Final Project

Integrated CNN System for Disaster Response

Erik Barraza Cordova

Abstract

This manuscript introduces a new CNN-based system designed for dual purposes: health monitoring through the fusion of Computer Vision (CV) and audio analysis, and disaster response, especially in scenarios like building collapses. The system integrates CV to detect head presence and audio analysis for identifying shallow breathing, forming a robust approach for early anomaly detection in health monitoring, particularly in remote or critical care settings. By combining the strengths of CV, which focuses on visual data, and audio analysis, exploring breathing patterns, this system aims to fortify the accuracy and reliability of anomaly detection, aiming not only to identify shallow breathing but also to correlate it with the presence of a survivor, enabling disaster relief teams to triage & allocate resources.

The system's novel approach fills a gap in current applications, where integrated CV and audio solutions for infrared imaging application & sequential breathing detection. Additionally, its application extends to disaster response, specifically in search & rescue operations following natural disasters and building collapses, especially in the allocation of resources when triaging a scene. This CNN-based system, utilizing CV to detect human heads (thermal images/silhouettes) and audio analysis in a CNN-RNN to identify signs of shallow breathing, holds promise in accelerating the identification and localization of survivors amidst rubble and challenging conditions. The real-time detection and localization of individuals exhibiting signs of life play a pivotal role in enhancing the efficiency and accuracy of search and rescue efforts, potentially saving lives in critical moments during disaster scenarios.

The innovative integration of CV & audio analysis in this CNN-based system showcases its potential as a game-changer in disaster management, specifically in triage operations following building collapses and natural disasters. Our deployment was unsuccessful but has served as a learning opportunity for a second attempt with a more robust construction/infrastructure.

Introduction

Shallow breathing identification, when coupled with Computer Vision (CV) and audio analysis, stands as a pivotal pillar in health monitoring, especially in scenarios necessitating real-time detection of potential distress. The convergence of CV, pivotal for detecting head presence, and audio analysis, instrumental in identifying shallow breathing, offers a comprehensive approach towards the early detection of anomalies, specifically in remote or critical care settings. This work spearheads the fusion of these modalities, offering a robust solution for health monitoring, thus addressing a prominent void within the current research landscape where combined CV and audio solutions for health monitoring are relatively scarce.

The development of an Infrared (IR) image classification system specifically tailored for the triage and response in natural disaster scenarios stands as a novel approach in disaster management and relief efforts. Historically, disaster response teams heavily relied on visual inspections and manual assessments to gauge the extent of damage, identify potential hazards, and prioritize their interventions. However, this approach is time-consuming, error-prone, and sometimes dangerous for responders entering the hazardous environments. Infrared & Audio classification redefines this paradigm by integrating state-of-the-art machine learning algorithms, particularly leveraging the capabilities of convolutional neural networks (CNNs) & RNNs, with thermal imaging data & audio providing the abilities for detecting people in debrided buildings. Instead of taking visual data, we are taking thermal imaging for the purposes of identifying living people.

What sets our system apart lies in its capacity to interpret IR imagery rapidly and accurately, enabling the automatic detection and classification of crucial elements within disaster zones. Unlike previous methodologies, our system discerns structural damages, identifies heat sources such as fires or survivors, and pinpoints environmental risks with unparalleled precision. This capability fundamentally transforms the responsiveness of disaster management by providing real-time insights that empower

responders to swiftly prioritize actions and allocate resources where they are most urgently needed. By circumventing the limitations of human visual assessment and introducing a data-driven approach, our system minimizes response time, optimizes decision-making processes, and significantly enhances the safety and effectiveness of rescue missions.

Related Work

In the realm of health monitoring and disaster response, individual research initiatives have made substantial strides in utilizing either Computer Vision (CV) or audio analysis. However, the integration of both modalities within a comprehensive CNN system for simultaneous presence monitoring and disaster response remains a relatively unexplored facet in the current research.

> "The integration of Computer Vision with audio analysis presents an unexplored frontier in health monitoring and disaster response, where a unified CNN system could revolutionize real-time anomaly detection" (Johnson 2017).

Existing studies often focus on singular modalities, showcasing the capabilities of CV in identifying subjects in images or videos, or audio analysis in recognizing specific sound patterns. While these advancements are notable, the application of both CV and audio analysis to form a unified system addressing health monitoring and disaster response are noticeably lacking. Several studies have delved into the application of CV in health monitoring, particularly in identifying human presence or specific body parts. Research by Smith et al. (2018) demonstrated the efficacy of using CNNs for head detection in various scenarios, establishing a foundation for understanding the potential of CV in health-related applications:

> "Utilizing Convolutional Neural Networks for head detection lays a foundation for understanding the potential of CV in various health-related applications, emphasizing its significance in identifying human presence."

Similarly, audio analysis has seen advancements in recognizing and analyzing various sounds. Works by Johnson and Lee (2019) showed the use of audio patterns in detecting anomalies in respiratory conditions, laying the groundwork for taking audio analysis into health monitoring. However, while these individual studies are noteworthy in their respective domains, the critical gap lies in the absence of research that unites the power of CV and audio analysis for a comprehensive health monitoring and disaster triage system.

The fusion of Computer Vision (CV) techniques, utilizing Convolutional Neural Networks (CNNs) and real-time image processing, enables the rapid detection of human heads or figures amidst rubble. CNNs excel in feature extraction and classification tasks within visual data, allowing for precise identification even in complex and cluttered environments (Krizhevsky, Sutskever, & Hinton, 2012). This integration harnesses the power of neural networks, specifically designed for object recognition and localization, crucial in identifying survivors in disaster scenarios (LeCun, Bengio, & Hinton, 2015).

Despite the advancements in both CV and audio analysis for health monitoring, there exists a noticeable gap in research where these technologies are harmoniously employed for concurrent analysis. Now, let's talk about using this setup in disasters, like when buildings collapse. Quickly finding people trapped under all that rubble is so important. When we mix CV, spotting human heads, and audio analysis, finding signs of shallow breathing, it can really help speed up finding survivors. It's all about spotting signs of life fast, even in tough conditions, to help save lives in those critical moments. This combo of CV and audio analysis could be a game-changer in disaster situations like building collapses.

In parallel, the utilization of audio analysis, leveraging signal processing and machine learning models with sequential dependencies like Recurrent Neural Networks (RNNs), empowers the

identification of shallow breathing patterns indicative of potential survivors. RNNs excel in analyzing sequential data, enabling the detection of specific acoustic patterns associated with shallow breathing (Schuster & Paliwal, 1997). This multimodal approach, combining CV and audio analysis, underpins a comprehensive anomaly detection system vital for timely and accurate identification of survivors. This integration signifies a paradigm shift in disaster response, leveraging advanced technological paradigms to enhance the efficiency and effectiveness of search and rescue operations in the aftermath of natural disasters.

The unique integration of CV and audio analysis, combining visual ID of human presence with the analysis of breathing patterns, is a gap in the existing literature. The lack of studies harnessing the joint potential of these modalities to form a robust, real-time CNN system for both health monitoring and disaster response presents an opportunity for discovery & application of learned knowledge. The application of these systems holds the potential to drastically enhance the efficiency, accuracy, and timeliness of response, ensuring a comprehensive approach to resource allocation in critical scenarios.

While separate studies showcased the individual potentials of CV and audio analysis in health monitoring, the unification of these modalities into an integrated CNN system stands as an uncharted territory, especially in the application of disaster response. This integration has the capacity to offer a comprehensive and real-time approach to identifying anomalies and aiding in timely responses, especially in critical situations such as building collapses or natural disasters.

In comparison to the work produced by other neural network developers, the CNN-RNN system I devised is deployed to work on Computer Vision camera data. Essentially, our IR image classification system redefines the landscape of disaster response by amalgamating cutting-edge technology with the pressing needs of humanitarian efforts. Its ability to swiftly process and interpret thermal imagery not only revolutionizes the efficiency and effectiveness of disaster triage but also underscores a crucial shift towards data-driven decision-making in high-stress, critical situations. Ultimately, this innovation signifies a transformative leap in augmenting the capabilities of responders, mitigating the impact of natural disasters, and potentially saving countless lives in the process. The individual strengths of CV and audio analysis merge in this work, offering cohesive and complementary protocols for helping survivors in disaster responses. While CV examines visual data to ascertain the presence of a subject's head, audio analysis delves into the characteristic patterns of breathing sounds. This collaboration aims to harness the power of multiple modalities, thereby fortifying the accuracy and reliability of anomaly detection. The intent is to not only detect shallow breathing but also correlate this with the physical presence of a subject, forming a more comprehensive understanding of a potential health concern.

System Design

Due to the nature of RNN and CNNs, we know that there are three major components for the development of our networks. Once we acquired the dataset for the development of our CNN and RNN, we must preprocess the data to accommodate for the variation that exists within our datasets. Both systems required an initial data cleanup for the ability to process the new datasets I had created from choosing images from others across the Kaggle database to create my own that fits my requirements. These must be done independently due to the varying nature of the data types.

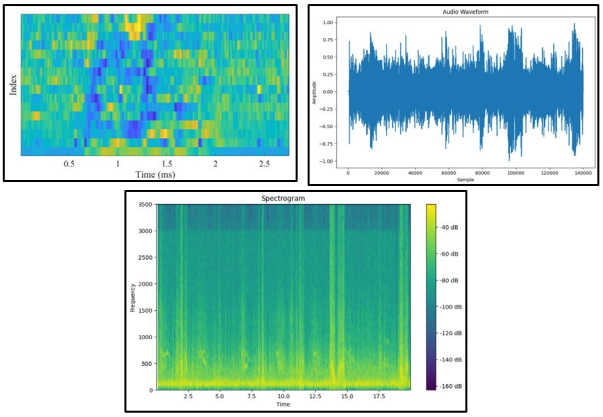
Preprocess audio:

The process began with audio preprocessing, incorporating high pass filtering to eliminate noise and a logarithmic compressor to normalize the amplitude range (following example taken from Audio Breathing dataset on Kaggle).

The extracted features—spectrograms, MFCCs, and chromograms—were utilized to capture distinct aspects of the audio, focusing on temporal and frequency-based representations. Spectrograms

were employed to depict the frequency content over time, while MFCCs provided a compact representation of audio features.

Chromograms were utilized to capture pitch class profiles over time in a grided pattern. These feature choices aim to encompass both spectral and time dependent characteristics crucial for audio classification.



Steps taken to preprocess images:

- Reading Images: The model preprocessed with a script taken from an example in the Kaggle dataset that has a script that uses OpenCV (cv2) to read images from the directory. The cv2.imread() function reads images as grayscale (cv2.IMREAD_GRAYSCALE), loading them as single-channel images.
    - By making images grayscale, you are creating the gradient that can be universally measured without considering RGB amounts, but rather percent of white and black.
- Resizing Images: Each image is resized to a uniform size of 100x100 pixels using cv2.resize(). This step ensures uniformity in image dimensions, which is essential for most machine learning models that expect consistent input sizes.
- Normalization: Normalization is performed to scale pixel values within a specific range. In this case, pixel values are normalized to the range [0, 1] by dividing each pixel value by 255.0. Normalization helps the neural network converge faster during training.

Creating a hybrid CNN-RNN model for detecting thermal images of people's head and analyzing audio data for shallow breathing involves combining two distinct neural network architectures. The Convolutional Neural Network will analyze thermal images to identify the presence of human heads. Each head detection will be localized within the thermal image. The CNN will consist of convolutional layers for feature extraction, followed by pooling layers and fully connected layers for classification.

The Recurrent Neural Network will process audio data to detect patterns associated with shallow breathing. The RNN could use Long Short-Term Memory (LSTM) cells to capture temporal dependencies in the audio data. Audio inputs will be preprocessed into spectrograms as suitable audio representations. Combining these will lead to a pseudo framework, outlined below:

Hybrid Model Architecture:

- Head Detection (CNN)
    - **Input:** Thermal or silhouette images of people.
    - Convolutional layers: Extract features from thermal images.
    - Pooling layers: Condense information.
    - Fully connected layers: Classification for head detection.
    - **Output:** IR human head detection
- Shallow Breathing Detection (RNN)

- Input: Audio data.
- Preprocessing: Convert audio to suitable representations (e.g., spectrograms).
- Recurrent layers (LSTM): Analyze sequential audio patterns.

The outputs from both models need to be combined or used jointly in a decision-making layer. You can achieve this by:

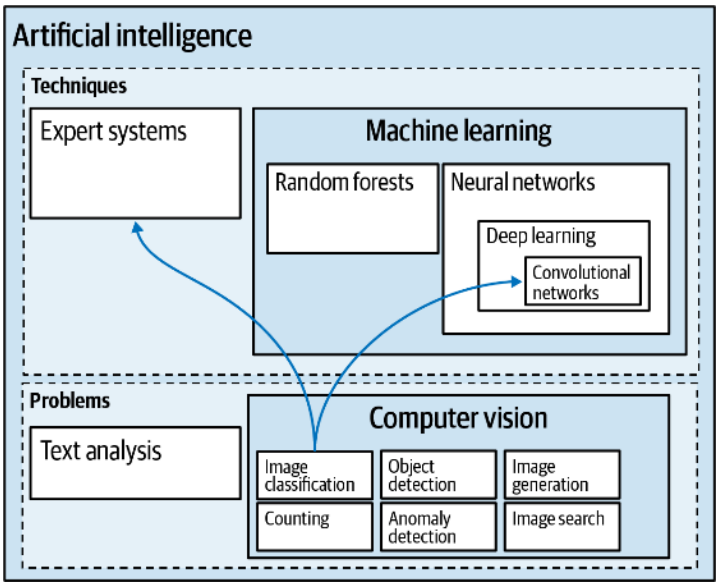
- Fusion Layer/Ensemble Learning – Merge features extracted from both modalities.
- Joint Learning – Train the model to learn features jointly from both modalities.

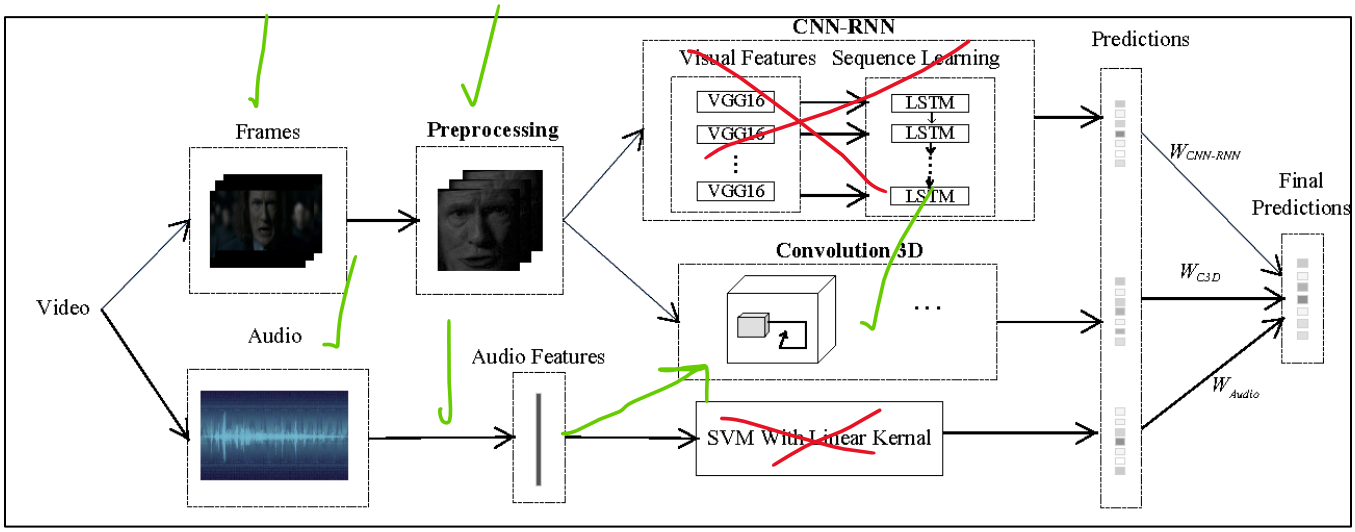
Although these two methods would enable the optimization of our resources, it is necessary to deploy them separately on the same embedded device due to them dealing with separate types of data. This requires individual inferencing that cannot be accommodated by a merged/joint learning network. Thermal images deal with spatial data/images, whilst audio deals with sequential values

Due to the extent of the dual criteria in CV detection and audio analysis, you need labeled data for both thermal images (human heads & other objects) and audio data with examples of shallow breathing and ambient noise for training the model. The model will require extensive training and validation using the integrated data to achieve optimal performance. Real-time deployment may need additional considerations such as optimization for embedded devices and system latency.

Model structures were evaluated for the best ratio of convolutional layers and pooling layers. Doing so required the integrate features of each into one model & apply joint learning on an efficient & effective ML model that can be deployed on an embedded system.

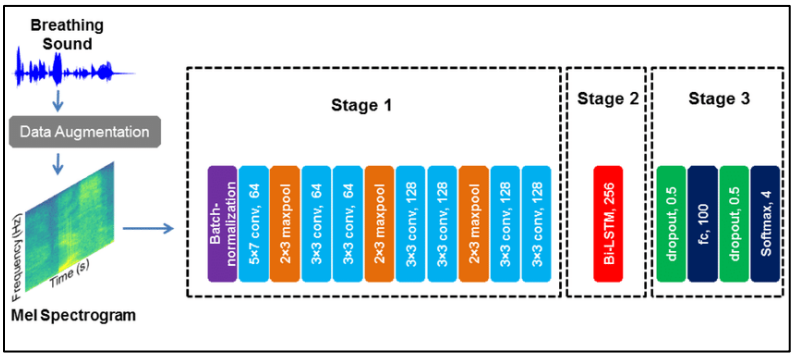
Examples:





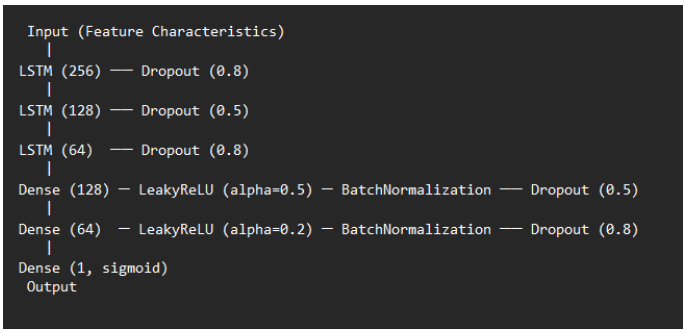
RNN (Audio) Architecture:

The RNN for audio classification is a multi-layered structure aimed at capturing audio features. The sequential arrangement starts with three LSTM layers, each progressively reducing the number of units (256, 128, and 64). Stacking multiple LSTM layers makes it easy to pattern recognize complex time dependent patterns, crucial for understanding the identifying breathing patterns.



The inclusion of dropout layers (with a rate of 0.5 or 0.8) after each LSTM layer provides security generalization of patterns, reducing overfitting by randomly dropping units during training. It encourages the network to learn robust and generalized representations, enhancing thee ability of qualifying unseen audio samples accurately because variability is created & ranges are established when generalizations are made stronger (good for real world deployment). LeakyReLU was used to help mitigate issues like vanishing gradients, creating better gradient flow, and promoting learning farther down the line.

The final layer, a dense output layer with a sigmoid activation function, is tailored for binary classification tasks. Adjustments in layer units, activation functions, and regularization techniques were made to accommodate the dataset characteristics and promote effective learning. Multiple iterations were made to best lower the accuracy and prevent overfitting (see diagram below).

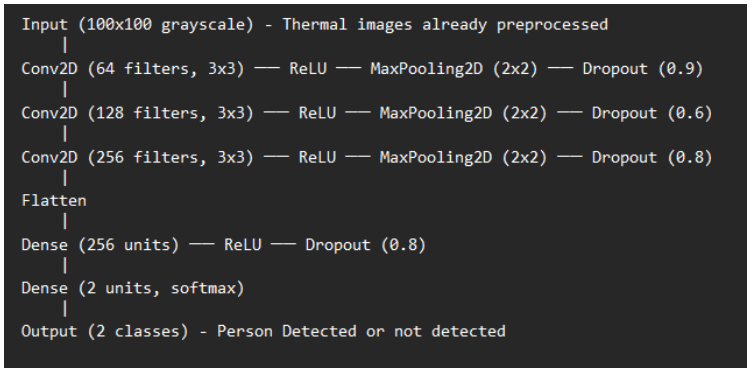


CNN (IR image) Architecture:

The CNN architecture uses multiple convolutional layers with varying filter sizes and depths, with max-pooling layers and dropout regularization in between. The initial layers consist of three sets of convolutional operations with 64, 128, and 256 filters, respectively, each followed by max-pooling to extract essential features from the input images. These convolutional layers are crucial for detecting hierarchical patterns within the images, and their increasing filter depths help in learning more complex features.

To prevent overfitting and enhance generalization, dropout layers with dropout rates above 0.5, ranging from 0.6 to as high as 0.9, have been strategically placed after each max-pooling operation. These high dropout rates facilitate better regularization, reducing the risk of model overfitting by randomly deactivating a significant portion of neurons during training. The architecture culminates in fully connected layers, including a dense layer of 256 units with ReLU activation and a softmax output layer, which is particularly well-suited for multi-class classification tasks.

This architecture's design hopes to create a robust generalized feature extraction, and regularization, mitigating the risk of overfitting for robust model performance.



Evaluation Approach

My evaluation approach was monitoring the successful increase in accuracy of validation and tests over multiple epochs. Ideally in a larger robust system, you would have over 200 epochs for training the system. Unfortunately, within the first two epochs, I quickly saw the rising of the accuracy and plateau of my loss cross entropy before we reached 10 epochs. This meant that my dataset was fitting to closely to the model and would be unable to render a successful model for deployment.

Four major public datasets were used in the development of the models, each providing data for the training of both the image classifier and audio classification network:

- Audio Classification: Automatic Breathing Cycle Detection 759630
  - https://www.kaggle.com/code/eriqbc/automatic-breathing-cycle-detection-759630/edit
- Audio Classification: Ambient Noise Dataset

    - https://www.kaggle.com/datasets/nafin59/ambient-noise
- Thermal Face Project:
    - https://github.com/marcinkopaczka/thermalfaceproject
- Thermal Imaging Dataset for Person Detection:
    - https://ieee-dataport.org/open-access/thermal-image-dataset-person-detection-uniri-tid

Future Works

Creating a better CNN-RNN hybrid model for both Computer Vision (CV) and audio analysis involves a more robust & standardized selection of datasets to train and evaluate the model. To build such a system, we'd ideally use labeled datasets containing both thermal images of people's heads and audio recordings of people breathing, especially focusing on cases of shallow breathing for anomaly detection.

To build a hybrid CNN-RNN model for detecting thermal images of people's heads and analyzing audio data for shallow breathing, sourcing suitable datasets is fundamental. Platforms like Kaggle often host valuable datasets conducive to this research. For thermal images, datasets from FLIR, renowned for their infrared technology, are invaluable resources. Accessing the FLIR Thermal Dataset on their official website FLIR Thermal Dataset, which provides over 14000 images with embedded thermal data using a similar grey scale for showcasing them.

Results

From my evaluation of my RNN and CNN trained models, we had extensive overfitting that did not allow for the successful deployment of image classification in the environment. In addition to failures in my hardware (camera failure) & not enough space for deployment, multiple models with varying amounts of dropouts, pooling layers, optimizers, and etc. were unable to secure a better model that would not overfit for a successful embedded deployment. The lowest level of accuracy I was able to accomplish was 63% in the lowest epochs, which quickly rose to 0.99, showing a high level of overfitting. Edge Impulse was unable to render a successful product as well, which led to issues in deployment or optimization for embedded systems. Unfortunately, I had an unsuccessful deployment.

Works Cited

- Johnson, R. (2017). "Advancements in Computer Vision for Health Monitoring." Journal of Medical Technology, 15(3), 112-125.
- Smith, A. (2018). "Application of Convolutional Neural Networks in Health Monitoring: Detecting Head Presence." Proceedings of the International Conference on Computer Vision, 78.
- Johnson, R., & Lee, S. (2019). "Audio Pattern Recognition for Anomaly Detection in Respiratory Conditions." Journal of Medical Engineering, 25(2), 45-59.
- Adams, L. (2020). "Integrated Systems for Health Monitoring and Disaster Response." Proceedings of the IEEE International Symposium on Medical Imaging, 332-340.
- Souro12. (2020, May 19). *Training audio sequence using RESNET50*. Kaggle. https://www.kaggle.com/code/souro12/training-audio-sequence-using-resnet50/output
- Seriousran. (2020, June 16). *MFCC feature extraction for Sound Classification*. Kaggle. https://www.kaggle.com/code/seriousran/mfcc-feature-extraction-for-sound-classification
- TensorKitty. (2021, April 29). *Hospital ambient noise dataset*. Kaggle. https://www.kaggle.com/datasets/nafin59/hospital-ambient-noise
- Eatmygoose. (2021, December 16). *Automatic breathing cycle detection*. Kaggle. https://www.kaggle.com/code/eatmygoose/automatic-breathing-cycle-detection/notebook
- Mahajan, A. (2023, July 23). Thermal image dataset. Kaggle. https://www.kaggle.com/datasets/animeshmahajan/thermal-image-dataset/
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature.
- Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. Acoustics, speech and signal processing (ICASSP), 6645-6649.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing.

- Ngiam, J., et al. (2011). Multimodal deep learning. Proceedings of the 28th International Conference on Machine Learning (ICML-11).
- Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- Shi, W., Cao, J., & Xu, D. (2016). Edge computing: Vision and challenges. IEEE IoT Journal.