Anomaly Detection in Raw Audio Using Extreme Learning Machine

2023 14th International Conference on Information & Communication Technology and System (ICTS) | 979-8-3503-1216-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTS58770.2023.10330852

Ratih Nur Esti Anggraini
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ratih_nea@if.its.ac.id

Ainurrochman
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
a1308.id@gmail.com

Riyanarto Sarno
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract* - Anomaly detection is a crucial problem that has garnered attention across various research areas and application domains. Numerous anomaly detection techniques have been developed, with specific focus on domains such as sound or speech recognition. Sound recognition applications are commonly employed in health system monitoring and control, making them particularly relevant in this study. The objective of this research is to detect anomalies in sound, enabling informed decision-making regarding the handling of such sounds. For this study, the TUT Rare Sound Events 2017 dataset is utilized, comprising 2987 audio files. These files encompass isolated sound events for each target class with the noise background noise from everyday acoustic scenes. The dataset is divided into a training and testing set with the split of 80:20. The Extreme Learning Machine (ELM) method is employed for the learning process. The accuracy and performance of the ELM method are evaluated through calculations. The results reveal that the ELM method demonstrates promising capabilities in detecting anomalies within raw audio data, achieving an accuracy of 93.98%. Notably, the highest accuracy of 92.13% is achieved when detecting the anomaly of baby cry. These findings highlight the effectiveness of the ELM method in anomaly detection within sound data.**

Keywords— Anomaly Detection, Sound Recognition, Deep Learning, Extreme Learning Machine, TUT Rare Sound Events 2017

I. INTRODUCTION

Anomalies can be defined as patterns within data that deviate from the expected notion of normal behavior. These anomalies can arise in the data for various reasons, including the presence of malicious activity [1-3]. Anomaly detection in sound (ADS) has garnered significant attention due to its potential to identify various types of anomalous sounds, including indications of errors, malicious activities, or critical events like gunshots [4]. Timely detection of such anomalies holds the potential to prevent potential issues. Presently, ADS finds application in diverse areas such as audio surveillance and product inspection, enabling enhanced security and quality control measures [5]. A recent study in this area was using deep learning [6,7]: Autoencoder, Variational, and Flow Based Model to calculate the anomaly score. Most of the research which related to anomaly detections mostly focus to the performance.

The primary focus of Anomaly Detection research lies in achieving higher performance levels. In their study, Yuma et al. utilized the TUT Rare Sound Events 2017 dataset to develop an unsupervised model [8]. This model employed an autoencoder for the detection of unknown anomaly sounds. The study successfully demonstrated the model's capability to detect anomaly sounds in real-world environments. The researchers plan to explore supervised approaches in their future work.

The dataset used in the study consists of isolated audio recordings belonging to three classes: baby cry, glass break, and gunshot. These isolated recordings were further mixed with background audio from acoustic scene recordings. This mixture generation technique not only served as data augmentation but also facilitated adjustable parameters such as background audio, event timing, and event-to-background power ratio.

A study [9] noted that Extreme Learning Machine (ELM) could surpass Support Vector Machines (SVM) by around 5%. In addition, it has 10 times faster training time. Agarwal et al. [10] also reported that ELM was able to outperform Neural Network performance by 20% by using Mel-Frequency Cepstrum (MFCC) as feature that has been extracted from the dataset. Therefore, this study purposed to build a model which able to recognize or detect an anomaly from TUT Rare Sound Event 2017 as dataset using supervised approach called Extreme Learning Machine. The building of the model used Python and the objective of the model was to be able to recognize 3 anomaly which contains in the dataset such as gunshot, baby cry, and glass breaking.

II. PREVIOUS WORKS

Previous studies have utilized various approaches to detect anomalies using the TUT 2017 dataset. One notable work by Mesaros et al. [11] developed a baseline system based on a multilayer perceptron (MLP) architecture. The frequency range was covered by 40 mel bands from 0 to 22.050 Hz. The network was trained using the Adam optimizer with a maximum of 200 epochs and a learning rate of 0.001. The MLP model achieved a training accuracy of 72.7% and an evaluation accuracy of 64.1%.

In a study conducted by Sophiya et al. [12], Audio Scene Classification was performed using the TUT dataset 2017. The researchers utilized Extreme Learning Machine (ELM) and Log Mel band features to represent the distinctive characteristics of the input audio scenes. The results demonstrated an accuracy of 75% for the classification task. The ELM model was constructed with two dense layers, each consisting of 50 hidden units per layer. The model underwent training for 200 epochs and employed the sigmoid function along with the Rectified Linear Unit (ReLU) activation function, known for its rectifying properties.

Valenzise et al. [13] reported that audio features such as temporal features such as Zero Crossing Rate (ZCR); energy features such as Short Time Energy (STE); spectral features

such as spectral moments, spectral flatness; perceptual features such as loudness, sharpness, or MFCC was very promising in scream and gunshot detection for audio surveillance system.

III. Research Method

This section explained the research methodology employed. The research scenario, depicted in Fig. 1, involves several key steps, namely feature extraction, normalization, data split, model training using the Extreme Learning Machine (ELM), and evaluation.

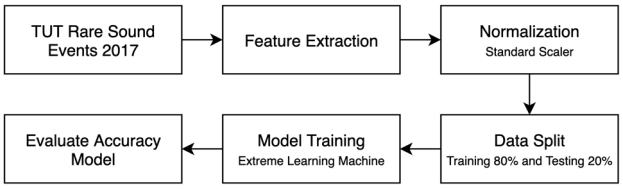

Fig 1. Research Method

A. Dataset

The experiment conducted in this study utilizes the TUT Rare Sound Events 2017 dataset. Each target sound event category is treated independently, resulting in distinct characteristics for each sound category. The dataset incorporates 15 diverse audio scenes, with a portion sourced from the dataset of TUT Acoustic Scenes 2016 as the background noise. This comprehensive dataset comprises of 2987 audio files encompassing anomalies such as Glass breaking, Gunshot, and Baby crying. It is accompanied by precise timing and labeling of the anomaly events. Although each filename denotes the sound group to which it belongs, it does not necessarily indicate the presence of an anomaly within that sound. Therefore, the onset and offset attributes in the dataset play a crucial role as identifiers for each event. To provide an overview, Table 1 presents a selection of sound examples from the dataset.

TABLE 1. TUT Rare Sound Event 2017 dataset

Filename	Onset	Offset	Label
babycry_000.wav	20.46	23.51	babycry
babycry_001.wav	20.76	22.89	babycry
babycry_002.wav	-	-	none
babycry_003.wav	-	-	None
glassbreak_491.wav	14.08	14.46	glassbreak
glassbreak_492.wav	18.71	19.21	glassbreak
glassbreak_493.wav	29.56	29.98	glassbreak
gunshot_003.wav	4.91	6.23	gunshot
gunshot_005.wav	10.97	11.73	gunshot
gunshot_006.wav	-	-	gunshot

B. Feature Extraction

Once all the dataset resources have been gathered, the subsequent step involves extracting sound features from the TUT Rare Sound Events 2017 data. In this study, we utilize several fundamental sound features, including MFCC, pitch, intensity, amplitude, energy, power, and harmonics to noise. These features are extracted specifically at the time of the anomaly event within each sound. For a comprehensive overview, Table 2 presents a list of the features employed in this experiment.

In the field of sound processing, Mel-Frequency Cepstral Coefficients (MFCC) serve as an acoustic model mimicking the functioning of the human ear, offering significant potential for accurate speaker recognition, particularly when a large number of coefficients are utilized [14]. These coefficients collectively constitute the Mel-Frequency Cepstrum (MFC), which represents the short-term power spectrum of a sound through a linear cosine transform applied to a logarithmic power spectrum quantized on a nonlinear mel scale of frequency [15]. Derived from a cepstral representation technique, MFCC enables the extraction of crucial speaker-related features from the audio clip. Figure 2 visually illustrates the distinctive characteristics of MFCC features, providing a graphical representation for enhanced understanding.

TABLE 2.Sound Features

ID	Feature	Sub Feature	Global Static	N
1	MFCC	-	-	20
2	Pitch	Strength	-	1
		Frequency	-	1
3	Intensity	-	- Mean - Std. Dev	2
4	Amplitude	-	-	1
5	Energy	-	-	1
6	Power	-	-	1
7	Harmonics to noise	-	-	1
			Sum	29

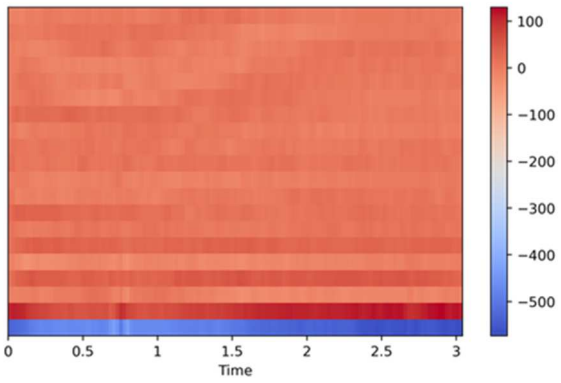

Fig 2. MFCC feature

Pitch refers to the perceptual property of sounds that enables their ordering on a frequency-related scale [16]. In a more common sense, pitch represents the quality of sound that allows us to discern whether a sound is higher or lower, particularly in the context of associated melodies. Specifically, pitch is defined as the frequency of a sine wave that human listeners perceive as matching the target sound [17]. While pitch and frequency are related, they are not equivalent. Frequency is an objective and measurable attribute, whereas pitch is a subjective perception of a sound wave that cannot be directly measured. Figure 3 visually

presents the Pitch features, depicted as blue dots, providing a graphical representation for enhanced understanding.

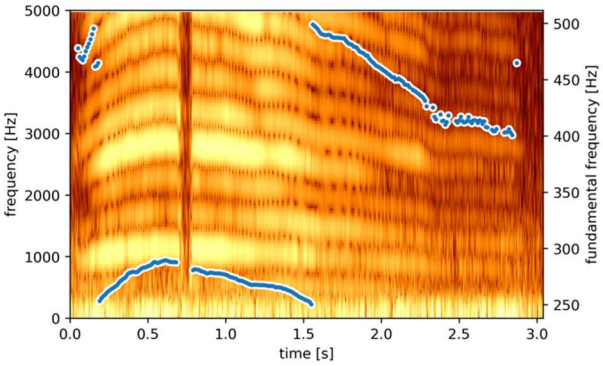

Fig 3. Pitch feature

Sound intensity refers to the amount of power carried by sound waves per unit area in a perpendicular direction. It is quantitatively defined as the temporal average of the product between sound pressure and acoustic particle velocity [18]. Sound intensity can be determined by either direct measurements using a sound intensity probe, which comprises a microphone and a particle velocity sensor, or through indirect estimation using a pressure-pressure (p-p) probe. The p-p probe approximates the particle velocity by integrating the pressure gradient between two closely spaced microphones [19]. Figure 4 illustrates the representation of intensity features, depicted as a blue line in the visualization.

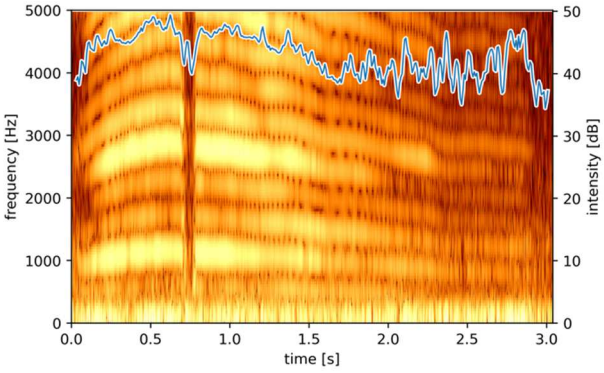

Fig 4. Intensity feature

The Harmonics to Noise parameter serves as a valuable indicator of the relationship between the harmonic and noise components, providing insights into the voice characteristics of a speech signal. This parameterization involves quantifying the relationship between the periodic and aperiodic components, typically expressed in decibels (dB) [20, 21]. The Harmonics to Noise value of a signal varies due to differences in vocal tract configurations, resulting in distinct amplitudes for the harmonics. This parameter relates to the energy carried by the voice signal through the glottal impulses and the energy of the glottic noise fraction once it propagates through the vocal tract. The noise component originates from turbulence generated during phonation as the airflow passes through the glottis [22]. The visualization of Harmonics to Noise features can be observed in Figure 5, while Figure 6 depicts the amplitude features.

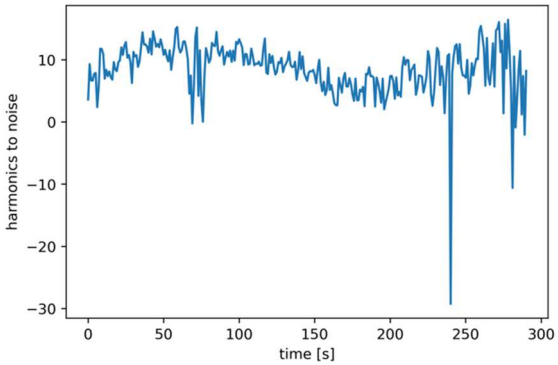

Fig 5. Harmonics to Noise feature

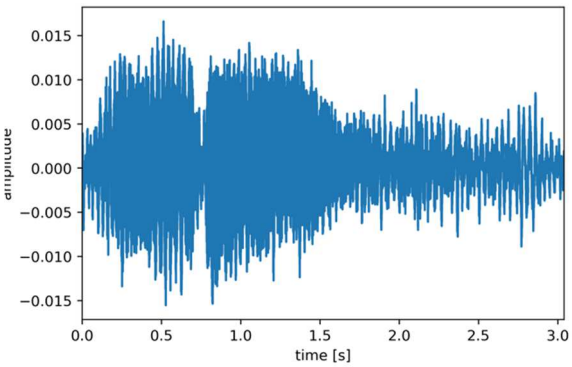

Fig 6. Amplitude feature

C. Normalization and Data Split

The Standard Scaler [23] is applied to normalize each feature by first subtracting the mean and then scaling it to achieve unit variance. By dividing all values by the standard deviation, the feature is scaled to have a variance of 1. The Standard Scaler makes the distribution mean 0. About 68% of the values will fall between -1 and 1. These numbers still depend on the amount and variability of the data.

Regression-type algorithms also benefit from normally distributed data with smaller data sample sizes. Standard Scaler distorts the relative distance between feature values, thereby minimizing imbalanced sample data. In TUT Rare Sound Events 2017, the sample data is considered imbalanced due to high variability, namely the anomaly in each sound clip can be empty.

Then, the data split process is carried out using the sklearn library. In this process, we split the data into two parts, namely data for training the machine (training data) and data for testing the machine (testing data), with a scale of 80% and 20%, respectively. Table 3 shows the details of the dataset that we used for this study.

TABLE 3. TUT Rare Sound Events 2017 dataset

Label	Training	Testing	N
Babycry	401	89	490
Gunshot	382	118	500
Glassbreak	406	94	500
None	1200	297	1497
Sum	**2389**	**598**	**2987**

D. Model of Extreme Learning Machine

The Extreme Learning Machine (ELM) is a type of feedforward neural network that can be utilized for tasks such as classification, regression, clustering, feature learning, and compression. It employs a single layer or multiple layers of hidden nodes to perform these operations. The hidden nodes can be randomly assigned and never updated. An according to their creators Gu et al. [24] ELM models is able to produce good generalization and learn thousands of times faster than networks trained with backpropagation. ELM has been advocated as a solution to address the challenges of slow training speed and overfitting. Due to its superior generalization ability, robustness, controllability, and rapid learning rate, ELM has found applications in diverse fields and domains [25].

Unlike traditional Neural Network methods that require various hidden neurons for efficient training, ELM demonstrates that the system can be trained without such iterative design. Notably, Aarushi et al. [26] have highlighted that the ELM classification algorithm exhibits remarkably fast training and attains commendable generalization performance, particularly when using infinite and different activation functions in the hidden layers. Table 4 provides a log summary of the ELM model's performance.

TABLE 4. Log Summary of ELM

Layer	Output Shape	Param
Dense	(None, 128)	4,224
Dense_1	(None, 128)	16,512
Dense_2	(None, 128)	16,512
Dense_3	(None, 128)	16,512
Dropout_3	(None, 128)	0
Dense_4	(None, 3)	516
Total params: 54,276 Trainable params; 50,052 Non-trainable params: 4,224		

The model in this experiment was created using a well-known library for machine learning and artificial intelligence, namely Keras. There are five dense layers, and a value of 0.3 as the dropout value. Each hidden layer, including dense, has a unit size of 128. Then, the ELM model was "trained" using the training data and the test data was then used as validation to calculate the model's accuracy value. This step is called the fitting process.

IV. Experiment and Results

This section describes the results of the experimental design, namely the value of anomaly prediction accuracy on ELM. In the fitting process, the feature that has been extracted and normalized in training data has been passed through the model as a parameter of the fitting process to train the ELM model. The training accuracy of the model in the fitting process shows in Figure 7 and the training loss shows in Figure 8.

The fitting process was performed with 150 epochs with 10 batch size. ELM model was able to get the highest training accuracy with the value of 98.12% and the highest testing accuracy in fitting process is 94.31%. was able to perform the classification process by having the overall accuracy of 93.98% which represent in detail on Figure 9.

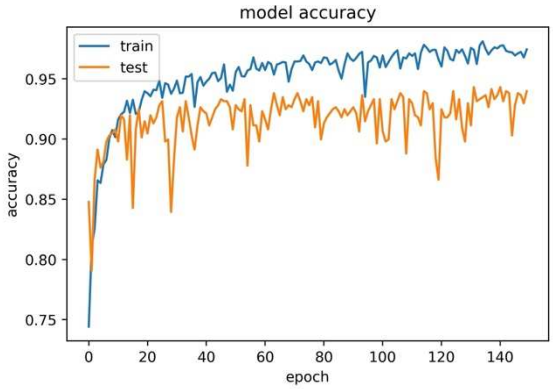

Fig 7. Model accuracy ELM

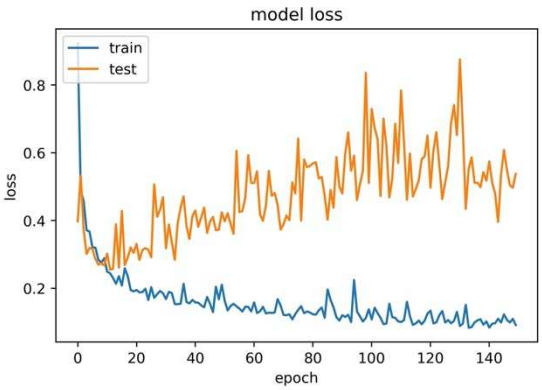

Fig 8. Model loss ELM

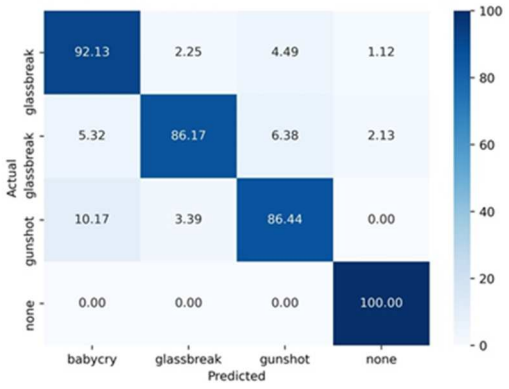

Fig 9. Confusion Matrix of the noise detection

Figure 8. indicates the accuracy of ELM model in recognizing the anomaly in the dataset. The type of anomaly is determined based on the highest probability value is assumed to be the variable *y_pred*. After getting the *y_pred* value, the *y_pred* value is compared with the original type of anomaly (in the *y_test* variable). The highest accuracy performance of the ELM model was on recognizing babycry anomaly with the value of 92.13% and 93.98% as the overall the accuracy. The ELM model was also able to predict the sound that doesn't have an anomaly with the accuracy of 100% which become an exception since the feature extraction was performed in the

anomaly event duration which makes the value of its feature as 0.

V. CONCLUSIONS

In this experiment, we carried out the process of identifying anomaly detection using the TUT Rare Sound Event 2017 dataset. The dataset contains an isolated sound event with records of everyday acoustic scene as background, but some of datapoint (sound file) does not contains an anomaly even it has been labelled as anomaly. ELM give sufficient results in emotion detection. The results of the training on ELM gave a figure of 98.12%, while the result of evaluation or testing gave a figure of 93.98% with the highest anomaly detection accuracy in baby cry anomaly as 92.13% and the lowest anomaly detection accuracy in glass break anomaly as 92.13%.

As a recommendation for further research, we chose Data Augmentation method as a step that must be executed, to avoid overfitting or underfitting such as Generate Adversarial Networks (GAN). The Data Augmentation method was also chosen to overcome the high variability of the data.

REFERENCES

[1] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, "Anomaly detection," *Computers, Materials and Continua*, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.

[2] P. Transfeld, S. Receveur, and T. Fingscheidt, "An acoustic event detection framework and evaluation metric for surveillance in cars," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, vol. 2015-January, pp. 2927–2931. doi: 10.21437/interspeech.2015-452.

[3] J. Baumann, T. Lohrenz, J. Franzen, and T. Fingscheidt, "Baby Cry Recognition in Vehicles." [Online]. Available: https://www.researchgate.net/publication/340338872

[4] J. Baumann, T. Lohrenz, A. Roy, and T. Fingscheidt, *Beyond The Dcase 2017 Challenge On Rare Sound Event Detection: A Proposal For A More Realistic Training And Test Framework*. IEEE, 2020.

[5] O. I. Provotar, Y. M. Linder, and M. M. Veres, "Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders," in *2019 IEEE International Conference on Advanced Trends in Information Theory, ATIT 2019 - Proceedings*, Dec. 2019, pp. 513–517. doi: 10.1109/ATIT49449.2019.9030505.

[6] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 1996–2000.

[7] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 313–317.

[8] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman-Pearson Lemma," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 1, pp. 212–224, Jan. 2019, doi: 10.1109/TASLP.2018.2877258.

[9] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," 2014.

[10] A. Agarwal, S. Chandrayan, and S. S. Sahu, "Prediction of Parkinson's disease using speech signal with Extreme Learning Machine," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 3776–3779.

[11] A. Mesaros *et al.*, "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," 2017. [Online]. Available: https://hal.inria.fr/hal-01627981

[12] E. Sophiya and S. Jothilakshmi, "Large scale data based audio scene classification," *International Journal of Speech Technology*, vol. 21, no. 4, pp. 825–836, Dec. 2018, doi: 10.1007/s10772-018-9552-3.

[13] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Anonacci, and A. Sarti, *Advanced Video and Signal Based Surveillance, 2007, AVSS 2007, IEEE Conference on : date, 5-7 Sept. 2007*. IEEE, 2007.

[14] Z. Wanli and L. Guoxin, *The Research of Feature Extraction Based on MFCC for Speaker Recognition*.

[15] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "HMM-Based Audio Keyword Generation."

[16] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2007.

[17] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *The Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3491–3502, 1996.

[18] F. Fahy, *Sound intensity*. CRC Press, 2017.

[19] F. Jacobsen and H.-E. de Bree, "A comparison of two different sound intensity measurement principles," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1510–1517, 2005.

[20] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, and J. P. Teixeira, "Harmonic to noise ratio measurement - Selection of window and length," in *Procedia Computer Science*, 2018, vol. 138, pp. 280–285. doi: 10.1016/j.procs.2018.10.040.

[21] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, 1993, vol. 17, no. 1193, pp. 97–110.

[22] H. Lutfiyya *et al.*, *Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection*.

[23] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006, doi: https://doi.org/10.1016/j.neucom.2005.12.126.

[24] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2011.

[25] G.-B. Huang, "What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle," *Cognitive Computation*, vol. 7, no. 3, pp. 263–278, 2015.

[26] A. Agarwal, S. Chandrayan, and S. S. Sahu, "Prediction of Parkinson's disease using speech signal with Extreme Learning Machine," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 3776–3779. doi: 10.1109/ICEEOT.2016.7755419.