JOURNAL OF BASIC SCIENCE AND ENGINEERING

DETECTION OF EMERGENCY SITUATION USING DEEP LEARNING SOUND RECOGNITION METHODS

Avijit Bose[1], Bidisa Chowdhury[2], Rajdeep Roy[3], Ishika Ghosh[4], Fazal Hussain[5], Ahana Mukherjee[6], Satyajit Chakrabarti[7]

[1,2,3,4,5,6,7,] Institute of Engineering & Management, Kolkata, India, University of Engineering & Management, Kolkata, India.

DOI: https://doie.org/10.10399/JBSE.2025141993

Abstract: Machine learning (ML) and deep learning (DL) technologies enable sound recognition systems to perform their functions.The detection of anomalous events across different environments depends critically on sound recognition technology which is especially important for security and household safety. Emergency detection within home environments demands special attention for elderly residents.The safety of elderly individuals and children remains a crucial issue as they may struggle to communicate in emergencies. Automated systems capable of real-time detection that recognize distress signals and accidents enable effective monitoring of household audio data.The proposed system monitors household audio to identify life-threatening situations through detection of incidents such as screaming or shattering noises or gunshots. Advancements in deep learning, particularly CNNs have created the capability to achieve highly precise recognition of sound patterns. By examining audio patterns researchers can differentiate between everyday household sounds and signals that indicate potential danger. Our research introduces a deep learning model which tackles the problem of recognizing emergency sounds within indoor spaces. The research offers a sound recognition model based on deep learning that monitors occupant safety and detects potential emergencies. Experiments are conducted using audio data originated from both real-world environments and online[18] sources. Our model achieved an accuracy of 90%, with promising precision and recall rates. The[27] proposed system shows potential to improve safety by generating timely alerts establishes a foundation for future research on sound recognition integration.

Keywords: Deep learning (DL), Anomalous event detection, Household safety, Emergency detection, Audio data monitoring, Convolutional Neural Networks (CNNs), Mel Spectrogram, Audio pattern analysis, Indoor environment monitoring, Hazardous sound classification

1. Introduction

There is still a major problem of keeping people safe in indoor environments, in particular for vulnerable groups including the elderly and children. Speeches, glass breaking, and gunshots are some sudden emergency events that can determine situations where the life is in danger and which need immediate action. Conventional surveillance systems work based on visual indications, which may not work in conditions with low visibility or arise on occasions when perilous events happen outside of the field of view of the camera[25]. While video data may be cumbersome and heavily overloaded, sound carries rich context information, and thus sound can serve as a prime modality for emergency detection within smart environments. The exploration of sound recognition as a real-time emergency monitor (REM) technology has grown rapidly with the advancement of deep learning and the extension[22] of Internet of Things (IoT) devices. Nevertheless, there are certain challenges associated with building a strong sound classification model: emergency sounds tend to be scarce, greatly varied, and can often resemble non-emergency events. In this case, a dropped object may mimic a shattering sound, whereas loud background noises may prevent accurate classification. To manage these challenges, a model must be able to learn distinct acoustic properties and generalize well in a range of contexts. A deep learning-based approach for detecting[19] emergency sound events in home environments is proposed in this paper. The system is trained and evaluated using a set of real-world and online audio recordings, including high-risk sounds like shouting, glass breaking, and gunshots. Class-weighting processes and data augmentation techniques are employed[23] to balance the dataset and improve performance for rare events. The proposed method aims to accurately distinguish emergency sounds from normal activities to enable proactive responses and

timely warnings.

2. Related Work

2.1. Different Deep Learning Approaches

Anomalous sound event detection has potential application in indoor security surveillance. Mnasri et al. highlight how audio modality can be more effective for emergency event detection compared to simple video surveillance [10] whereas Borna S et al. suggests how Artificial Intelligence and Machine learning can be useful and efficient tools to extract speech-specific features from the recorded voice[17], to recognize pain [5]. Mnasri et al. performed a survey which shows how deep learning techniques perform better to produce state of the art results compared to traditional approaches like Gaussian Mixture Model(GMM) and Hidden Markov Model(HMM) [10]. Zhao et al. demonstrates the use of CNN to detect abnormal sound events using noiseless abnormal sound dataset and proposed new mixed feature image to increase training speed and accuracy of results [2]. While Salamon et al. also suggests deep CNN architecture for environmental sound classification [14], Mustaqeem et al. proposed a Deep Stride Convolutional Neural Network(DSCNN)[17] model inspired by the concepts of plain nets[17],mostly used for computer vision,image classification etc to achieve high accuracy [7].The DSCNN model has less number of convolution layers and small kernel size which reduces computational complexity.The model uses adaptive threshold-based preprocessing to eliminate the noise and unnecessary parts from the[17] speech signal. A CNN method for audio-based emergency identification during myocardial infarction was also proposed by Mohan et al.[6].Their work demonstrates the approach to use Deep learning to detect audio-based distress signals.

2.2 Classification of Audio data

Data classification in the context of emergency event detection using sound recognition involves the process of assigning predefined labels to audio samples, indicating the presence or absence of specific emergency events. In order to identify emergency situations in single-person households, Kim et al. concentrated on creating a deep learning-based sound recognition mode[17][1].They divided the emergency sound events into two categories-Primary events included ‘Impact Sound’, ‘Glass-breaking’, ‘Explosion’, ‘Burning’; Secondary events included ‘Screaming’ and ‘Vomiting’. Heittola et al. proposed a system which utilizes context information like humans do to predict different sound events [13].Pasquale Foggia et al. suggested a method of audio detection with high accuracy,adapted to deal with high noise level and loud background sounds. ‘Scream’, ‘Glass-breaking’, ‘Gun-shots’ -these three different classes of audio are detected using ‘Bag of Words’ approach [15]. Yasin et al. categorized an infant's cries according to a variety of factors, such as fatigue, anxiety, difficulty finding their parent, blood loss, pain in the abdomen, etc. [3].They proposed to classify the reason of toddler’s cry with the help of the different mouth noises a toddler makes and the acoustic information of infant screaming. Infant cry was regarded by Mudiyanselage et al. as a binary classification comprising cry and not-cry categories. [16] [4]. It has been observed that neural network-based methods operate well in hygienic and limited settings. However, classifiers are sensitive at the boundary and are easily confused and overlapped with noise signals in noisy situations and with insufficient training data.[16] Chang et al. proposed a new method that uses acoustic feature engineering and variable selection to classify cries into three groups: hunger, sleep, and discomfort [8].

2.3 Data Augmentation

Data augmentation is one of the most important tools to reduce the scarcity of labled data.Wei et al. drew a comparative analysis among different data augmentation techniques for audio classification [9].They also proposed a new simple and effective data augmentation method named “Mixed Frequency Masking”.The researchers compared various data augmentation methods like “Time stretch”, “Pitch shift”, “Add nose”, “Cut-out”, “Mix-up”, “Spec-augment”, “Spec-mix” etc along with the proposed method and evaluated their effectiveness.They experimentally proved that usage of data augmentation can be beneficial for audio classification using spectrogram. Additionally, in their work, Salamon et al. applied data augmentation techniques such as time stretching, pitch shifting, introducing background noise, etc.[14].

2.4 Feature Extraction

A key role is played by feature extraction in the effectiveness of sound recognition models for emergency event detection in household environments, especially when focusing on the safety of elderly individuals and children. The extraction of meaningful and relevant features from audio[21] data is essential for enhancing the discriminative power of deep learning models. Afendi et al. selected log-mel energies for convolutional recurrent neural network (CRNN) with long short term memory (LSTM), whereas Zhao et al. retrieved audio features as MFCC for use as input in their CNN model [2] [12]. However, Abbasi et al. offered a method that utilizes the best feature extraction technique by removing MFCCs from the audio data and then uses principal component analysis (PCA) to choose the fewest best-performing features for optimal performance. [11].

3. Methodology

This paper proposes a system that automatically identifies and distinguishes sound events deviating significantly from expected acoustic patterns. The core challenge is to separate normal day to day sounds from anomalies, and this task is often complicated by the complexity of real world audio environments, and the scarcity of data especially for anomalous events. Here, the main aim is to build deep learning model capable of identifying common sound features and detecting deviations, this and many times requires working on a real-time base. Successfully demarcating "anomaly" in the specific domain and addressing challenges such as noisiness, environmental variability, and the need for strong generalization are all crucial for success.

3.1 Data Collections

Collecting a balanced and varied dataset is crucial for building a strong and reliable sound recognition system for identifying emergency situations. To make sure the model is exposed to a broad range of sounds that exist in the real-world, this study combines publicly available audio datasets with specifically collected recordings. Both emergency and non-emergency sounds are included in the dataset, which shows common scenarios in smart personal home (SPH) environments. Audio samples are from public online repositories and real-world environments.
The dataset is divided into four main categories:
● Screams: A sign of agony, fear, or peril.
● Shattering sounds: An indication of glass breakage or falling objects.
● Gunshots: A sign of a possibly deadly circumstance.
● Non-emergency sounds: Common household noises like clinking utensils background chatter, or doors closing.

3.1.1 Data Preprocessing and Augmentation:

Data augmentation techniques were implemented to increase the dataset's diversity because of the imbalance in sound event occurrences (e.g., fewer samples of gunshots or shattering sounds).
● Time stretching: Slowing down or speeding up the audio without altering its pitch.
● Time shifting: Randomly shifting the audio waveform forward or backward in time to simulate variations in sound occurrence.
● Adding background noise: Overlaying low-level noise to make the model more resilient to real-world conditions.

3.1.2 Dataset Splitting:

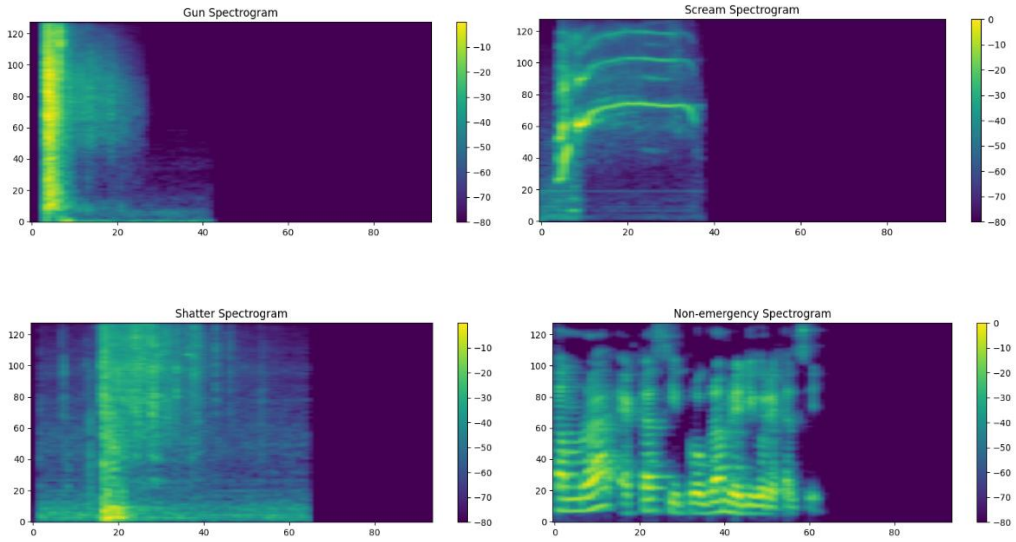
To avoid overfitting and guarantee generalizability, the dataset is separated into training, validation, and test sets. By carefully curating and augmenting the dataset, this study ensures the developed model can accurately classify emergency events, even in noisy or complex environments.

3.2 Feature Extraction

Mel spectrograms are the input features used in this study to transform raw audio into a format that can be used for deep learning. A Mel spectrogram is close to human hearing perception, it shows the intensity of different frequencies across time, mapped to the Mel scale. A 2048-point Fast Fourier Transform (FFT)

with 512 samples hop is used to process the audio signals to extract fine-grained frequency information. The frequency space is compressed while maintaining the auditory characteristics using 128 Mel filter banks. The spectrogram is further converted to a logarithmic scale

to highlight subtle sound variations, making it easier for the model to discern quiet emergency events (e.g., a distant glass shatter) from louder ambient noises. The feature 6 extraction stage is important because it converts the audio waveform into visual pattern that the convolutional neural network can learn to distinguish between emergency and non-emergency sounds.



3.3 Model Architecture

Because of its hierarchical interpretation of the data, the Convolutional Neural Network (CNN)[28], an efficient deep learning model[28], is used in the system to handle spectrograms. Due to the fact that frequency components are represented in one direction and time components in another, the spectrograms are recorded as grayscale images during feature extraction. This method enables the CNN to extract both spectral and temporal characteristics that are imperative for differentiating between sophisticated auditory occurrences.

The model consists of a series[17] of convolutional layers followed by batch normalization and ReLU activation[17] functions. The convolutional layers use 3x3 kernels to distil the local features of[17] spectrograms like frequency peaks, sound patterns. Batch Normalization makes sure that learning is stable and then using ReLU brings in non-linearity in the model which is used to learn complex patterns. Max pooling layers are added after each convolutional block to downsample feature maps, reducing dimensionality and highlighting the most prominent features, thus preventing overfitting and speeding up training.

The feature maps are flattened and sent via fully connected layers after the convolutional blocks. These layers act as high-level feature classifiers by learning complex combinations of the lower-level audio properties that the convolutional layers have detected. Dropout layers, which randomly deactivate a fraction of neurons during training, are introduced to prevent overfitting and guarantee that the model generalizes to unseen audio samples.

The last output layer is a softmax activation function,which creates a probability distribution for each of the four possible classes: gunshot, yelling, shattering, and non-emergency. The class with the highest probability is[17] used to[17] select the predicted label[26].

The model is optimized using the Adam optimizer and compiled with the categorical cross-entropy loss function, which is appropriate for multi-class classification applications. The model is trained using a weighted loss approach to address class imbalances, ensuring that less frequent emergency events (like shattering glass) receive sufficient attention during training.

Such an architecture strikes a balance between the depth of the model and computational efficiency, facilitates the learning of subtle acoustic patterns without incurring overfitting, and leads to a high-performing solution to differentiate emergency related sound events from the daily background noise.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 126, 32)	320
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0
dropout (Dropout)	(None, 63, 63, 32)	0
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0
dropout_1 (Dropout)	(None, 30, 30, 64)	0
flatten (Flatten)	(None, 57600)	0
dense (Dense)	(None, 128)	7,372,928
dropout_2 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 4)	516

Total params: 7,392,260 (28.20 MB)
Trainable params: 7,392,260 (28.20 MB)
Non-trainable params: 0 (0.00 B)

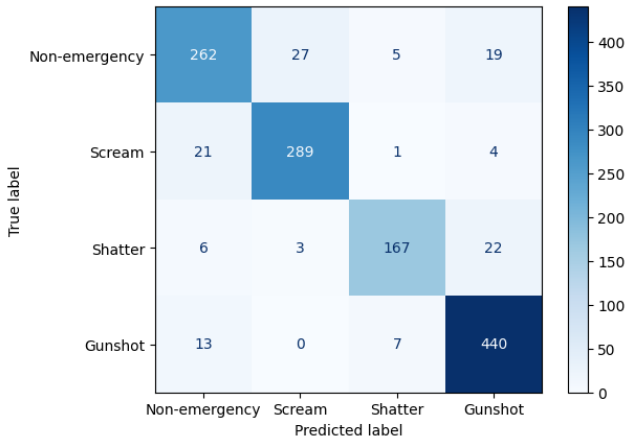
Vol. 22, No. 1, (2025)
ISSN: 1005-0930

4. Experimental Results and Analysis

Based on the trial results, the global accuracy of[24] the model was 90%, and class-specific precision and recall values were very reliable. As there were fewer samples in the dataset, the model slightly underperformed on shattering noises, but performed very well on gunshots (93% F1-score) and crying (91% F1-score).

	precision	recall	f1-score	support
Non-emergency	0.87	0.84	0.85	313
Scream	0.91	0.92	0.91	315
Shatter	0.93	0.84	0.88	198
Gunshot	0.91	0.96	0.93	460
accuracy			0.90	1286
macro avg	0.90	0.89	0.89	1286
weighted avg	0.90	0.90	0.90	1286

We designed a confusion matrix to evaluate misclassifications and determine which areas need improvement. The model showed a strong recall for emergency sounds, identifying most important events accurately. Yet some sounds were incorrectly classified as breaking sounds and non-emergency sounds, likely because they have similar spectral profiles.



The use of Mel spectrograms as input features significantly enhanced classification performance. The model was better able to distinguish between different sound events by capturing both temporal and spectral information through spectrogram-based feature extraction.

The CNN's ability to recognize frequency variations helped distinguish sharp transient sounds like gunshots from more sustained sounds like screams. Since the dataset had an imbalance (fewer shattering and gunshot samples compared to non-emergency sounds), class weighting was applied during training. This technique helped mitigate bias toward the majority class and improved recall for less frequent emergency events.

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

5. Conclusion and Future Work

In this study, we suggested a system based on deep learning[28] for anomalous sound event detection, demonstrating real-time capability and robustness to noise. We addressed the challenges of environmental variability through novel model architectures. Our results indicate the potential of this approach for industrial monitoring and healthcare. However, limitations such as generalization to unseen environments and explainability remain. Future work should focus on improving model robustness, exploring multimodal data fusion, addressing data scarcity through synthetic data generation. Ultimately, this project contributes to the advancement of Anomalous Sound Event Detection, paving the way for more reliable and efficient systems that can enhance safety and efficiency in various real-world applications. Future research can enhance the proposed anomalous sound event detection model by expanding dataset diversity to include more real-world emergency recordings, improving generalization across environments. Real-time deployment for real-time alerts will be enabled through integration with Internet of Things-based home security systems and smart home automation devices. With multi-modal approaches, such as sound along with motion sensors, thermal imaging or video, robustness may be enhanced and false alarms minimized. Adaptive noise removal methods, i.e., specialized denoising autoencoders, can allow better recognition within noisy environments. Model optimization towards real-time low-latency inference using quantization and pruning will facilitate deployment feasibility. Personalized audio recognition can be investigated to better fit into home environments, enhancing distinction between mundane activities and actual emergencies. Furthermore, integrating explainable AI (XAI) methods can boost transparency and user confidence. The model can also be extended to other areas, including healthcare monitoring, industrial safety, and city surveillance, with cross-domain transfer learning for further applications.

References

1. Jinwoo Kim, Kyungjun Min, Minhyuk Jung, Seokho Chi,Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition,Building and Environment,Volume 181,2020,107092,ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2020.107092

2. Jie Zhao Anomalous Sound Detection Based on Convolutional Neural Network and Mixed Features 2020 J. 10.1088/1742-6596/1621/1/012025 Phys.: Conf. Ser. 1621 012025 ,DOI https://iopscience.iop.org/article/10.1088/1742-6596/1621/1/012025/meta

3. Yasin, Sana & Draz, Umar & Ali, Tariq & Shahid, Kashaf & Abid, Amna & Irfan, Muhammad & Huneif, Mohammed & Almedhesh, Sultan & Alqahtani, Seham & Abdulwahab, Alqahtani & Alzahrani, Mohammed & Alshehri, Dhafer & Abdullah, Alshehri & Rahman, Saifur. (2022). Automated Speech Recognition System to Detect Babies' Feelings through Feature Analysis. Computers, Materials and Continua. 73. 4349-4367.10.32604/cmc.2022.028251. 10 https://www.researchgate.net/publication/362062523_Automated_Speech_Recognition_System_to_Detect_Babies'_Feelings_through_Feature_Analysis J

3. Ji, C., Mudiyanselage, T.B., Gao, Y. et al. A review of infant cry analysis and classification. AUDIO SPEECH MUSIC PROC. 2021, 8 (2021). https://doi.org/10.1186/s13636-021-00197-5

5. Borna S, Haider CR, Maita KC, Torres RA, Avila FR, Garcia JP, De Sario Velasquez GD, McLeod CJ, Bruce CJ, Carter RE, Forte AJ. A Review of Voice-Based Pain Detection in Adults Using Artificial Intelligence. Bioengineering (Basel). 2023 Apr 21;10(4):500. doi: 10.3390/bioengineering10040500. PMID: 37106687; PMCID: PMC10135816. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10135816/

6. H. M. Mohan and S. Anitha, "Real Time Audio-Based Distress Signal Detection as Vital Signs of Myocardial Infarction Using Convolutional Neural Networks," Journal of Advances in Information Technology, Vol. 13, No. 2, pp. 106-116, April 2022.http://www.jait.us/index.php?m=content&c=index&a=show&catid=211&id=1203

7. Mustaqeem, and Soonil Kwon. 2020. "A CNN-Assisted Enhanced Audio Signal Processing for

Speech Emotion Recognition" Sensors 20, no. 1: 183.
https://doi.org/10.3390/s20010183

8. Chuan-Yu Chang, Sweta Bhattacharya, P. M. Durai Raj Vincent, Kuruva Lakshmanna, Kathiravan Srinivasan, "An Efficient Classification of Neonates Cry Using Extreme Gradient Boosting-Assisted Grouped-Support-Vector Network", Journal of Healthcare Engineering, vol. 2021, Article ID 7517313, 14 pages, 2021.
https://doi.org/10.1155/2021/7517313

9. Shengyun Wei et al A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification 2020 J. Phys.: Conf. Ser. 1453 012085 DOI 10.1088/1742-6596/1453/1/012085
https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085

10. Mnasri, Zied & Rovetta, Stefano & Masulli, Francesco & Cabri, Alberto. (2022). Anomalous sound event detection: A survey of machine learning based methods and applications. https://www.researchgate.net/publication/357574500_Anomalous_sound_event_detection_A_survey_of_machine_learning_based_methods_and_applications

11. A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska and U. Tariq, "A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics," in

IEEE Access, vol. 10, pp. 38885-38894, 2022, doi: 10.1109/ACCESS.2022.3166602.
https://ieeexplore.ieee.org/abstract/document/9755147/authors#authors

12. Amirul Sadikin Md Afendi , Marina Yusoff Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia 2Advanced Analytic Engineering Center (AAEC), Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.
http://download.garuda.kemdikbud.go.id/article.php?article=1494433&val=151&title=Review%20of%20anomalous%20sound%20event%20detection%20approaches 11

13. Heittola, T., Mesaros, A., Eronen, A. et al. Context-dependent sound event detection. J AUDIO SPEECH MUSIC PROC. 2013, 1 (2013).
https://doi.org/10.1186/1687-4722-2013-1

14. J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.
https://ieeexplore.ieee.org/document/7829341

15. Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, Mario Vento,Reliable detection of audio events in highly noisy environments,Pattern Recognition Letters,Volume 65,2015,Pages
https://doi.org/10.1016/j.patrec.2015.06.026

16. https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00197-5

17. www.mdpi.com

18. www.researchgate.net

19. https://dblp.dagstuhl.de/pid/46/113.bib?param=1

20. https://dblp.dagstuhl.de/pid/46/113.bib?param=1

21. Speech Emotion Recognition using feature fusion: a hybrid approach to deep learning
https://doi.org/10.1007/s11042-024-18316-7

22. University of Whales, Bangor on 2024-05-30

23.Discriminative importance weighting of augmented training data for acoustic model training
https://doi.org/10.1109/ICASSP.2017.7953085

24. Heritage, Culture and Society

https://doi.org/10.1201/9781315386980

25. https://pure.rug.nl/ws/files/32628117/Complete_thesis.pdf
26. https://link.springer.com/article/10.1007/s11042-023-17964-5?code=87b7701a-0f37-4824- a6a9-
27. University of Zulululand on 2024-11-11
28. https://www.raghuenggcollege.com/documents/AQAR/criteria-3/3.4.4-Proofs-2022-23.pdf