Surveying Sidewalk Materials For and By Individuals Who Are Blind or Have Low Vision: Audio Data Collection and Classification *

Jiawei Liu[1,2][0009−0001−1684−6823], Wayne Lam[3][0009−0008−4148−6656], Hao Tang[1,2][0000−0002−6197−0874], and Zhigang Zhu[1,3][0000−0002−9990−1137]

[1] The Graduate Center of the City University of New York, New York, USA
jliu9@gradcenter.cuny.edu
[2] Borough of Manhattan Community College - CUNY, New York, USA
htang@bmcc.cuny.edu
[3] The City College of New York - CUNY, New York, USA
wlam001@citymail.cuny.edu, zzhu@ccny.cuny.edu

Abstract. Navigating safely and independently presents considerable challenges for people who are blind or have low vision (BLV), as it requires a comprehensive understanding of their neighborhood environments. Our user study reveals that understanding sidewalk materials and objects on the sidewalks plays a crucial role in navigation tasks. This paper presents a pioneering study in the field of navigational aids for BLV individuals. We investigate the feasibility of using auditory data, specifically the sounds produced by cane tips against various sidewalk materials, to achieve material identification. Our approach utilizes machine learning and deep learning techniques to classify sidewalk materials solely based on audio cues, marking a significant step towards empowering BLV individuals with greater autonomy in their navigation. This study contributes in two major ways: Firstly, a lightweight and practical method is developed for volunteers or BLV individuals to autonomously collect auditory data of sidewalk materials using a microphone-equipped white cane. This innovative approach transforms routine cane usage into an effective data-collection tool. Secondly, a deep learning-based classifier algorithm is designed that leverages a dual architecture to enhance audio feature extraction. This includes a pre-trained Convolutional Neural Network (CNN) for regional feature extraction from two-dimensional Mel-spectrograms and a booster module for global feature enrichment.

* The work is also supported by the National Science Foundation (NSF) through Awards #2131186 (CISE-MSI), #1827505 (PFI), and #1737533 (S&CC), the US Air Force Office of Scientific Research (AFOSR) via Award #FA9550-21-1-0082, and the ODNI Intelligence Community Center for Academic Excellence (IC CAE) at Rutgers University (#HHM402-19-1-0003 and #HHM402-18-1-0007). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Air Force Office of Scientific Research, and the Office of the Director of National Intelligence.

Experimental results indicate that the optimal model achieves an accuracy of 80.96% using audio data only, which can effectively recognize sidewalk materials.

Keywords: Auditory · Assistive technology · Blindness and low vision · Deep learning · Material recognition · Navigation.

1 Introduction

The World Health Organization (WHO) estimates that there are 285 million people with visual impairment worldwide, among whom 39 million are totally blind [1]. People who are blind or have low vision (BLV) face many challenges in their daily lives, including the difficulty of navigating safely and independently [2]. To navigate effectively, individuals with BLV need to acquire as much spatial information as possible from their surroundings, including information about sidewalk materials and defects [3]. Regrettably, most existing advanced applications [4–6] do not provide sufficient functionality to help BLV people collect landmark information and understand sidewalk conditions. Mobile navigation applications with GPS and mapping services (such as Google Maps and Apple Maps), mainly focus on providing efficient, short navigation routes, which is insufficient for BLV individuals [7, 8]. Therefore, their preferences have to tilt towards paths rich in tactile landmarks and minimal sidewalk defects, prioritizing safety and reliability over shorter distances.

As for the BLV individuals, they often rely on white canes to explore their surroundings, via auditory feedback that enhances their spatial awareness and assists in self-localization [3]. For example, they can follow the tactile shoreline in their travel by identifying surface material changes, such as grass edges or raised curbs. Many street intersections are equipped with tactile pavements of varying materials and patterns, designed to aid BLV people in identifying important locations, such as street crossings, bus stops, and the direction of streets. These surface materials serve as effective landmarks and hence their inclusion in accessible maps is crucial. Maps including sidewalk materials would be very useful to BLV people, facilitating real-time navigation and trip planning. To gain a deeper understanding of the challenges faced by BLV individuals in their lives, we have conducted an informal user study with BLV individuals [13]. This study has revealed that materials and objects on sidewalks play a crucial role in navigation tasks. Moreover, BLV individuals highlight the critical role of audio signals in identifying sidewalk landmarks and ensuring safe travel in urban areas.

This study conducts a preliminary investigation into the use of non-visual, audio-based data, specifically the sounds produced by the cane tips of visually impaired individuals rubbing against different sidewalk materials, for the identification of materials that are challenging to differentiate by sight. Leveraging machine learning (ML) and deep learning techniques, this research centers on the classification of sidewalk materials using exclusively auditory cues. This inquiry lays the groundwork for a future in which BLV individuals can independently

gather data on sidewalk materials during their routine travels, transforming the mundane act of cane usage into an opportunistic data collection method. Such a paradigm not only fosters autonomy among BLV individuals but also augments the navigational data repository with their unique, experientially-rich insights. As BLV individuals navigate diverse urban terrains, their canes evolve into dual-purpose instruments, serving both personal navigational needs and the communal objective of enhancing a dynamic, adaptive mapping infrastructure responsive to the intricacies of urban settings.

In pursuit of this innovative future, our study undertakes the development of a novel data collection methodology, enabling individuals with blindness or low vision (BLV) to autonomously gather auditory data. Additionally, this research introduces a deep learning-based classification system focused on categorizing these auditory signals. The key contributions of this paper are:

1. The design of a lightweight data collection method for BLV individuals to acquire non-visual information on sidewalk materials. We equipped the white cane with a microphone that captures auditory feedback through audio data as the cane contacts the sidewalk surface. Additionally, we have generated an auditory dataset regarding sidewalk material using the proposed method.
2. The development of a deep learning-based classifier algorithm to identify different sidewalk materials using only audio data. We proposed a dual architecture for feature extraction: (a) a pre-trained Convolutional Neural Network (CNN) model was used to extract the regional characteristics from the two-dimensional Mel-spectrogram; and (b) a booster module aimed at global feature extraction from the Mel-spectrogram representations to enhance the audio feature extraction. The encouraging outcomes of our experimental evaluations underscore the robustness and practicality of this algorithm.

The organization of the paper is as follows: Section 2 delves into the related work, providing context and background. Section 3 outlines the data acquisition approach in detail. Section 4 expounds on the classification methodology. Section 5 discusses the results derived from our experiments. Finally, Section 6 offers concluding observations and remarks.

2 Related Work

2.1 Material Recognition

Most existing research on material recognition relies heavily on visual cues. One notable study [14] achieved significant results by focusing on three key elements: material image datasets, contextual influences, and unique descriptors of material appearance. In addition, numerous studies have explored the utility of light field (LF) images for material identification [15]. An alternative view of material recognition has been proposed using a combination of acceleration and images, and a fully convolutional network has been deployed for joint surface material

recognition [16]. In contrast, our project mainly utilizes non-visual data, specifically audio data. Our deep learning classifier aims to use non-visual forms of data to discriminate sidewalk materials, thus providing a new perspective in the field of material recognition.

2.2 Feature Engineering

A prevalent trend in the acoustic community involves the preprocessing of raw audio data to convert it into spectrograms, including Mel-Spectrogram and Mel-frequency cepstral coefficients (MFCC). These characteristic visual representations then serve as inputs to intricate network models for training. Several studies have affirmed the effectiveness of CNN based models when applied to spectrograms [11, 12]. Remarkably, most state-of-the-art results have been achieved through transfer learning, employing pre-trained CNN models like ResNet50 [10]. Interestingly, one notable study indicated that CNNs pre-trained with regular images, such as ImageNet, remain proficient at extracting critical features from audio spectrograms [17]. Additionally, through a series of tests we have found that features derived from the mean, minimum, and maximum values of Mel-Spectrogram frequency bands have discernible decision boundaries. Our approach aims to synergize both forms of audio data: taking advantage of deep learning to extract rich and effective features from spectrograms of audio data, while using global features derived from statistical techniques to "boost" training on the audio data.

3 Data Acquisition

This section presents the overview of the data collection process. A modified white cane was used to capture the unique acoustic feedback of different sidewalk materials. The following subsections detail the equipment used and the methods of data collection, including both static and continuous modes, to ensure a diverse dataset. The sidewalk material audio data acquisition was performed by 23 volunteer students who embarked on an expansive data collection mission across 4 of the 5 boroughs of New York City. This audio data inventory provided us with a basis for a training dataset for our proposed classifier, which is further detailed in Section 4. We will introduce the audio data collection equipment and inventory in the following subsections 3.1 and 3.2.

3.1 Audio Data Collection Equipment

A lightweight data collection system is designed to acquire acoustic feedback when the cane contacts with sidewalk surfaces using a modified white cane (Fig. 1).

As a white cane interacts with different materials, distinct acoustic signals are produced by the cane tip (a metal tip is used in the system). To capture those differences, a wired microphone was positioned near the cane tip clipped to a

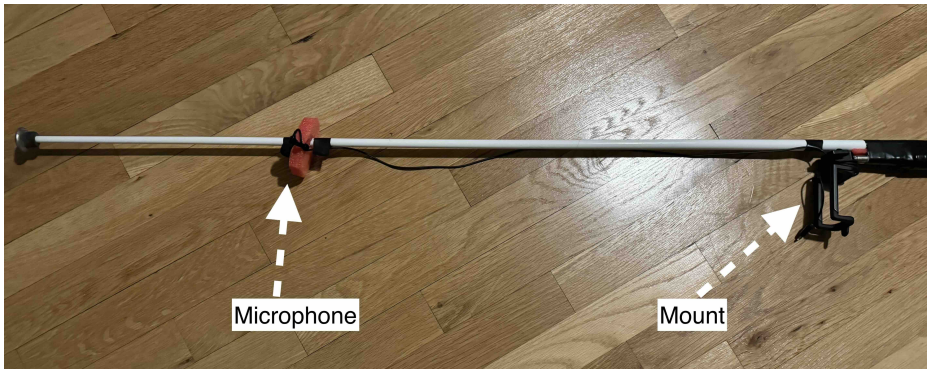

Fig. 1. The modified white cane for sidewalk material audio data collection.

foam ring to maximize the clarity of recorded sounds while minimizing ambient noise and cane vibration noise. Additionally, a mount for a phone was installed to aid in recording video data utilized as a reference for annotations.

3.2 Audio Data Collection

To assemble the sidewalk material audio data inventory, a large-scale data collection effort was initiated involving student volunteers. The data collection was acquired in two different modes: static and continuous.

- **Static data.** Twenty-three (23) sighted volunteers collected a substantial amount of sidewalk material data. Each data record contains a duration of 30+ seconds of a singular surface material category. This was done in order to approximate BLV individuals stopping at certain sidewalk landmarks and repeatedly surveying the material with a white cane in order to ensure they have arrived at a location (e.g. tactile pavement strips on sidewalk intersections).
- **Continuous data.** Four (4) sighted volunteers collected sidewalk material data walking along longer strips of sidewalk to better emulate the standard walking conditions of BLV users. In this mode of data collection, each data record would often contain multiple sidewalk landmarks such as manhole covers and subway grates, providing us with data for multiple categories. With the annotation tool Label Studio [24], each data record was manually annotated by labeling delimiting points of sidewalk materials in the accompanying video.

To ensure accuracy and relevance, the collected sidewalk material audio data was classified according to the criteria specified in the New York City Street Design Manual [9] and the Guidebook for Accessible Sidewalk and Street Intersection Information [18].

The audio data inventory contains several hours of static data and continuous data. This dataset represents a diverse and comprehensive encapsulation of 11 categories including concrete, asphalt, dirt, grass, metal, manhole, granite, tactile pavement, brick, subway grate, and cellar door.

4 Material Classification Approach

4.1 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project for transforming raw data into a form that machine learning models can learn effectively. In this project, the goal of data preprocessing is to slice the audio data into manageable, trainable pieces, and to convert these pieces into a format suitable for deep learning classifiers to learn. Fig. 2 provides a schematic diagram of the data preprocessing pipeline used in this study. The pipeline consists of three main components: data preparation (Fig. 2, Part I), data slicing (Fig. 2, Part II), and data transformation (Fig. 2, Part III). These components are briefly described in the following subsections for completion even though we mostly follow existing approaches.

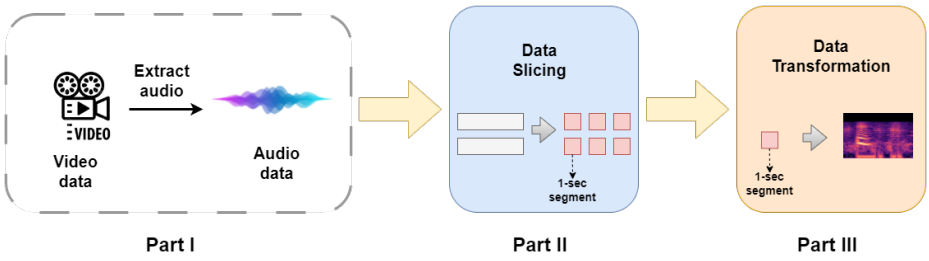

Fig. 2. Schematic diagram of data preprocessing pipeline. **Part I** illustrates the module for audio data extraction from video data; **Part II** depicts the module for slicing data into 1-second time slices; **Part III** shows the module for data transformation of audio bitstreams to Mel-spectrograms.

Data preparation. In the initial stage of data preparation, as depicted in Fig. 2, Part I, our methodology involves extracting audio data from the corresponding video recordings. This crucial step is followed by a meticulous process of resampling the audio data to a frequency of 44 kHz. We employ the Sinc interpolation method for this purpose, a technique renowned for its efficacy in handling nonuniform sample rates [19]. Additionally, the audio data undergoes a rechanneling process to convert it into a stereo format. This rechanneling serves a dual purpose: firstly, it aligns the data with the common standards of auditory

processing, and secondly, it facilitates a more nuanced analysis by preserving spatial characteristics of the sound, which could be crucial in distinguishing between different types of sidewalk materials. Stereo audio, with its dimensional quality, offers a richer dataset for the subsequent stages of processing and analysis [23].

Data slicing. The next step is data slicing (Fig. 2, Part II), which is the process where we decompose the original audio data into manageable segments that are suitable for training our deep learning classifier. A sliding window technique establishes a fixed-length window that moves across the data sequence with a determined step size. Each shift of this window generates a new data segment, enabling the extraction of localized features from the time series data.

It is worth noting that each data segment slice has a corresponding annotation label. With static collection data, data segment labels are consistent across all slices from the original audio data. However, with continuous collection data, data segment labels first reference the manual annotations for the audio data and then select the category with the greatest duration within a particular data segment (1 second is used in our experiments).

Data transformation. The final step is data transformation (Fig. 2, Part III) where we performed a Mel-spectrogram transformation on the segmented audio and acceleration data. This transformation maps raw data onto a two-dimensional grid, with the horizontal axis representing time and the vertical axis denoting frequency. The Mel-spectrogram efficiently captures both the spectral and temporal properties of the raw audio signal, which are crucial for our sidewalk material classification task.

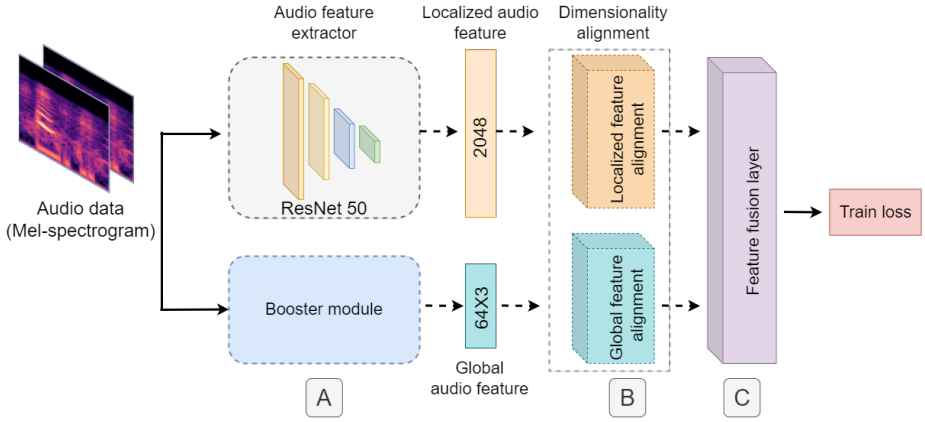

Fig. 3. The booster model for audio sidewalk material classification.

4.2 Classifier Architecture

The core of our material classification algorithm lies in the sophisticated architecture of our deep learning-based classifier (Fig. 3). The proposed architecture has been carefully designed to proficiently process and analyze audio data to accurately identify different sidewalk materials. At the core of its efficacy is a dual-mechanism approach that includes two key modules: the feature extraction module and the booster module. These two modules work in tandem to meticulously extract and analyze features from the Mel-spectrogram representation of audio data. Together, these modules form the backbone of our classifier's architecture, playing an instrumental role in converting the auditory nuances, captured through our innovative data collection methods, into meaningful and practical insights. Subsequent subsections will delve into the specifics of each module, illustrating their respective functions and their synergistic operation in our classification algorithm.

Feature extraction module. Given the richness of the audio data acquired, the raw data has been transformed into Mel-spectrograms, which are essentially visual representations of the spectrum of frequencies in a sound signal as they vary with time. The utilization of Mel-spectrograms for feature extraction has been consistently corroborated by numerous studies in the field, highlighting its efficacy in capturing pertinent audio information [17]. These Mel-spectrograms are excellent candidates for the application of deep learning-based feature extraction methods, as also expressed by [10] in their pioneering work on audio classification.

With the Mel-spectrogram representations in hand, we harness the power of pre-trained deep neural network architectures to extract robust features. This approach, while novel in our specific application, stands on the shoulders of previous studies which have emphasized the robustness of pre-trained models in extracting meaningful features from complex data [17]. Specifically, we employ a transfer learning approach using the ResNet model [21, 22]. Leveraging the representational power of ResNet, we discern and isolate the most important localized audio features from the Mel-spectrogram, preparing the audio data for subsequent stages of the classification pipeline.

Booster module. The booster module in our classifier architecture plays a pivotal role in augmenting the feature set extracted from the Mel-spectrogram representations. This module is intricately designed to process the Mel-spectrogram across the time axis, capturing the {minimum, maximum, and mean} values of each intensity band. This operation is executed for each of the 64 Mel frequency bands. As a result of this process, the engineered data assumes a structured shape of (64, 3), where the three channels correspond to the minimum, maximum, and mean values for each of the 64 Mel frequency bands, which serve as global features. Global features, in contrast to the local or regional features extracted by the pre-trained ResNet50 model, encapsulate overarching patterns

and trends present in the entire audio sample. They provide a macroscopic perspective of the data, capturing broad, holistic properties that are not bound to specific time frames or localized spectral regions. The global features, in synergy with the regional characteristics extracted by the ResNet50 model, create a more robust and nuanced feature set. This enriched feature set helps to improve the accuracy and reliability of the classifier, ensuring a more efficient and detailed interpretation of the audio data for material classification.

Feature fusion. The most straightforward method of fusion is concatenation. After the feature extraction module and the booster module (as in Fig. 3, block A), two discrete feature vectors, denoted V_{local} of 2048 dimensions and V_{global} of 192 (64×3) dimensions, are obtained. To attain consistency in dimensions, a dimensionality alignment layer was designed where regional characteristic features V_{local} will be fed into the localized feature alignment module, where it will conduct down-sampling to v_{local} of 128 dimensions. While the global feature V_{global} will be fed into the global feature alignment module where it will resample to v_{global} of 128 dimensions as well (as in Fig. 3, block B). They are then synthesized into a single, unified, lengthened feature vector v within the local and global feature fusion layer (as in Fig. 3, block C).

$$v = v_{region} \oplus v_{global} \tag{1}$$

5 Experimental Results

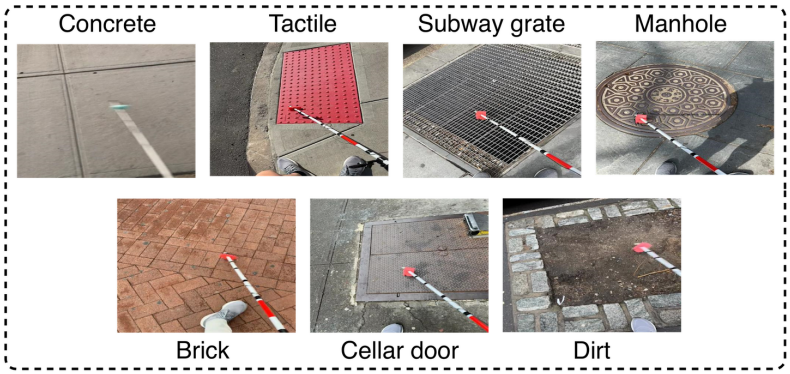

Fig. 4. Typical image of primary categories utilized in training dataset.

5.1 Dataset

Training a deep learning model requires a large amount of data. In this study, we selected seven (out of eleven) categories from the audio data inventory that were

most commonly found on New York City sidewalks. These categories (Fig. 4) include: *concrete, tactile pavement, subway grate, manholes, bricks, dirt, and cellar doors.*

For training data selection, data with portions of the missing audio due to lost microphone connection or other audio signal errors were removed from the dataset. Additionally, static audio data from the most prevalent categories was removed to prevent a heavy class imbalance. Fig. 5 illustrates the data distribution between static and continuous training data of each category and among the seven categories. Each of these categories encompasses data with almost 60 minutes of duration, creating a robust foundation for our model training.

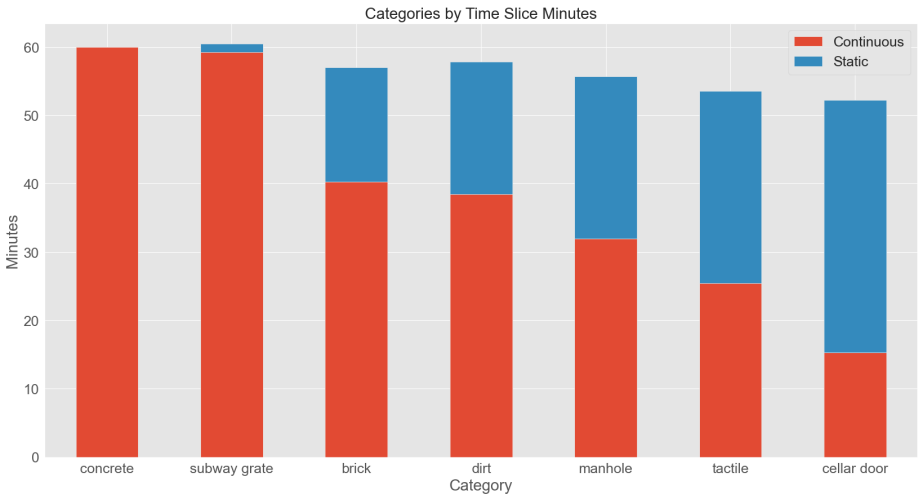

Fig. 5. Data duration for each utilized primary category.

5.2 Implementation Details

Mel-spectrogram transformation and parameter selection. The conversion of audio data to Mel-Spectrograms is a critical step in our preprocessing pipeline, enabling the effective extraction of features relevant to our classification task. In this study, the selection of key parameters was guided by empirical testing, with the chosen settings optimizing for both accuracy and model efficiency. Key parameters in this transformation include:

- **Number of Mel bands.** 64, chosen to capture a wide range of frequencies while maintaining computational efficiency.
- **Frame length and hop length.** 1024 and 256 respectively, balancing temporal resolution and frequency resolution.

Model architecture adaptations. As mentioned above, we utilized ResNet50 to conduct localized feature extraction from Mel-Spectrograms in conjunction with a booster module for global feature extraction from Mel-Spectrograms as well. Later, the localized and global features will be aligned by their alignment modules respectively.

Within the network's architectural scaffolding, the Rectified Linear Unit (ReLU) was employed as the activation function. Further, we also integrated Batch Normalization layers to stabilize and accelerate training by normalizing intermediate feature maps, and applied the dropout with the probability of 0.2 where it randomly zeros some of the input tensors to improve the model's generalizability and robustness.

In this study, the proposed model is implemented in Pytorch and the overall model size is 24.7M parameters. The detailed implementation of each module is listed as follows.

- **Audio feature extractor.** To adapt the ResNet50 model, we initially experimented with several variants, toggling the number of frozen blocks in the architecture. As the crux of deep learning is finding the right amount of transfer versus fine-tuning, our experimentation revealed that freezing just the initial block led to an optimal balance, outperforming other configurations in terms of classification efficacy. Post this freeze, the terminal classification layer was excised, thus enabling the network to produce a feature vector of length 2048, encapsulating the richer semantics of our data without an unwarranted imposition of specificity.
- **Dimensionality alignment.** In the dimensionality alignment module, there are two main components, namely localized feature alignment and global feature alignment. For the localized feature alignment component, it includes two feedforward layers; the first layer is used to down-sample the features from 2048 to 512 dimensions while the second layer is used to down-sample the intermediate features from 512 to 128 dimensions. Likewise, the global feature alignment component is also a feedforward layer that resamples the global features from 192 to 128 dimensions aligned with localized features. Notably, all these layers are followed by the pre-defined Relu, batch normalization and dropout to avoid the overfitting problem.
- **Training procedures.** Our training procedures include an initial learning rate of 1×10^{-5}, reduced by a factor of 10 upon plateauing of validation loss using the Pytorch learning rate scheduler [25]. The proposed model is trained in 50 epochs with a batch size of 128. Adam optimizer with weight decay of 1×10^{-5} was employed for its efficiency in handling sparse gradients and adaptive learning rate capabilities. In terms of the loss function, we employ a cross-entropy given its effectiveness in handling multi-class classification problems.
- **Inference time.** In order to test inference times for the model, we ran model inference over 3000 samples 3 times with an "off-the-shelf" CPU (Intel i9-13900KF, 3.00 GHz, 32 GB RAM) and an "off-the-shelf" GPU (Nvidia GTX 4090, 24 GB RAM), the average inference time per 1-second audio segment

were 13.3 ± 0.005ms and 8.1 ± 0.002ms, respectively, making it suitable for near-real-time applications.

Evaluation metrics. Given the inherent imbalance in our dataset, traditional accuracy metrics would provide a skewed representation of the model's prowess. To counter this, macro accuracy, which computes accuracy for each class and then averages it, was utilized. Complementing this, the macro F1 score was also used, which provides a harmonized mean of precision and recall, thus giving a balanced view of the model's performance across diverse classes.

Validation strategy. To ensure our model wasn't merely memorizing the idiosyncrasies of our dataset, we implemented a K-fold cross-validation approach [20]. The entire dataset was meticulously partitioned into eight distinct folds. However, given the computational overheads and our endeavor to remain time-efficient, we eschewed exhaustive validation across all folds. Instead, a representative subset of three randomly selected folds was earmarked for cross-validation. This approach ensured a rigorous assessment while balancing computational feasibility.

5.3 Ablation Study and Model Comparison

To discern the efficacy and contribution of each modality in our task, we first tested the feasibility of utilizing the minima, maxima, and average of Mel frequency bands by training a Multilayer Perceptron with the data. We then undertook an ablation study comparing a ResNet50 model, and a ResNet50 plus audio booster model. Additionally, a selection of standard machine learning models (K-Nearest Neighbor, Naive Bayes, RandomForest, and Support Vector Machine) trained on flattened one-dimensional arrays of the Mel-spectrogram images derived from the audio data was utilized as a basis for comparison.

Table 1. Ablation study results in terms of macro accuracy and macro F1-score.

Model	**Macro Accuracy**	**Macro F1-Score**
K-Nearest Neighbor	41.11%	32.24%
Naive Bayes	42.30%	38.41%
Multilayer Perceptron	46.54%	45.20%
RandomForest	50.79%	49.98%
Support Vector Machine	65.77%	65.14%
ResNet50	78.31%	78.64%
ResNet50 + Audio Booster	**80.96%**	**80.85%**

Our findings indicate substantially greater classification accuracy and F1-score with the {ResNet50} and {ResNet50 + Audio Booster} models compared to the standard machine learning models. Whereas the Multilayer Perceptron

trained on the average, minima, and maxima of Mel frequency bands places in the middle of the standard machine learning models.

The {ResNet50 + Audio Booster} model appears to have a 2 percentage-point increase for accuracy and F1-score compared to the base ResNet50 model and greater than 20% increase for accuracy and F1-score compared to the Multilayer Perceptron (Table 1).

The empirical results supported our hypothesis: leveraging features engineered from the audio data with statistical techniques in order to boost the training of a CNN deep learning model enhances the model's robustness and accuracy.

6 Conclusion and Future Work

Drawing on observations of the independent travel experiences of visually impaired individuals, our study has explored the use of auditory data from cane tips against different sidewalk materials for surface identification. We have generated a novel audio dataset and developed a model with dual-mechanism for material classification, achieving a promising 80% accuracy.

While the model was developed with a focus on accuracy and robustness rather than real-time classification capabilities, an average inference time of 13.3 ± 0.005ms (CPU mode) per 1-second audio segment, the possibility of classification in real-time in order to help BLV individuals be alerted to materials and obstacles ahead might be worth exploring in the future. An attempt to balance inference time and accuracy might be another pathway worth exploring.

In addition, we plan to explore a crowdsourcing framework, further enabling BLV users to contribute to sidewalk material data collection during their independent travels. This expansion not only aims to refine our existing model but also seeks to actively involve the BLV community in our research process, which could improve the assistive navigation technology.

References

1. World Health Organization: Global data on visual impairment. https://www.who.int/blindness/publications/globaldata/en. Last accessed 30 Mar 2023
2. Donaldson, N.: Visually impaired individuals' perspectives on obtaining and maintaining employment. Ph.D. thesis, Walden University (2017)
3. Herman, J.F., Chatman, S.P., Roth, S.F.: Cognitive mapping in blind people: Acquisition of spatial relationships in a large-scale environment. Journal of Visual Impairment & Blindness **77**(4), 161–166 (1983)
4. Blindsquare. http://www.blindsquare.com/. Last accessed 30 Mar 2023
5. Nearby. https://www.aph.org/nearby-explorer/. Last accessed 30 Mar 2023
6. Seeing. http://www.senderogroup.com/products/seeingeyegps/index.html. Last accessed 30 Mar 2023
7. Ariadne GPS. https://www.ariadnegps.eu. Last accessed 30 Mar 2023
8. Google Map. https://www.google.com/maps. Last accessed 30 Mar 2023

9. Transit street design guide. https://nacto.org/publication/transit-streetdesign-guide/transit-lanes-transitways/lane-elements/pavement-materials. Last accessed 30 Mar 2023
10. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135. IEEE (2017)
11. Maccagno, A., Mastropietro, A., Mazziotta, U., Scarpiniti, M., Lee, Y.-C., Uncini, A.: A CNN approach for audio classification in construction sites. In: Progresses in Artificial Intelligence and Neural Systems, pp. 371–381. Springer (2021)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
13. Liu, J., Tang, H., Seiple, W., Zhu, Z.: Annotating Storefront Accessibility Data Using Crowdsourcing. Journal on Technology and Persons with Disabilities **10**, 154-170 (2022)
14. Trémeau, A., Xu, S., Muselet, D.: Deep Learning for Material recognition: most recent advances and open challenges. arXiv preprint arXiv:2012.07495 (2020)
15. Hu, Z., Chen, X., Yeung, H.W.F., Chung, Y.Y., Chen, Z.: Texture-enhanced light field super-resolution with spatio-angular decomposition kernels. IEEE Transactions on Instrumentation and Measurement **71**, 1–16 (2022)
16. Zheng, H., Fang, L., Ji, M., Strese, M., Özer, Y., Steinbach, E.: Deep learning for surface material classification using haptic and visual information. IEEE Transactions on Multimedia **18**(12), 2407–2416 (2016)
17. Palanisamy, K., Singhania, D., Yao, A.: Rethinking CNN models for audio classification. arXiv preprint arXiv:2007.11154 (2020)
18. Federal Highway Administration U.S. Department of Transportation Accessible Sidewalks and Street Crossings. https://highways.dot.gov/. Last accessed 30 Mar 2023
19. Wolfe, P.J., Howarth, J.: Nonuniform sampling theory in audio signal processing. Journal of the Audio Engineering Society (2004)
20. Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S.: The 'K' in K-fold Cross Validation. In: ESANN, pp. 441–446 (2012)
21. Liu, S., Tian, G., Xu, Y.: A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. Neurocomputing **338**, 191–206 (2019)
22. Du, H., He, Y., Jin, T.: Transfer learning for human activities classification using micro-Doppler spectrograms. In: 2018 IEEE International Conference on Computational Electromagnetics (ICCEM), pp. 1–3. IEEE (2018)
23. Liu, T.,Yan, D.:Identification of fake stereo audio. arXiv preprint arXiv:2104.09832 (2021).
24. Label Studio: Open Source Data Labeling https://labelstud.io/. Last accessed 30 Jan 2024
25. Pytorch Optimization Algorithms: Reduce Learning Rate on Plateau https://pytorch.org/docs/stable/index.html. Last accessed 30 Jan 2024