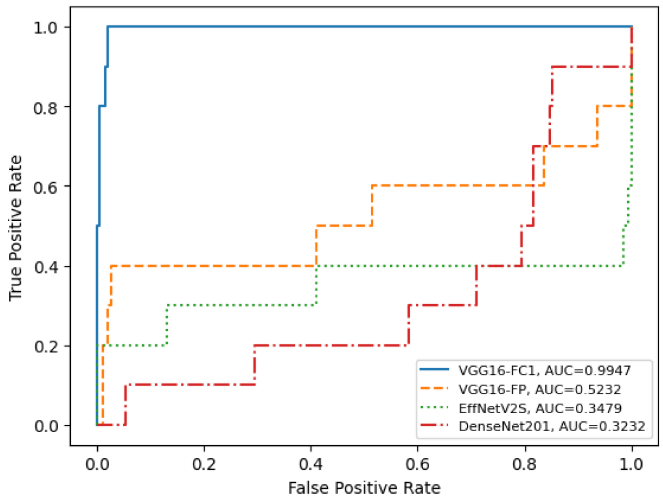

Fig. 4. AUC (Area under the ROC Curve) Curve: Phase 1 results.

produce the best accuracy possible, but rather provide a common baseline for evaluating efficacy of feature extraction. It consisted of 10 dense layers with the maximum number of internal units never reaching over 100. The results from this experimentation are provided in Table IV. The metrics presented in this table are defined below:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

$$FOR = \frac{FN}{FN + TN}, \tag{5}$$

where TP is the counts of true positives, FP is the counts of false positives, FN is counts of false negatives, FOR is the False Omission Rate (FOR), and AUC is the Area under the ROC Curve.

The VGG16-FC1 extraction method clearly distinguished itself from its counter-parts with excellent metric results. Note that these results were also achieved with a naive topology from an autoencoder perspective. Even so, outliers are clearly being distinguished from the normal distribution as shown in Fig. 4.

B. Phase 2: Sensor Loss Investigation

Once the method of feature extraction and sensor fusion is chosen, model optimization and sensor loss training will follow. This piece was chosen due to so little research being available on the effects of sensor loss to anomaly detection systems. It was known that the failure of sensors would likely have significant effects on model accuracy, but we desired to quantify that effect for future reference. To accomplish this at feature extraction, a single sensor is dropped from the total. This is repeated for all 14 sensors. Effectively, this increases the training sample size by a factor of 14. To train the model, 7,500 samples were utilized. This model was deepened to increase retention and then trained on the 7,500 samples. For testing purposes, the model was first evaluated against the test set in Phase 1.

The AUC results for this model on this test set were 0.821, representing a significant decline when compared to the model trained on only 500 samples of good data without sensor loss. Next 14 test sets were created consisting of 190 samples of normal data and 10 samples of anomalous data, with a different failed sensor in each set respectively. The model trained on the lost sensor data, along with the baseline model from Phase 1 were both evaluated against these 14 test sets. The results from these tests are available in Table V. These results illustrate just how difficult the sensor loss problem is. The AUC results may be somewhat misleading in that on all 14 sensor loss data sets, the VGG16-FC1: PH2 model was able to discriminate at least two anomalies out of the 10 present. That said, model performance in all cases was severely degraded by the loss of a single sensor. This scenario represents an open area of research with very little documentation available. The intent of this section is to bring the issue to the forefront to instigate further research.

C. Benchmark Comparison

Finally, the individual models developed and trained in Phase 1 and Phase 2 were evaluated in the same manner as described in [7]. Note, that all training was completed in an offline manner. The models were evaluated against a previously unseen test set composed of 81 normal samples and 50 anomalous samples. The metrics produced by the models were then directly compared with the results obtained in [7]. To maintain similarity, the FOR metric (defined in (5)) of all the schemes was evaluated.

As can be seen from Table II, the VGG16 Phase 1 model performed quite well on all accounts. In fact, on the metric that transfers best for comparison (AUC), the model developed in Phase 1 actually outperformed all other options by nine percentage points. This is significant as it demonstrates the efficacy of using a multi-modal approach to anomaly detection. By fusing these sensors together, increased performance is attained. Note also that although model VGG Phase 2, performed better than some prior methods it clearly does not distinguish as well as the model trained on only good data. Although not directly investigated here, it is implied from the results that sensor loss included in the training set will degrade model performance. Overall though, the proposed framework outperforms all other methods in almost all of the metrics evaluated in the experiment.

V. Future Works

The most difficult problem associated with the work is the handling of lost sensors. This represents a key gap in literature and is a core issue for companies implementing these Predictive Maintenance solutions. A proposed method for

TABLE II
BENCHMARK RESULTS

Metric	VGG-PH1	VGG-PH2	LSTM	GRU	CLSTM	IF	SVM	LOF	RF	DT	KNN	CNN	CNN-AE	FNN-AE
Precision	**0.96**	**0.97**	0.91	0.90	0.90	0.77	0.41	1.00	0.35	0.39	0.30	0.30	0.91	0.91
Recall	**0.94**	**0.70**	0.88	0.87	0.87	0.75	0.83	0.77	0.73	0.43	1.00	0.78	0.80	0.76
Accuracy	**0.96**	**0.88**	0.87	0.85	0.85	0.69	0.58	0.82	0.87	0.82	0.89	0.87	0.80	0.76
F1	**0.95**	**0.81**	0.90	0.88	0.88	0.76	0.55	0.87	0.47	0.41	0.47	0.44	0.85	0.83
FOR	**0.04**	**0.30**	0.20	0.22	0.22	0.42	0.14	0.48	0.03	0.10	0.00	0.02	0.38	0.48
AUC	**0.99**	**0.76**	0.90	0.89	0.89	0.67	0.63	0.76	0.66	0.64	0.65	0.64	0.76	0.69

TABLE III
FEATURE EXTRACTION MODELS TESTED

Model	Output Layer	Output Shape
VGG16 [14]	Final Pooling	7, 7, 512
VGG16 [14]	Layer FC1	1, 4096
DenseNet201 [15]	Average Pool	1, 1920
EffNetV2S [16]	Top Drop Out	1, 1280

TABLE IV
FEATURE EXTRACTION EXPERIMENTATION RESULTS

Model	Accuracy	Precision	Recall	F1	AUC
VGG16-FP	0.94	0.40	0.40	0.40	0.52
VGG16-FC1	**0.98**	0.80	**0.80**	**0.80**	**0.99**
DenseNet	0.90	0.0	0.0	0.0	0.32
EffNetV2S	0.96	**1.0**	0.20	0.33	0.35

TABLE V
SENSOR LOSS TESTING

Model	VGG16-FC1: PH1	VGG16-FC1: PH2
Accuracy (Best)	0.92	0.92
Accuracy (Average)	0.91	**0.92**
AUC (Best)	**0.54**	0.51

addressing this issue is through synthetic data augmentation. This method would allow much higher training volumes, improving discrimination. Another item of investigation would be leveraging audio specific CNNs; since spectrograms are native to these models.

VI. CONCLUSIONS

In conclusion, this paper investigated the use of a sensor fusion techniques which allows a single model to learn on multiple modes of data. A variety of feature extraction techniques were compared, while the VGG16 CNN model clearly distinguished itself as the optimal method. Even in initial training without optimization, the results were excellent. Next, a thorough investigation with regard to sensor loss was performed, which demonstrated the difficulty associated with this task and clearly defines a future area of research. Finally, the models developed by this work were tested against prior techniques used on this dataset. Our model clearly distinguished itself as an optimal technique, outperforming past methods by many basis points, demonstrating the efficacy of a multi-modal approach.

ACKNOWLEDGMENTS

This work is supported in part by the NSF under Grant CNS-2107190, and by the Wireless Engineering Research and Education Center at Auburn University, Auburn, AL, USA.

REFERENCES

[1] "How manufacturers achieve top quartile performance," *The Wall Street Journal*, 2017.
[2] P. Kamat and R. Sugandhi, "Anomaly detection for predictive maintenance in Industry 4.0–A survey," *E3S Web Conf.*, vol. 170, p. 02007, 2020.
[3] H. Nizam, S. Zafar, Z. Lv, F. Wang, and X. Hu, "Real-time deep anomaly detection framework for multivariate time-series data in industrial IoT," *IEEE Sensors Journal*, vol. 22, no. 23, pp. 22 836–22 849, Dec. 2022.
[4] F. Cordoni, G. Bacchiega, G. Bondani, R. Radu, and R. Muradore, "A multi–modal unsupervised fault detection system based on power signals and thermal imaging via deep autoencoder neural network," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104729, 2022.
[5] C. Yang, X. Wang, and S. Mao, "TARF: Technology-agnostic RF sensing for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 636–647, Feb. 2023.
[6] M. I. Radaideh, C. Pappas, and S. Cousineau, "Real electronic signal data from particle accelerator power systems for machine learning anomaly detection," *Elsevier Data in Brief*, vol. 43, p. 108473, Aug. 2022.
[7] M. I. Radaideh, C. Pappas, J. Walden, D. Lu, L. Vidyaratne, T. Britton, K. Rajput, M. Schram, and S. Cousineau, "Time series anomaly detection in power electronics signals with recurrent and ConvLSTM autoencoders," *Elsevier Digital Signal Processing*, vol. 130, p. 103704, Oct. 2022.
[8] K. Doshi, "Audio deep learning made simple - why Mel spectrograms perform better," Feb. 2021. [Online]. Available: https://ketanhdoshi.github.io/Audio-Mel/
[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. ICASSP 2017*, New Orleans, LA, Mar. 2017, pp. 131–135.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, May 2017.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR 2016*, Las Vegas, NV, June/July 2016, pp. 770–778.
[12] N. Japkowicz, C. Myers, and M. Gluck, "A novelty detection approach to classification," in *Proc. IJCAI 1995*, Montreal, Canada, Aug. 1995, pp. 518–523.
[13] C. Yang, X. Wang, and S. Mao, "Unsupervised detection of apnea using commodity RFID tags with a recurrent variational autoencoder," *IEEE Access Journal*, vol. 7, no. 1, pp. 67 526–67 538, June 2019.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, Apr. 2015. [Online]. Available: https://arxiv.org/abs/1409.1556
[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *arXiv:1608.06993*, Jan. 2018. [Online]. Available: https://arxiv.org/abs/1608.06993
[16] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. 38th International Conference on Machine Learning*, Virtual Conference, July 2021, pp. 10 096–10 106.