# NANOTECHNOLOGY ENGINEERING

## NE 459 Engineering Project Course

# Audio Classification Using Acoustic Triboelectric Nanogenerators: A Comparative Study

**Karim Habashy**

**Written to: Dr. Dayan Ban, Dr. Pengcheng Xi, Dr. Youngki Yoon**

**Submitted: 10/04/2023**

# Abstract:

Data availability has led to significant development in the domain of Deep Learning. Traditionally, classification methods have relied on microphones for signal acquisition. However, with the rise of edge computing, there has been a growing interest in energy efficient solutions that can deliver the same functionality. Acoustic Triboelectric Nanogenerators (TENGs) are promising candidates for this application. These devices leverage the mechanical motion induced by acoustic waves to actuate and create a signal that can be sensed and processed.

This research project investigates the fabrication process of acoustic TENGs and their potential for sound classification using deep learning models. To do this, their frequency response is captured using a frequency sweep, providing a transfer function to be applied to the ESC-50 dataset. The selected models for fine-tuning are the Hierarchical Token Semantic Transformer, the Self-Supervised Audio Spectrogram Transformer, and the Bidirectional Encoder representation from Audio Transformers framework. The performance of these models is compared, providing context to their robustness and usability of their internal representations in a modified domain.

This interdisciplinary research, carried out by the author and members of Professor Dayan Ban's lab, aims to integrate self-powered acoustic edge devices with deep learning audio classification applications. The structure of the report outlines the design process of the device, the creation of the transfer function, model fine-tuning and subsequent results and recommendations.

# Acknowledgements:

I would like to first and foremost thank Dr Dayan Ban for involving me in this highly interdisciplinary project. Because of this project, I was granted the opportunity to explore a topic at the intersection of my interests in the field of Nanotechnology and the Artificial intelligence. I would like to thank Dr. Pengcheng Xi for not only providing guidance and advice throughout this project, but for also helping make me a better researcher over the past year. I would also like to thank Md. Masud Rana and Majid Bagheri for their help and expertise on the device fabrication aspect of the project. Finally, I am grateful to Professor John Saad for his assistance in signal measurement, and Dr. Pascal Poupart for his advice on future work in this project.

# Table of Contents

## Table of Figures

# Introduction

In recent years, the field of signal processing within the domain of deep learning (DL) has been an area of active development due to the growing availability of data. As a consequence, applications of audio DL such as speech recognition and environmental sound analysis/classification have garnered special attention [1]. Traditionally, classification has relied on the use of microphones for signal acquisition, which was then processed and analysed using hand crafted features, extracting the relevant features for classification. However, these techniques would often-times be over-reliant on quality and low-level (non-semantic) characteristics of the captured audio signals, leaving them dependant of the specific microphone used.

Another promising and rising field of study is that of edge computing, were data is processed close to its source, enabling low-latency computing solutions. To this end, alternative acoustic sensing technologies with improved energy efficiency compared to traditional microphones are being explored. Edge computing devices have the imposed limitation of energy availability, so creating devices with low-power consumption or even self-sustaining devices with energy-harvesting capabilities are desirable properties.

One such device is the acoustic triboelectric nanogenerator (TENG), which is a self-powered device that generates electric signals in response to incident acoustic waves. Since they require no external source of energy, they serve as good candidates for edge-computing input devices. One issue of primary concern is the frequency response of the fabricated devices, as they may vary significantly from conventional microphones. In this report, the performance of transformer based DL audio classification models is investigated. A comparative analysis of model performances on a fine-tuning dataset is applied between the pre-processed dataset and one with the applied frequency response of a fabricated acoustic TENG device.

First, the acoustic TENG device is fabricated. Is frequency response is characterized using signals covering frequencies of the audible range between 20 Hz to 20kHz. Both a frequency sweep and pink noise are applied. From this, a transfer function is obtained, representing the device's response to an incident sound. Testing is done to verify the validity of the computed transfer function by comparing its outputs to the device's actual response to various environmental sounds. The transfer function is then used to augment the ESC-50 dataset by simulating the TENG-generated audio signals that would be encountered if should the device be deployed as an edge computing device.

Given this dataset, three pretrained transformer based state-of-the-art models are selected. They are then fine-tuned using the original and augmented ESC-50 dataset. The objective of this report is to assess the models' robustness to the change in frequency responses. While the frequency response of the acoustic TENG is different than that of regular microphones, the semantic content of the audio is not changed through augmentation. The goal of DL is to capture high level semantic information of data. Ideally, then, it follows that the model performances should remain invariant to the change in low-level changes found in the acoustic TENG's frequency response.

This research covers advancements made by me and members of Professor Dayan Ban's lab in this highly interdisciplinary project of Deep Learning (DL) sound classification and self-powered acoustic triboelectric device fabrication. This research direction is novel and paves the way for integration of self-powered acoustic edge devices. The remainder of the report is structured as follows. First, a background on the triboelectric effect is given with a summary of previous approaches involving acoustic TENGs. This is followed by a background on audio DL, specifically leveraging transformer-based models. Secondly, the methodology section presents the device fabrication process, calculation of the transfer function, and model fine tuning. Finally, the results are discussed and future directions for the research project are outlined.

# Background

## Acoustic Triboelectric Nanogenerators

The triboelectric effect is a phenomenon of electrical signal generation as a result of contact friction between materials having different affinities for electrons [2]. In this project, the contact friction is created through the response of a thin membrane to mechanical vibrations induced by acoustic waves. The contact friction of this membrane with another layer of opposing electronegativity matches the frequency of the incident sound, resulting in voltage/current readings that can be translated into audio. Effectively, the actuating membranes act as a capacitor with a variable separation distance, and open circuit voltage can be measured as:

$$V_{oc} = \frac{\sigma x(t)}{\varepsilon_0} \qquad (1)$$

Where σ is the surface charge density of the friction interfaces, $\varepsilon_0$ is the vacuum permittivity, and $x(t)$ is the displacement distance of the actuating layer.

The triboelectric effect consists of the following steps. Initially, the layers are separated. This causes a charge imbalance between the layers due to their varying affinities, resulting in a voltage difference. The imbalance induces a current through electrodes connecting the devices, known as electrostatic induction. Through the incident compressions of the audio, contact electrification occurs, where the 2 materials experience a charge transfer, establishing the original different charges. The layers are then separated, and the process repeats [3].

### Multi-tube Helmholtz Resonator Based Triboelectric Nanogenerator

In Zhang et al., a multi-tube parallel Helmholtz based TENG (MH-TENG) is created mainly as an energy-harvesting system [4]. In this system, the TENG consists of 50 micron thick aluminum plate (tribo-positive layer) that comes into contact with a 30 micron thick Fluorinated Ethylene Propylene (FEP) (tribo-negative layer) film. This device is placed at the backside of a Helmholtz resonator made of polylactic acid (PLA), a box that isolates specific frequencies at which the acoustic waves reach resonance, amplifying the power at those frequencies. A variable number of hollow tube openings to the resonator box let in sound. In the design of the resonator, the relative placements and diameters of the tubes are optimized, leading to a device with the capacity to extract appreciable amounts of electrical energy from low frequency input acoustic energy. It was also found that the use of 4 tubes led to an optimal amplification of sound pressure within the resonator while maintaining the widest frequency bandwidth.

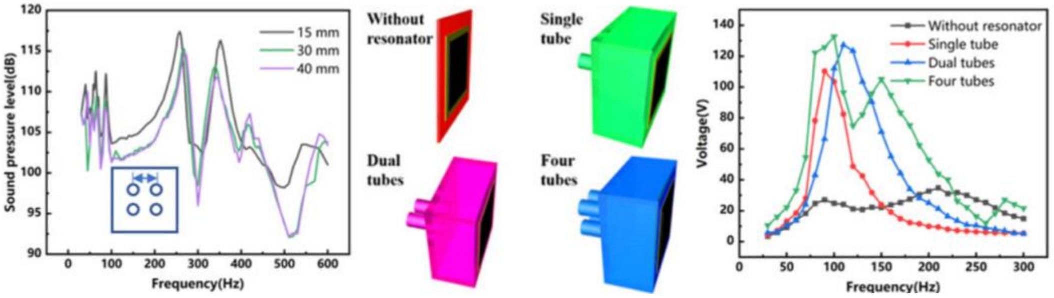

*Figure 1: Multi-Hole Helmhotlz Resonator frequency response and device*

In Figure 1 above, outputs of the fabricated device can be seen. In Figure _A, responses of sound pressure levels inside of resonators with variable tube distance are compared. In Figure _B, one can see different helmoltz resonators variations, followed by their corresponding electrical signal generation performance in Figure _C.

### A Triboelectric-based Artificial Basilar Membrane to Mimic Cochlear Tonotopy

In an alternative application of the triboelectric effect, an artificial basilar membrane (TEABM) is created to mimic cochlear tonotopy as a step towards self-powered cochlear implants [5]. In this design, the TENG device consists of 15 micron thick aluminum foil and a 25 micron thick Kapton film. These are clamped at the edges, and are allowed to separate at the center as a response to incident acoustic waves, leading to triboelectrification.

The main difference between this proposed TENG system and the one mentioned above, however, is that this approach incorporates 8 beams of varying dimensions. The discrete beams range from 8.2 to 32 mm in length and 6 to 8 mm in width. Accordingly, the different beams have differing mechanical properties, and therefore resonant frequencies. Consequently, they collectively cover a much wider bandwidth than any single beam, ranging from 294-2311 Hz. The assortment of the beams as well as their frequency responses can be seen in Figure 2 below.

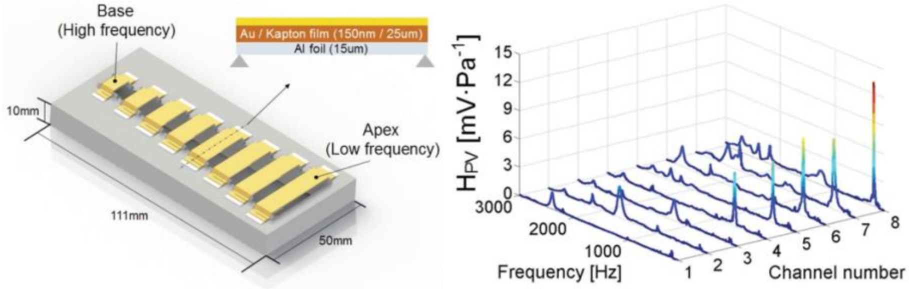

*Figure 2: Cascaded TEABM Design with waterfall plot of frequency response for each channel*

The TEABM was used with deafened Hartley guinea pigs by the electrical response of their spinal ganglial neurons. With stimulating pulses between 1 and 10VPP, correlating to the input sound pressure level from 70 to 85 dB, the eABR signals were identified only when the input acoustic stimulus corresponded to the resonance frequency of that channel. Measurements of eABR were recorded at a latency of 0.7 ± 0.04 ms from the initial acoustic stimulus.

### SATURN: Thin and Flexible Self-Powered Microphone Leveraging Triboelectric Nanogenerator

In Arora et al. [6], triboelectricity was leveraged in a similar fashion to the previously discussed mechanisms. In this approach, instead of leveraging a resonator or an array of sensors, a single large sensor was created. This led to both more contact electrification (higher produced voltage) and higher frequency bandwidth (due to there being multiple modes of vibration). Figure 3 shows the device design

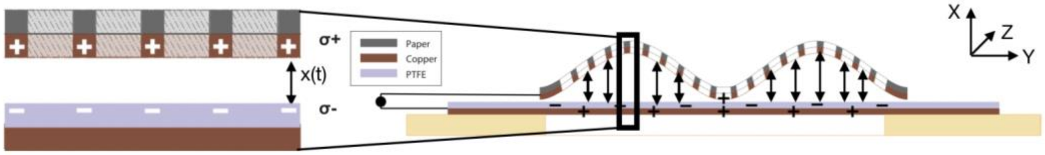

*Figure 3: SATURN device design and actuating principle*

The device consisted of a thin film polytetrafluoroethylene (PTFE), that comes repetitively in contact with copper due to incident acoustic signals. PTFE is an inherently negatively charged material due to high electronegativity. On the other layer, copper was sputtered on a sheet of paper directly, acting as an electrode and tribo-positive layer. Using glue dots, 9 fixed points on the surface of the sheet were set to hold the top and bottom layers together. To reduce air friction during device operation, a laser cutter was used to perforate the surface of the paper prior to sputtering. This step allowed air flow between the layers, leading to larger separation distance and maximum achievable voltage.

To evaluate and optimize the device, variations in the design were applied, and the frequency response of the device was measured. Firstly, hole size, separation of holes, and hole pattern were modified. In addition to improved air flow, perforation also reduces the paper's stiffness, leading to less resistance to deformation. On the other hand, more holes resulted in less contact area, meaning less surface charge density. Effect on maximum and effective separation distance, as well as frequency response was measured and optimized for, resulting in 400 µm holes 200 µm apart with grid spacing (rather than concentric). To continue, surface charge density was further optimized by treating the PTFE film with plasma via O2 plasma etching, effectively leading to a 10 dB improvement in sensitivity. Finally, other parameters such as sensor shape (circular versus square), spacing of the glue dots, presence of back support, and patch size were all studied and optimized for.

As this paper showed the best performance in terms of frequency response, it is the best proposed TENG-based acoustic sensing device. In the lab, the methodology proposed in this paper is implemented.

## Audio Based Deep-Learning

The second component of the project relies on intelligent detection of sound events. More specifically, given input audio signals detected using triboelectricity, the signals can be used as inputs to models for processing and classification. To this end, DL approaches have found great success over recent years due to their ability to learn hierarchical representations from raw data without requiring hand-crafted feature extraction. Convolutional and Recurrent Neural Networks (CNNs & RNNs) have been especially effective in tasks such as audio classification, speech recognition, event detection, among others.

Transformer based models quickly rose to fame following the introduction of the attention mechanism in Vaswani et al.'s paper "Attention is all you need" [7]. Originally proposed for Natural Language Processing (NLP), transformers enabled learning statistical long range dependencies in data sequences. The attention mechanism, at the core of the transformer architecture, allowed for the calculation of mutual information across tokens using scaled dot-product attention seen in Equation 1 below.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Tokens, which are individual components of sequences, are split into query (*Q*), key (*K*), and value (*V*) vectors. A softmax calculation is employed to calculate the dot product of *Q* and *K*, scaled by the dimensionality $d_k$ of the key vectors. It is important to note that the power of this approach is its invariance to the modality of the inputs, as it can, and has been applied to language, vision, and audio. [8][9] [10] . For the audio domain, the base model architecture is the Audio Spectrogram Transformer (AST), seen in Figure 4. It accepts spectrograms as input, which are time-frequency representations of sound. In this project, three audio transformer models – all based on the AST - were compared. They are described below.

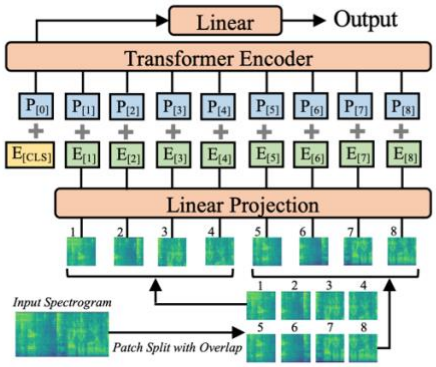

*Figure 4: Model structure of the Audio Sprectrogram Transformer*

## HTS-AT: A Heirarchical Token-Semantic Audio Transformer for Sound Classification

The first implementation is the Hierarchical Token-Semantic Audio Transformer (HTS-AT) [11]. In this paper, Chen et al. introduce an audio transformer with a hierarchical structure, inspired by the SWIN Trasnformer [12]. Their approach was focused on the reduction of model size and training time compared to existing audio transformers. To do this, the original AST is modified by reshaping the outputs of each vertical attention layer (see Figure 5).

While the attention calculation itself is unchanged, it is applied to windows of the original input spectrograms. Each window, consisting of patches, calculates attention scores only within the set window size, instead of globally. A patch-merge layer is used at the end of each group of transformer encoder blocks to reduce the sequence size. This effectively lowers the GPU memory consumption, as there are less computations needed. From here, the windows are shifted, allowing for attention to be calculated across embeddings not previously within the same windows.

The model also includes a token-semantic module that maps the final outputs into class feature maps, making it suitable for audio event detection. In all, the resulting model applies supervised learning to form representations, similar to [10]. However, due to model architecture efficiency, it is able to do so with roughly a third of the number of parameters. Because the embeddings are resized following each encoder block, the model scales well with higher resolution spectrograms, and can theoretically encode longer audio segments with higher frequency resolutions.

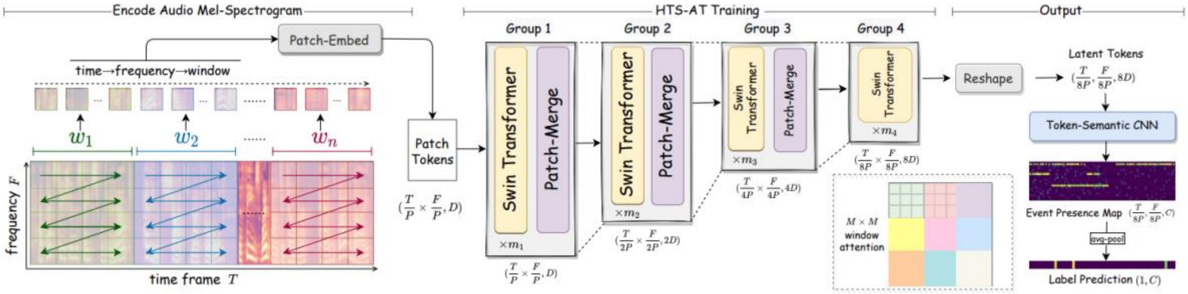
Encode Audio Mel-Spectrogram
Patch-Embed
time→frequency→window
$w_1$
$w_2$
$w_n$
frequency $F$
time frame $T$
Patch Tokens
$(\frac{T}{P} \times \frac{F}{P}, D)$
HTS-AT Training
Group 1
Group 2
Group 3
Group 4
Swin Transformer
Patch-Merge
$\times m_1$
$\times m_2$
$\times m_3$
$\times m_4$
$(\frac{T}{P} \times \frac{F}{P}, D)$
$(\frac{T}{2P} \times \frac{F}{2P}, 2D)$
$(\frac{T}{4P} \times \frac{F}{4P}, 4D)$
$(\frac{T}{8P} \times \frac{F}{8P}, 8D)$
$M \times M$ window attention
Output
Reshape
Latent Tokens
$(\frac{T}{8P}, \frac{F}{8P}, 8D)$
Token-Semantic CNN
Event Presence Map $(\frac{T}{8P}, \frac{F}{8P}, C)$
avg-pool
Label Prediction $(1, C)$

*Figure 5: Model structure of the HTS-AT*

### SSAST: Self-Supervised Audio Spectrogram Transformer

Within Deep Learning, a major development has been the self-supervision paradigm, which allows models to learn their complex representations without the use of labelled data. Two main methods have been employed to achieve this: The first leverages masking random patches from the input, and training the model to predict the token that correctly matches the masked location through reconstruction (generative). In the second method, patches of the inputs are also masked, and the model is tasked with matching the masked patch embeddings given the list of embeddings of all masked samples [13].

In the Self-Supervised Audio Spectrogram Transformer (SSAST), the architecture of the AST is used without change. The output embeddings of the transformer are then used in *both* a generative and discriminative task to learn representations.

### BEATs: Audio Pre-Training with Acoustic Tokenizers

The approach taken by Chen et al. addresses a drawback of reconstruction tasks as a means for learning valuable representations [14]. The authors argue that the task is too focused trying to reconstruct all of the information in the input rather than the semantically essential information, leading to lower robustness to noise and other low-level features, as well as a large number of parameters being wasted on the prediction of these low level features.

Instead, what is proposed is the creation of a tokenizer for audio, as an equivalent to those used in NLP. In language, tokenizers contain all the possible tokens that can be predicted. Consequently, when a token is masked, the predicted token is an existing token in a vocabulary of possible words. These tokens are generated at the level of words, not letters [8]. Similarly, in the speech domain, tokenizers have been created with phonemes as the fundamental units of speech and are used by models for transcription.

In general audio, however, the task is relatively cumbersome, as it is not obvious what the fundamental units of environmental sounds are. In general audio, phonemes are but a small subset of the possible tokens. To address this, the authors employ an iterative pretraining approach for optimizing an acoustic tokenizer along a self-supervised learning model. The model initially uses a randomly initialized codebook to train a base audio spectrogram transformer encoder like the one in  [13], then use the resulting encoder as a teacher to create a better tokenizer. The resulting tokenizer is then used to improve the a subsequent model. This loop is repeated three times, after which the final encoder is fine tuned on downstream tasks. The approach taken is known as self-distillation, and its capacity to produce better student models than the teachers is not yet fully understood [15]. Tokenizer and model training can be seen in Figures 6 A and B, respectively.

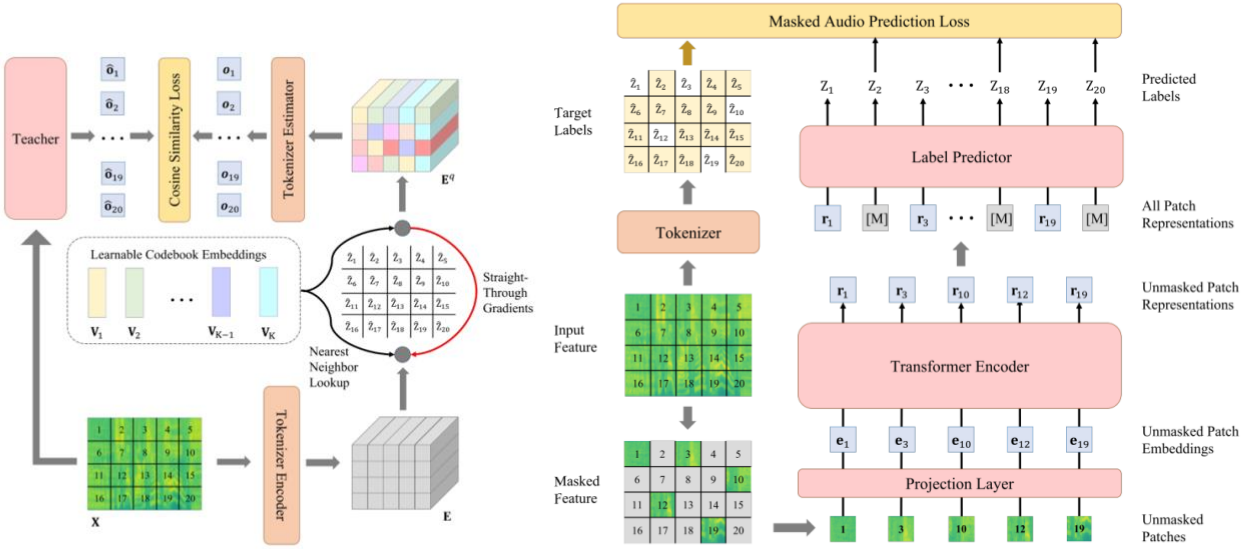

*Figure 6: BEATs Framework. Left: Tokenizer training loop. Right: Self-supervised model training loop*

# Methodology

## Device Fabrication

As mentioned, the approach taken by Arora et al. was the one employed for the creation of the Acoustic TENG. First, the various components of the TENG needed to be fabricated. From there, the device was put together and its voltage response to touch as well as audio were tested and prepared for signal processing.

The fabrication process of an acoustic TENG device began with the preparation of materials, including PTFE (Polytetrafluoroethylene) film, aluminum, paper, PDMS (Polydimethylsiloxane), and copper tape. For the bottom electronegative plate fabrication, a thin layer of copper was deposited on the PTFE film and the paper using sputtering. The Al-coated PTFE served as the bottom plate of the device, with the Al facing down and functioning as an electrode, connected to copper tape for conduction. The paper for the top plate served as a means to hold the deposited thin layer of aluminum, which faced downward and came into contact with the PTFE. The deposited Al functioned both as a tribo-positive layer and electrode. Holes were patterned into the top electrode using a CO2 laser cutter, creating perforations on the Al-deposited paper of 0.04 mm in diameter spaced 0.04mm apart. This allowed air to bounce between the layers, resulting in a better resonance response and improved actuation performance. Copper tape was then applied to the middle facing deposited aluminum. Both pieces of copper tape were connected to the high impedance 100x BNC wire and then to the oscilloscope for voltage and current readings.

For the PDMS separator preparation, PDMS was mixed with the base and curing agent and poured into a square mold to cure. The cured PDMS was then cut into lines and placed along the perimeter of the device. During assembly, the PDMS lines were placed along the perimeter of the bottom electronegative plate, ensuring no overlap between the lines. The top actuating plate was carefully aligned and placed on top of the PDMS lines, allowing the middle section to freely actuate. The copper tape was connected to the exposed aluminum sections on both the bottom and top plates, creating an electrical connection. The other ends of the copper tape were attached to an oscilloscope or other measuring equipment, completing the external wiring for the device. A schematic of the final device shape can be seen in Figure 7 below:

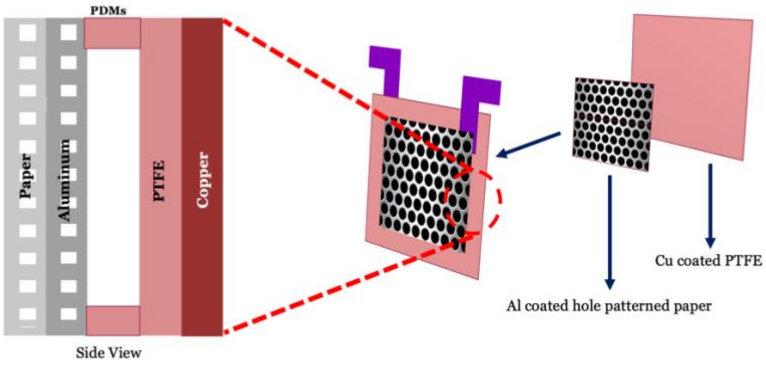

*Figure 7: Device design to be fabricated*

A 3D printed box similar to the one discussed in the first paper was created, with a removable lid with 2 cylindrical holes serving as Helmhotz tubes. Amplification of sound within the box was tested.

Finally, the assembled device was tested and optimized for actuation performance using finger taps (forcing the layers together without the need for sound) and an audio frequency sweep ranging from 20 Hz to 20 kHz.

## Signal Processing

Once the device was prepared, the following process was applied to obtain the transfer function:

1- Prepare a frequency sweep signal from 20 Hz to 20 kHz. This range roughly corresponds to the range of frequencies of humans. Also, all selected DL models were using a sampling frequency of 16 kHz, which is within the covered range.

2- Set the acoustic TENG beside a reference microphone. The reference mic serves as a good baseline, allowing for identification of anomalies in the custom mic's response. It also compensates for environmental factors that may negatively impact the readings such as the acoustics of the room. In this case, an iPhone was used.

3- Play and record the frequency sweep from both devices.

4- Load the responses, ensuring both signals have a matching sampling rate and length.

5- Calculate the Fourier Transform of both signals to obtain their frequency responses.

6- Calculate the ratio of the obtained arrays. The output is the transfer function. Division in the frequency domain is deconvolution in the time domain. Effectively, this process "undoes" the convolution operation of the acoustic TENG's impulse time response from that of the reference microphone. This is done to isolate the characteristics of the TENG device from those of the reference.

## ESC-50 Dataset and model Fine Tuning

The Environmental Sound Classification (ESC-50) dataset is a collection of 2000 environmental sound recordings. The data is manually selected from the publicly sourced Freesound Project, a collaborative database where users upload a wide variety of audio [16]. The classes, including animal, human-related, urban, and natural sounds are suitable for environmental classification. The dataset is widely used to benchmark audio DL models. To this end, the data is pre-arranged into 5 folds for cross validation purposes.

An important assumption made in this project is that the audio of the dataset is not collected using professional equipment, and that most files originate from videos filmed by phones of users. For this reason, the reference microphone used for transfer function calculation is an iPhone. Normally, to create a transfer function, the target microphone is compared to a reference mic with a flat frequency response.

By using an iPhone for the transfer function, the variations of audio quality in the recording environment of the dataset were more accounted for and were removed in the deconvolution operation. Effectively, the transfer function removes the effects of the microphone used in recording the sample (assumed to be an iPhone) and applies the frequency response of our created TENG device.

Using the resulting transfer function, the ESC-50 dataset is augmented. To do this, a python function was created that loads in each sample of the dataset, takes the FFT of the sample, applies the transfer function (through multiplication of the transfer function – convolution in the time domain), and takes the iFFT. The new TENG simulated sample would then be loaded into another folder containing all TENG simulated samples. The entire dataset was processed through the transfer function in this way and saved as a second dataset. From here, all three models were fine-tuned on both datasets, and their performances are compared.

Fine-tuning, or transfer learning, is the process of using a pretrained model as a basis for learning a new task (different from the pretraining task) that involves similar information to what the model had originally learned. In this process, instead of using a model with random/uninitialized weights, the pretrained weights are used to process the data, but slightly modified to match the demands of the new task. In the vision domain, an example of this would be the fine tuning of a model pretrained on general images on a dataset of specific animal species. Since the pretrained model already has information about the structure of images, as well as some of the semantic content of what an animal should look like, fine tuning only lightly tweaks the weights. Generally, the more difference there is between the pretrained and fine-tuning data, the poorer the performance of the fine-tuned model.

In this project, all models are pretrained on the AudioSet dataset, a collection of over 2 million audio recordings. They are all fine-tuned on the ESC-50 dataset, which shares significant similarity in the classes of the sounds. However, the ESC-50 dataset augmented using the transfer function has differences in the lower-level information. More specifically, the frequency response of the TENG-device differs from that of generally used microphones (higher frequencies having lower power output due to higher frequency modes being more difficult to actuate). It is possible that the new dataset is different enough from the original un-augmented version (and pre-training dataset), that the TENG augmented fine-tuned models perform weakly.

## Results and Discussion

In this section, the outcomes of the acoustic TENG device fabrication and assembly is assessed, the resulting transfer function used in ESC-50 dataset augmentation is discussed, and model fine-tuning for sound classification is evaluated and compared among the three models selected.

### Device fabrication and assembly

The fabricated device can be seen in Figure 8. When creating the various devices, a variety of design decisions were considered to optimize device performance. Device performance was compared with and without laser cutting the holes on the deposited aluminum. Various spacings of patterned holes were also compared in terms of efficiency. As mentioned in the background section, this parameter was critical in determining the optimal surface charge accumulation while maintaining actuation voltage.

To continue, the back rest of the device was also modified to examine the effect of the actuating environment. First, the device was placed on a solid back support. It was theorized that this would hinder actuation since it would make the device unable to achieve its full range of motion during actuation. It was then placed on a hollow HDPE back support (used in the figure below) to address this issue. Finally, the Helmholtz resonator setup was also tested for signal amplification.

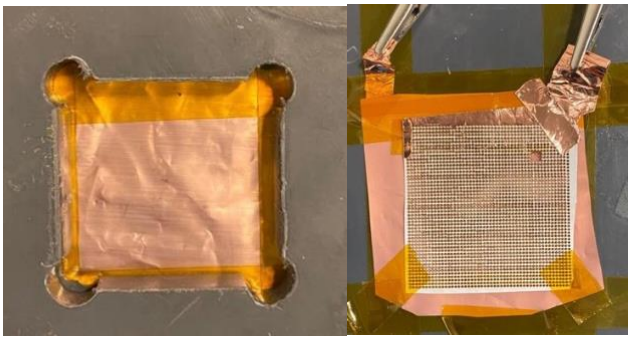

*Figure 8: Front and backside of the fabricated TENG device*

Before sound response testing, hand taps on the devices were applied to ensure the device was properly assembled and could transmit signal through significant mechanical actuation. In Figure 9, voltage and current measurements from the hand taps can be seen. As expected, these tests led to much higher response magnitudes than through sound actuation, seen further down.

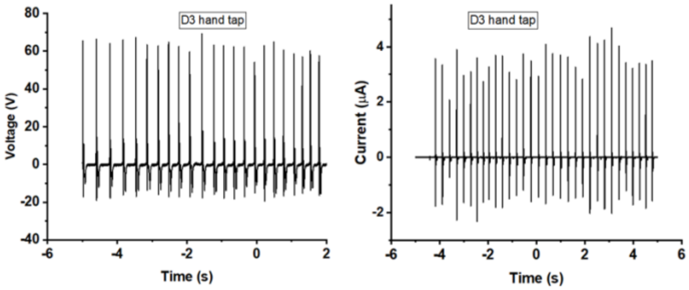

*Figure 9: Hand tap voltage (left) and current (right) responses of the TENG device*

The results of the different devices were difficult to recreate and were rather inconsistent. On some runs, a peak-to-peak voltage was measured up to 4V. On others, it was found to be as low as 1V peak-to-peak. This made it difficult to compare across devices. It was however easier to examine the effects of different setups using the same device. For example, the voltage response to the frequency sweep in Figure 10 can be seen to change depending on whether the resonator box was open or closed. Specifically, in the closed configuration, the voltage response was approximately half of its open counterpart. Regardless of configuration, however, the trend of the voltage response (and consequently frequency response), was consistent.

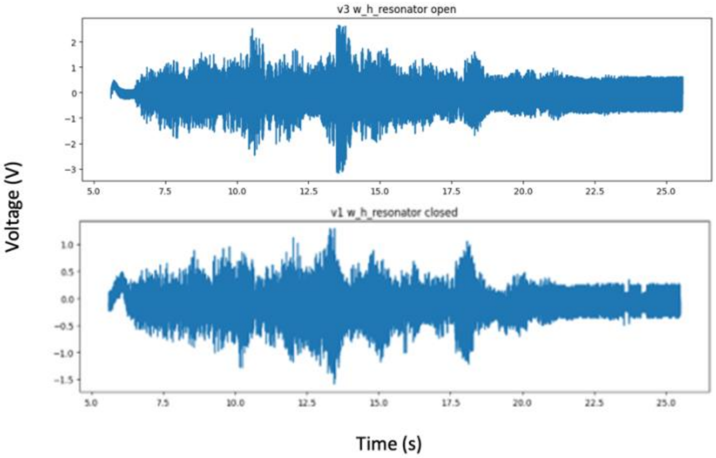

*Figure 10: Voltage response to 20 Hz -20 kHz frequency sweep*

## Transfer Function

The best configuration of the device was judged to be the one using 0.04mm diameter holes 0.04 mm apart, with the back free to actuate and the open box (likely used as a waveguide for the sound). This setup was used to record a signal alongside an iPhone recording the frequency sweep. Results of the spectrograms and transfer function of this setup can be seen in Figure 11. One of the problems encountered when measuring the voltage response was the low sampling frequency of the oscilloscope. To address this, another reading was taken at the 20 kHz sample rate on a different oscilloscope.

This however did not change the results. This is because sampling rate of the oscilloscope was not the limiting factor of the device. Instead, it was the limited rate of micro-vibrations of the plates. When measuring the frequency response, it was seen that resonance occurred around 400 Hz, with a sharp drop afterwards. By the 1 kHz mark, the signal was already 100 dB lower than the reference mic. This result is significant, as the device needs to be capturing a much wider bandwidth than this to allow for effective model classification.

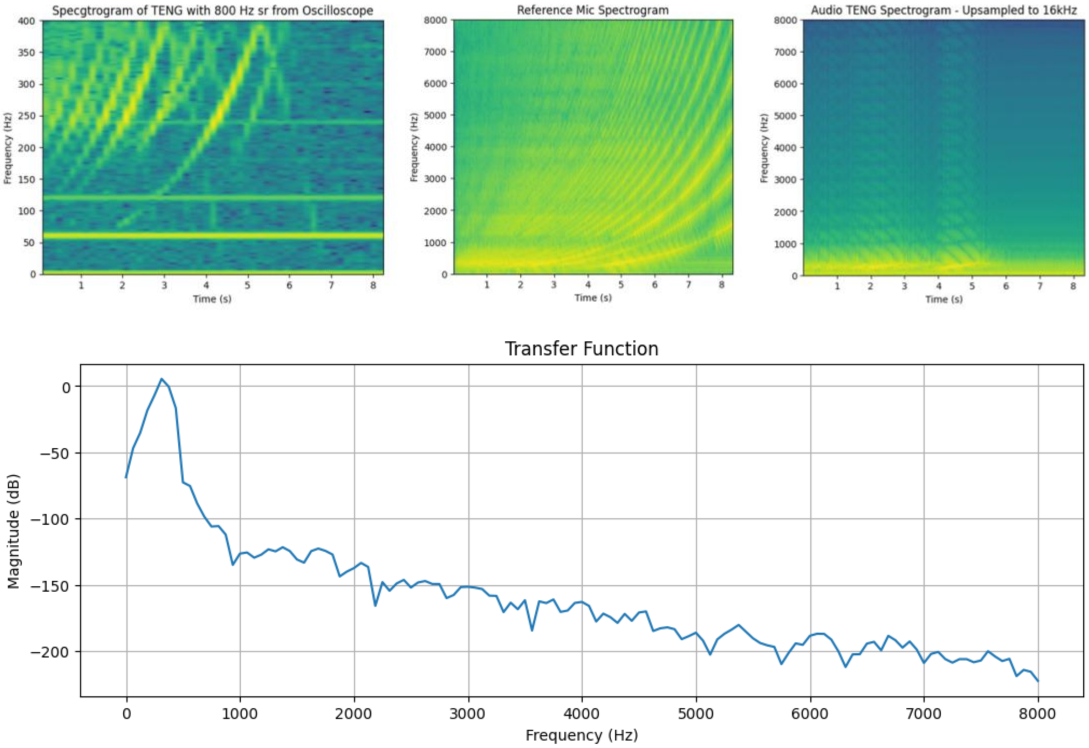

*Figure 11: Spectrograms of the frequency sweep of a) TENG, b) reference iPhone mic, c) upsampled TENG, Transfer function can be seen in the bottom graph*

The spectrograms above show the frequency response over time of the TENG device and the reference microphone. Ideally, the response should be a single line increasing exponentially from low to high frequency over the duration of the recording. While the exponential trend behaviour is somewhat observable in the signals, resonant effects can be identified in both devices as multiple stacked lines are represented. The TENG device reading has additional frequencies at the 60 and 120 Hz due to the power line hum. To match the models' sampling frequency requirements, the reference microphone was downsampled to 16 kHz from 48 kHz. Similarly, the TENG device's signal was upsampled to 16 kHz. The result is the rightmost figure. Evidently, there is a clear loss of information at higher frequencies. The implication of this in terms of the dataset performance is that low frequency sounds should in theory significantly outperform higher frequency ones in terms of classification score. This is tested in the next section.

## Model fine tuning: robustness to dataset augmentation

From the model fine tuning process, the different models showed very different results. Initially, all models were fine-tuned on the original ESC-50 dataset. This served as a baseline for comparison with both the augmented dataset as well as the findings of the papers proposing the models. One of the models was self-supervised (SSAST), one was supervised (HST-AT), one was self-supervised with additional supervised training including a tokenizer (BEATs). Among these models, the learned representations during pretraining were hypothesized to have a significant effect on the augmented dataset fine tuning performance. Accuracy results of the model fine tuning can be seen in Table 1 below.

| Model | | Claimed | Reproduced | TENG augmented dataset |
|---|---|---|---|---|
| Name | Type | | | |
| SSAST | Self-supervised | 88.8 | 83.5 | 5.25 |
| HTS-AT | Supervised | 97.0 | 96.8 | 10.7 |
| BEATs | Hybrid with tokenizer | 98.1 | 95.0 | 21.0 |

From this table, a few conclusions can be drawn. First, all reproductions of the fine tuning of models on the ESC-50 dataset gave worse results than the claimed performance. This was especially the case with the SSAST and the BEATs models. This is likely due to unoptimized parameters in the fine-tuning pipeline. For example, the learning rate scheduler in the SSAST often would get stuck at a minimum around the 80% mark, with the diminishing learning rate meaning it could not leave said minimum. Looking at the performance of the TENG augmented dataset, the very low relative performance is not surprising due to the poor quality of the data being processed. It should be noted that a prediction based on pure chance would lead to a model with 2% accuracy.

What is interesting, however, is the relative performance of the 3 models on this dataset. While all deep learning models store their representations in the model weights, the nature of these representations varies between supervised and self-supervised models in important ways. On one hand, supervised models explicitly learn the relationships between input data and output labels, with the models learning the most discriminative features of the dataset and using those for classification. The labels are tied to the pretraining task, and consequently so are the learned representations. The weakness of these models is that they do not explicitly learn the structure of the data but instead rely on the association between the data and labels for classification. On the other hand, self-supervised models are forced to learn the underlying structure and statistics of the data, since, by definition, they have no labels for supervised training. In order to have good performance on downstream tasks, the representations need to be robust and generalizable. Generally, their limitation is the resources wasted during pretraining where the model learns fine-grained, low-level representations in reconstruction-based tasks [14].

In relation to the augmented dataset, the fine-tuning task was not changed. However, the data itself changed significantly, losing most of the high frequency information contained in the audio. Consequently, the structural information learned during self supervised model pretraining could not prove be useful, producing a model (SSAST) that struggled more to learn the necessary associations during the fine-tuning task. Comparatively, the supervised model (HTS-AT) was relatively able to adapt by making associations between the labels and the modified data. This explicit mapping would have allowed the model to leverage more of the information learned during pretraining.

From here, it followed that a model with the pretraining task of predicting patch level tokens would have the best performance, not only due to its hybrid supervised and self-supervised pretraining, but also due to its ability to learn more semantic content at the patch level (rather than the pixel level). Figure 12 shows a comparison between the original and augmented dataset fine tuning performance. From a spectrogram processing standpoint, the bottommost patches containing the lowest frequency information were relatively unchanged from those in the pretraining task, retaining their usability in the fine-tuning process. This would have been especially helpful in the case of the BEATs model, as each patch is fed through the tokenizer for a patch level classification.

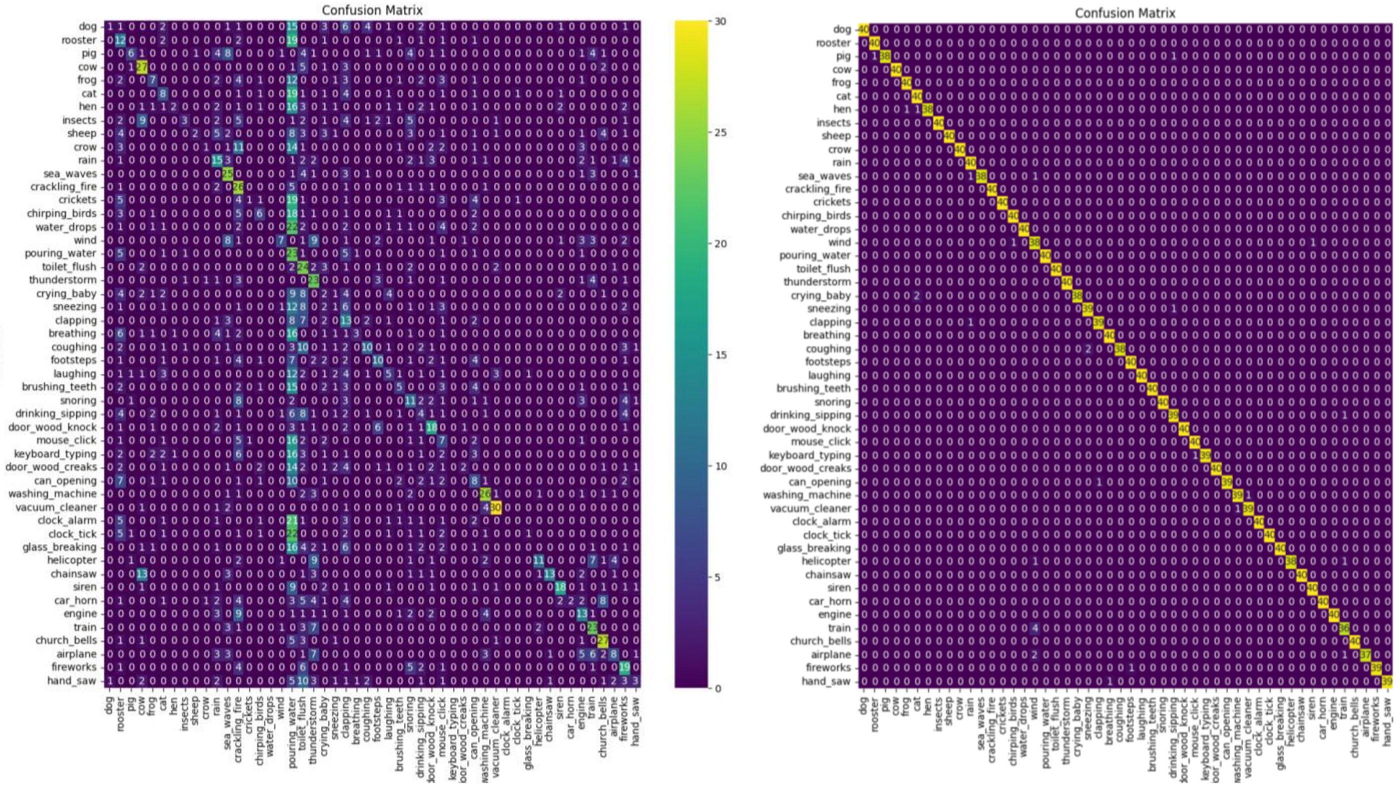

*Figure 12: Confusion matrix comparing BEATs model fine tuned onTENG augmented data (left), and original dataset (right)*

Looking at the confusion matrix above, we can assess the model performance following the fine tuning on the ESC50 TENG and base datasets. While the right-handside results may be unsurprising, the above chance performance of the TENG augmented dataset model is of above what one would expect with the given data. The main diagonal of matching true and predicted categories shows adequate performance. Interestingly, the most guessed classes are the vacuum cleaner, pouring water, washing machine, and crackling fire. As mentionned, this can be expected, as these classes consist of sounds with low frequency information that may have remained relatively intect during the processing through the transfer function.

## Recommendations

This report covers the first full pass through the creation pipeline of a classification enabled acoustic triboelectric nanogenerator. This project is deeply interdisciplinary, with the design incorporating aspects relating to the physical device such as material selection and methods of assembly. Signal processing is needed to obtain a frequency response from the actuating device for the creation of a transfer function to simulate the Triboelectric device. One can then modify existing datasets without the requirement of manually recording thousands of samples. Finally, the resulting dataset can be used to fine-tune a number of audio models on the downstream task of TENG simulated data classification. Therefore, there are refinements to be made at each step that can lead to significant progress and improvements in performance.

The quality of the TENG device voltage output is of paramount importance. Essentially, the quality of the produced signal is the parameter that cascades through the pipeline and affects overall performance the most. To improve the fabrication of the acoustic TENG, several modifications can be considered in the design and assembly process. This includes a wider range of material consideration for the tribolelctric layers, changing Helmholtz box/resonator geometry, quality of device assembly, etc. During assembly, PDMS placement can be modified: instead of pre-dried PDMS, one can apply the material while it is still in its liquid phase and let it dry over while in contact with the 2 plates at the desired locations. The fabrication process itself can be modified, where both top and bottom plates can be patterned differently to explore the effects of modifying the interaction of the air between the layers and improving representable frequencies in actuation. Furthermore, hole placement geometry can also be modified, exploring various configurations such as random, grid, or varying diameters to maximize the device's efficiency. Further simulations should be employed. Finally, as proposed by Jang et al., a broader range of resonant frequencies can be covered if a system of multiple nanogenerators is built with varying actuator dimensions [5].

On the machine learning classification side, several observations are made. First, it is evident from the performance plots that the representations learned by the DL models during pretraining do not lend themselves useful during classification. This is heavily affected by the audio quality of the TENG-modified ESC-50 fine tuning dataset. As it stands, the data generated by the TENG device is significantly different from that of the pretraining set. Unless the transfer function can be changed enough to produce meaningful data at higher frequencies, it is recommended that the Transfer Learning paradigm from large model pretraining be substituted for one of two alternatives.

The first one would be a generative approach. Specifically, the training task is audio super-resolution. This would be similar to stable diffusion modelling, where original model inputs are corrupted through added noise (in our case, this is through the transfer function causing loss of high frequency information) [17]. The training task of such models is to learn to recuperate the original signal given the corrupted signal. Practically, the model would be learning the relationship between low frequency patterns and the high frequencies they correspond to over time. A U-Net like design architecture would be leveraged, where the TENG transfer function would be the output of the encoder [18]. To obtain the intermediate states between the original input and the outputs, an interpolation metric can iteratively be applied back to a flat original frequency response (input signal). The model would essentially have to learn the inverse of this “noising” or, in our case, low pass filtering of the signal.

The second option would be to fully train a smaller model. This would be done to allow the model to learn more general patterns given the lack of higher frequency information. In terms of overall project scope of the project, one significant advantage it would have would be its speed and efficiency. Considering the end goal is smart detection, a smaller model can be deployed to continuously monitor its environment without large power requirements. This may be a meaningful implementation, as one could potentially train a small model of a few million parameters (as opposed to the hundreds of millions of parameters of the DL models) and have it run on a Raspberry Pi or other low latency chip for a fully self-powered system [19]. The use case for such a model is a passive energy harvesting device that can periodically sample its environment and transmit a classification result on the sample. Based on the output class, one can use the result as a wake signal for an energy intensive model with a better equipped microphone.

# Conclusion

In this report, the interdisciplinary field of deep learning based audio classification and self-powered acoustic triboelectric nanogenerators was explored. This work encompassed the various aspects of TENG design, fabrication, signal processing, and state of the art audio transformer based model fine tuning. The TENG-augmented dataset was revealed to be significantly different from the original ESC50 pretraining dataset, significantly impacting the performance of the models on the task of fine tuning. Nevertheless, their robustness to the effects of the fabricated device was compared, revealing that the representations gained by jointly having supervised and unsupervised pretraining lead to the most resilient model.

Various recommendations were provided at every level of the project to create a better end-to-end prototype. In conclusion, this research paves the way for integrating self-powered acoustic edge devices into various applications requiring intelligent classification. Future work should focus on refining the TENG design and fabrication process, as well as exploring other avenues within the machine learning domain. Ultimately, this would lead to the deployment of these devices in low power settings while enabling intelligent monitoring.

# Works Cited

[1] Y. L. A. G. H. YOSHUA BENGIO, *Deep Learning for AI,* https://doi.org/10.1145/3448250, 2021.

[2] Z. L. Wang, "Triboelectric Nanogenerator (TENG)—Sparking an Energy and Sensor Revolution," *Advanced Energy Materials,* vol. 2000137, pp. 1-6, 2020.

[3] Z. Z. Shuaihand Pan, "Fundamental theories and basic principles of triboelectric effect: A review," *Friction 7,* pp. 2-17, 2019.

[4] X. Z. W. Y. L. L. Y. H. W. H. a. X. M. Zhang Q, "Multi-Tube Helmholtz Resonator Based Triboelectric Nanogenerator for Broadband Acoustic Energy Harvesting," *Frontiers in Materials,* vol. 9, 2022.

[5] J. L. J. H. J. a. H. C. Jongmoon Jang, "A Triboelectric-Based Artificial Basilar Membrane to Mimic Cochlear Tonotopy," *Advanced Healthcare Materials,* vol. 5, no. 19, pp. 2481-2487, 2016.

[6] N. a. Z. S. L. a. S. F. a. O. D. a. W. Y.-C. a. G. M. a. W. Z. a. S. T. a. W. Z. L. a. A. G. D. Arora, "SATURN: A Thin and Flexible Self-Powered Microphone Leveraging Triboelectric Nanogenerator," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 2, no. 2, pp. 1-28, 2018.

[7] N. S. N. P. J. U. L. J. A. N. G. Ł. K. I. P. Ashish Vaswani, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[8] M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, Minneapolis, 2019.

[9] L. B. A. K. ,. D. W. ,. X. Z. ,. T. U. M. D. M. M. G. H. S. G. J. U. N. H. Alexey Dosovitskiy, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in *ICLR*, 2021.

[10] Y.-A. C. J. G. Yuan Gong, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021.

[11] X. D. B. Z. Z. M. T. B.-K. S. D. Ke Chen, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *ICASSP*, 2022.

[12] Y. L. Y. C. H. H. Y. W. Z. Z. S. L. B. G. Ze Liu, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *ICCV*, 2021.

[13] Y. L. a. O. V. Aaron van den Oord, "CPC: Representation Learning with Contrastive Predictive Coding," 2018.

[14] Y. W. C. W. ,. S. L. ,. D. T. Z. C. F. W. Sanyuan Chen, "BEATS : Audio Pre-Training with Acoustic Tokenizers," 2022.

[15] C. B. a. K. M. Linfeng Zhang, "Self-Distillation: Towards Efficient and Compact Neural Networks," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* vol. 44, no. 8, pp. 4388-4403, 2022.

[16] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane, 2015.

[17] A. B. D. L. P. E. B. O. Robin Rombach, "High-Resolution Image Synthesis with Latent Diffusion Models," in *CVPR*, 2022.

[18] E. M. a. V. Välimäki, "A TWO-STAGE U-NET FOR HIGH-FIDELITY DENOISING OF HISTORICAL RECORDINGS," in *ICASSP*, 2022.

[19] Applied Brain Research, "Time Series Processor Chips," ABR, 2023. [Online]. Available: https://appliedbrainresearch.com/hardware/time-series-processor-chip. [Accessed 09 04 2023].