A Semantic and Adaptive Two Staged Analysis for Civil Aviation Event prediction using TF-IDF

Deva Hema[1], Ankith Kishore[2], Akash M[2], Sathyaa Govindharajan[2]

[1]Assistant Professor, [2]Student

Department of computer science Engineering, SRM institute of Science and Technology Chennai, Tamil Nadu, India

ABSTRACT

This project aims to develop a Natural Language Processing (NLP) system for improving the analysis of aviation safety occurrence reports. The existing system for analyzing aviation safety occurrence reports relies heavily on human effort to categorize and label reports, which can result in interrater variability and may miss important information. The proposed NLP system addresses these issues by leveraging machine learning techniques to automatically classify, extract, and generate insights from reports. The objectives of this project are to improve the accuracy, efficiency, and consistency of safety data analysis in aviation and to identify precursors to adverse events, which can help operators take preventive measures to avoid potential safety incidents. The scope of the proposed system is to provide a more efficient and accurate method for analyzing aviation safety occurrence reports, which can ultimately improve the safety and security of commercial airline operations. As a result whereas the performance of the human factors model was noticeably worse, the model accurately predicted the labels of the papers in the test set. High- quality predictions require high-quality, homogeneous annotations in the training data.

INDEX TERMS: HFS - Hybrid Feature Selection, ICAO – International Civil Aviation Organization, NTSB - National Transportation Safety Board, NLTK - Natural Language Toolkit, ASRS - Aviation Safety Reporting System

I. INTRODUCTION

A key component of safety management systems is incident reporting, which gives safety analysts information about the frequency of hazards and accidents so they can identify, evaluate, and reduce risks. Service providers are required by the International Civil Aviation Organisation to report events and accidents, which are then looked at by impartial agencies like the NTSB (National Transportation Safety Board) and TSB (Transportation Safety Board). Airline Safety Reports and the NASA Aviation Safety Reporting System are two examples of organisations that also gather and process reports outside of the mandated system. Despite the enormous volume of reports that are gathered every year, considerable work is needed to effectively categorise and study occurrences in order to monitor risks and extract insightful data.natural Language Processing techniques offer a promising approach to reducing the workload of processing reports and providing additional insights to safety analysts. The proposed system aims to leverage NLP techniques to automate the categorization and analysis of occurrence reports to facilitate efficient risk management in safety-critical systems. The proposed algorithm will extract information from reports using various NLP methods, including named entity recognition, topic modeling, and sentiment analysis. The proposed system's results will be compared with existing algorithms using metrics such as precision, recall, and F1 score. The proposed system has significant social relevance, as it can help improve safety in critical systems like aviation, where the consequences of errors can be catastrophic. It also has potential applications in other safety- critical domains, such as healthcare and transportation.

II. LITERATURE SURVEY

Yagya Raj Pandeya et al(2020) [1] had proposed Visual Object Detector for Cow Sound Event Detection.This study investigated rare sound event detection through a comparative analysis of conventional CNN with visual object detectors. The experiment was performed using two synthetic datasets for cow sounds and urban sounds. It is efficient to Discriminate features learned by the model and improve the sustainability of the system but it had some drawbacks such as Tedious message updating and for high cardinality, the feature space can explode.

Leveraging Phase Transition of Topics for Event Detection in Social Media was suggested by Pedro H. Barros et al (2020) [2]. This study models the occurrence of an event in social media using entropy. We found that entropy of the bigrams extracted from social media alters its dynamics during the occurrence of an event, and we observed a continuous phase transition of the entropy dynamics. While it is effective to reduce the size of datasets in order to eliminate pointless and duplicate features, it is challenging to construct models and speed up processing at the risk of decreasing detection accuracy.

For the purpose of detecting short video events, Ping Li and Xianghua Xu (2020) [3] suggested Recurrent Compressed Convolutional Networks. This research focuses on fresh subjects like short video event detection, which has not yet been explored but is crucial to intelligent video analysis. In order to comprehend the information of brief videos, this paper presented an RCCN architecture, or recurrent compressed neural network. This had the advantage of minimising loss and locating the best solution while yet being sensitive to the particulars of the training data.

Diego De Benito-Gorrn et al (2021) [4] had proposed A Multi-Resolution CRNN-Based Approach for Semi-Supervised Sound Event Detection.This work presents a method to better modelling the different temporal and spectral characteristics of sound events in the task of Sound Event Detection and hypothesized the features extracted using different time-frequency resolution parameters are able to represent certain event categories in a more recognizable way. It has an efficiency of Eliminating superfluous features and filtering out unnecessary data, it reduces the complexity of the data and its dimensions but it takes huge time and economic cost to construct.

The Chinese Event Detection Based on Multi-Feature Fusion and BiLSTM was proposed by Guixian Xu et al (2019) [5]. To produce output vectors comprising contextual information at the sentence level, the word vectors in the phrase are progressively entered into the BiLSTM model. To identify the trigger words, the BiLSTM output vectors are finally fed through the Softmax classifier.

Yanping Chen et al (2020) [6] had proposed A History and Theory of Textual Event Detection and Recognition. This article provide a comprehensive survey about textual event detection and recognition. The methodologies for building events for information exploration are divided into three types: documental events, frame events and graphic Documental events are processed at the document level, where the document has the smallest granularity. It has an efficiency of reduction in features to improve the quality of prediction. Capable of further reducing the required level of human effort. But had drawbacks like it Cannot be implemented real time and Cannot reduce the variance of predictions.

Saeed Afshar et al (2020) [7] had proposed Event-Based Object Detection and Tracking for Space Situational Awareness. In this work, the (Rhetorical Structure Theory) RST event-based space imaging dataset was presented. The labeled dataset, augmented with a larger unlabeled dataset, provides a test bench for investigation of event-based algorithms for the unique and challenging space imaging environment. Had drawbacks like high cardinality, the feature space can explode, decrease the generalizability of the model.

Aymen Yahyaoui et al (2021) [8] had proposed READ-IoT: Reliable Event and Anomaly Detection Framework for the Internet of Things. This paper introduces a Reliable Event and Anomaly Detection Framework for the Internet of Things (READ-IoT for short). The designed framework supports outliers management in IoT. It handles events and anomalies to a common and integrated rule-based and machine learning- based detection.

Search on Semi-Supervised Sound Event Detection Based on Mean Teacher Models Using ML-LoBCoD-NET had been proposed by Jinjia Wang et al (2020) [9]. In this paper, the MRNN-Att network is suggested for the job of weakly labelling sound event recognition. The location information of the target item is lost during the CNN pooling operation. The ML-LoBCoD-NET, which is powered by the ML-LoBCoD algorithm, is the foundation of the MRNN-Att network.

Wenzhong Yang et al (2019) [10] had proposed Sina Weibo Bursty Event Detection Method. In this paper, the problem of detecting Sina weibo bursty events in a complex social network environment. Introduce the key microblog computing model into the eld of event detection. The user behavior, user attributes and text content are used to solve the problem of insufficient burst features. It has an advantage of Achieving better performance metrics than alternative state-of-the-art models. But had some drawbacks like, the collected dataset is not reliable and needs address more features, High prediction complexity for large datasets.

III. SYSTEM ARCHITECTURE

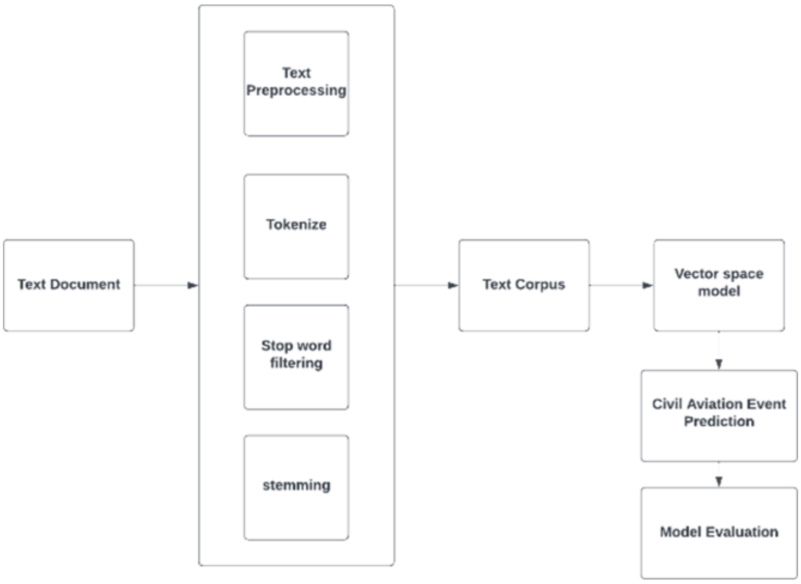


Fig 5.1 System Architecture

Text Preprocessing:

Text preprocessing is an essential step in natural language processing (NLP) that involves cleaning and transforming raw text data into a format that is easier to work with for further analysis. The preprocessing phase typically involves several steps, including tokenization, stop word removal, stemming, and lemmatization.

Tokenization:

Tokenization is the process of splitting text into individual words, or tokens, by separating them at spaces, punctuation marks, and other delimiters. This step is necessary because most NLP algorithms require input in the form of tokenized text.

Stop Word removal:

Stop word removal involves eliminating commonly used words such as "the", "and", and "is" that do not carry much meaning and can interfere with the accuracy of analysis. These words are often removed because they are considered noise, or irrelevant information, in the text.

Stemming:

Stemming is the process of reducing words to their base form by removing suffixes, prefixes, and inflections. For example, the stem of the words "walking", "walked", and "walks" is "walk". This step is useful because it reduces the number of different forms of a word that need to be considered, making it easier to identify patterns in the text.

Text Corpus:

Text Corpus In natural language processing, a text corpus refers to a large and structured set of texts or written materials that are collected and stored for linguistic analysis. A corpus can consist of various types of written material, including books, articles, web pages, social media posts, and more. Text corpora are used for various tasks, such as training language models, identifying patterns and relationships between words, and building applications that require natural language understanding.

Vector Space Model:

The Vector Space Model is a mathematical model used to represent text documents as vectors in a high-dimensional space, where each dimension corresponds to a unique word in the corpus. By representing documents in this way, we can measure their similarity and perform various operations, such as ranking search results or clustering similar documents. The most commonly used approach to create these vectors is the Term Frequency- Inverse Document Frequency (TF-IDF) scheme, which assigns a weight to each word based on its frequency in the document and its inverse frequency across all documents.

IV. METHODOLOGY

Proposed System:

The proposed system introduces a method for defining the inflection point and slope for input values that is smaller than the inflection point for activation, which prevents gradient saturation and reduces computational complexity. It also considers class weights during backpropagation to address the negative effects of imbalanced datasets. This algorithm has demonstrated significant improvements in average class accuracy and processing time for classification tasks in textual content, achieved by combining a modified leaky rectified linear unit function with a modified loss function.

Advantages:

Effectiveness for distributed optimization Improve the operational efficiency. Lightweight and fast model. Minimal time cost and memory usage. Can be adopted for better prediction in industrial applications. Increasing the precision and recall performance. Scale to incredible model quickly, easily, and cheaply

Proposed Algorithm:

Term Frequency - Inverse Document Frequency (TF-IDF)

This factor measures the frequency of a term in a document. It is calculated by dividing the number of times a term appears in a document by the total number of terms in the document. Inverse Document Frequency (IDF): This factor measures the rarity of a term in a collection of documents. It is calculated by taking the logarithm of the total number of documents in the collection divided by the number of documents that contain the term.

The formula for TF-IDF is:

TF-IDF = TF * IDF

The TF-IDF value for a term in a document increases as the frequency of the term in the document increases, but decreases as the term appears in more documents in the collection. This ensures that common words like "the" and "and" have low TF-IDF values, while rare and important words have high TF-IDF values. TF-IDF can be used for various applications, including document classification, information retrieval, and content-based recommendation systems. It is a powerful tool for analyzing text data and extracting meaningful insights from large collections of documents.

Advantages:

Reduce the weight of tokens which appear frequently compared to tokens which appear rarely. You have some basic metric to extract the most descriptive terms in a document. Makes rare words more prominent and effectively ignores common words.

Long Short-Term Memory (LSTM) Algorithm

Deep LSTM networks, which can recognise even more intricate patterns in sequential data, can be made by stacking LSTMs. Other neural network topologies, such as the bidirectional recurrent neural network (BRNN), can also be utilised in conjunction with LSTMs.

Forget Gate: The forget gate purges the data that is no longer relevant in the cell state. The gate receives two inputs, x_t (input at the current time) and h_t-1 (prior cell output), which are multiplied with weight matrices before bias is added. The output of the activation function, which receives the outcome, is binary. If a cell state's output is 0, the piece of information is lost, however if it is 1, the information is saved for use in the future.

Input gate: The input gate updates the cell state with pertinent information. To start, the inputs h_t-1 and x_t are used to regulate the information using the sigmoid function and filter the values that need to be remembered similarly to the forget gate. Then, a vector containing every possible value from h_t-1 and x_t is created using the tanh function, which produces an output ranging from -1 to +1. To extract the useful information, the vector's values and the controlled values are finally multiplied.

Output gate: The output gate's job is to take meaningful information out of the current cell state and deliver it as output. The tanh function is first used to the cell to create a vector. The data is then filtered by the values to be remembered using the inputs h_t-1 and x_t, and the information is then controlled using the sigmoid function. The vector's values and the controlled values are finally multiplied and supplied as input and output to the following cell, respectively.

Advantages:

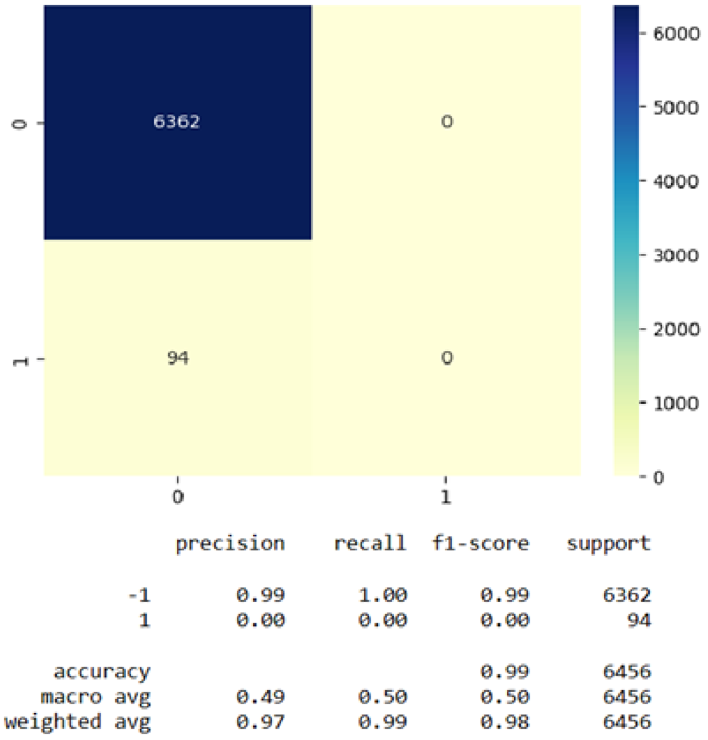
Can be used when dealing with large sequences and accuracy is concerned. Improved method of back propagating the error. Explicitly designed to deal with the long-term dependency problem.

Bidirectional Recurrent Neural Network (BRNN)

BRNN is a neural network commonly used in machine learning to process sequential data, including natural language, speech, and time series data. Unlike standard RNNs, which analyze input sequences in a unidirectional manner, from the start to the end, BRNNs process sequences in both directions, enabling them to consider past and future inputs in the sequence. This attribute helps BRNNs capture long-term dependencies and identify patterns in both directions.

V. RESULT AND DISCUSSION

The accuracy rate for A Semantic and Adaptive Two Staged Analysis for Civil Aviation Event prediction using TF-IDF can vary depending upon the various classification algorithms such as Random Forest Classifier, AdaBoost classifier, logistic regression classifier. The Logistic Regression (Figure 1) is a type of algorithm that is used for classification problems, where the predicted output is either a 0 or 1 (binary classification). It is based on a linear combination of features that result in the output of either 0 or 1. The Random Forest Classifier (Figure 2)is an ensemble learning method used for classification. It combines multiple decision trees and outputs the class that is the mode of the classes (classification) of the individual trees. The AdaBoost Classifier (Figure 3) is also an ensemble learning method used for classification. It combines multiple "weak" learners to create a single "strong" learner that yields better performance. The outcomes of using these algorithms, they are able to provide more accurate results than traditional machine learning algorithms. They are also able to identify patterns and anomalies in the data and provide better predictions of future decisions

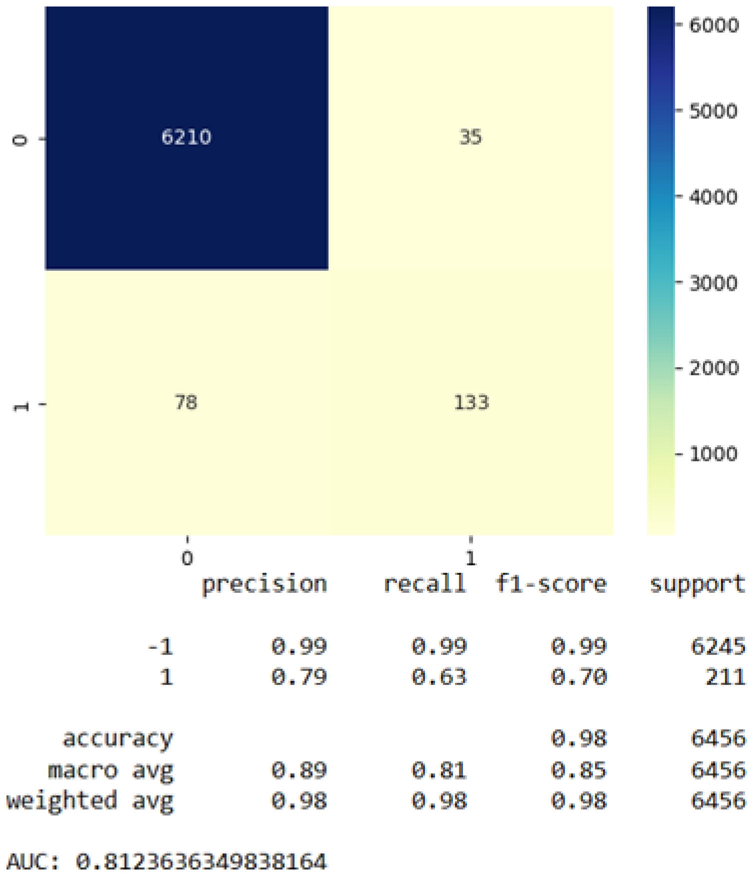


```
              precision    recall  f1-score   support

          -1       0.99      1.00      0.99      6362
           1       0.00      0.00      0.00        94

    accuracy                           0.99      6456
   macro avg       0.49      0.50      0.50      6456
weighted avg       0.97      0.99      0.98      6456
```

Figure 1: The Logistic Regression



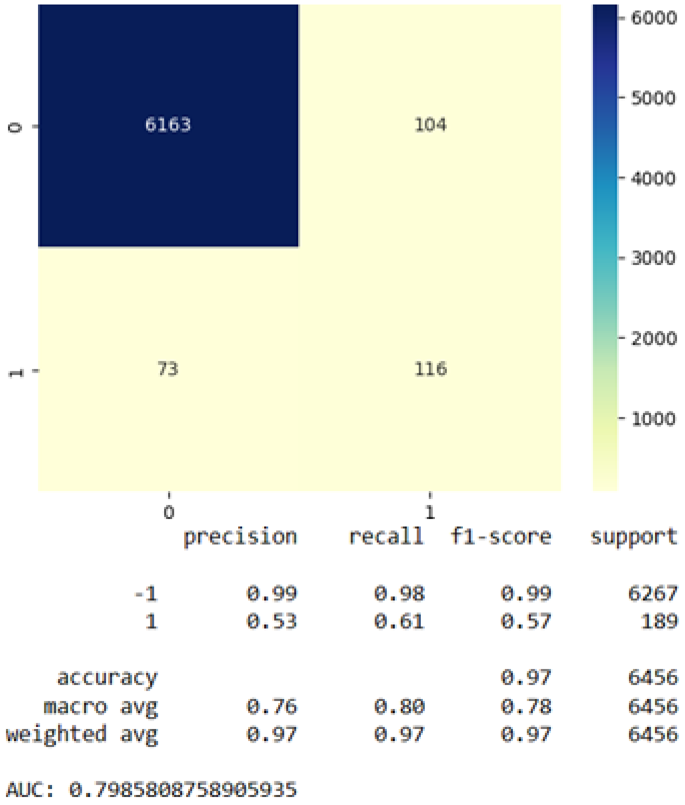
```
              precision    recall  f1-score   support

          -1       0.99      0.99      0.99      6245
           1       0.79      0.63      0.70       211

    accuracy                           0.98      6456
   macro avg       0.89      0.81      0.85      6456
weighted avg       0.98      0.98      0.98      6456

AUC: 0.8123636349838164
```

Figure 2: Random Forest Classifier



```
              precision    recall  f1-score   support

          -1       0.99      0.98      0.99      6267
           1       0.53      0.61      0.57       189

    accuracy                           0.97      6456
   macro avg       0.76      0.80      0.78      6456
weighted avg       0.97      0.97      0.97      6456

AUC: 0.7985808758905935
```

Figure 3: AdaBoost Classifier

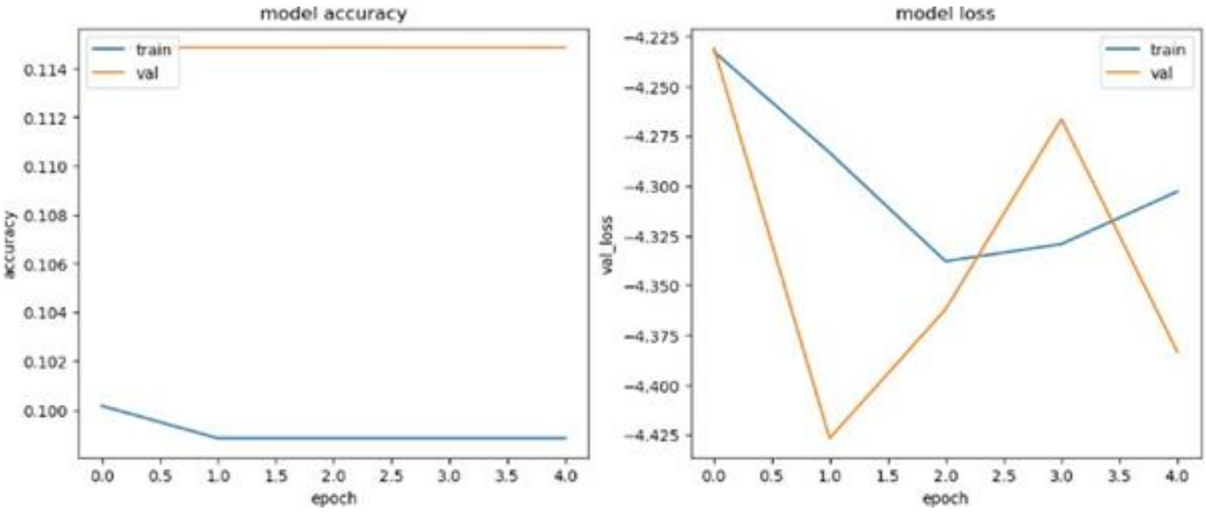


Figure 4: BRNN

The model used here is a Bidirectional Recurrent Neural Network (BRNN) (Figure 4) with a Time Series Data and a SoftMax activation function. The model consists of an Embedding layer, a SpatialDropout1D layer, a bidirectional GRU (Gated Recurrent Unit) layer, a GlobalAveragePooling1D layer and a Dense layer as the output layer.

VI. CONCLUSION

We conducted our analysis by applying the TF-IDF vectorizer on the text in the data, which resulted in vectorized representations of all the words in the corpus. For accurate predictions, high- quality and uniform annotations in the training data are necessary. These include the lack of quality and uniformity of data, limited depth of information in documents, the complexity of applied taxonomies, and the lack of uniformity and correctness in processing and labeling existingdocuments. Furthermore, the presence of low-quality annotated data may corrupt possible training data. We compared our results with a standard support vector machine, achieving over 90% accuracy. By automating labeling and processing, machine learning can help identify patterns and trends in safety incidents, leading to improved safety measures and reduced risk.

VII. FUTURE WORK

Investigate the possibility of using similar methods to visualize conventional deep neural networks decision boundaries. We only use the Voronoi diagram in the D plane in this research and better data dimension reduction methods and visualization methods specifically focus on the hyper spherical plane would be proposed.

VIII. REFERENCES

[1]. T. G. Puranik, N. Rodriguez, and D. N. Mavris, Towards online pre- diction of safety-critical landing metrics in aviation using supervised Art. no. , .

[2]. Airplane Flying Handbook (FAA-H-, -, A), Skyhorse Publishing, New York, NY, USA, ,

[3]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efcient gradient boosting decision tree, in Proc.

[4]. D. Wang, K.-L. Tsui, and Q. Miao, Prognostics and health manage- ment: A review of vibration based bearing and gear health indicators,

[5]. J. Luo, K. R. Pattipati, L. Qiao, and S. Chigusa, Model-based prognostic

[6]. C. Ordez, F. S. Lasheras, J. Roca-Pardias, and F. J. de Cos Juez, A hybrid ARIMASVM model for the study of the remaining useful Jan

[7]. A. Z. Hinchi and M. Tkiouat, Rolling element bearing remaining useful life estimation based on a convolutional long- short-term mem- doi

[8]. A. Z. Hinchi and M. Tkiouat, Rolling element bearing remaining useful life estimation based on a convolutional long- short-term mem- do

[9]. J. Wang, S. Li, Z. An, X. Jiang, W. Qian, and S. Ji, Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis

[10]. L. Bottou, Stochastic gradient descent tricks, in Neural Networks: Tricks , _, .

[11]. D. Hutchison, Evaluation of pooling operations in convolutional architectures for object recognition, in Articial Neural Networks

[12]. S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, Confusion matrix-based feature selection, in Proc. Midwest Artif. Intel. Cognit.

[13]. S. Zelei, T. Dunbing, and Z. Kun, Beat contract network protocol and its application in job shop AGV scheduling, Trans. Nanjing Univ. Aeronaut.

[14]. A. Devarakonda, M. Naumov, and M. Garland, AdaBatch: Adaptive batch sizes for training deep neural Networks

[15]. M. Lin, Q. Chen, and S. Yan, , Online, . Network in network

[16].