Classification of Audio Signals using Transformer's Layer-wise Features

Kabilan Mahathevan
Department of
comp. Science and Engineering
University of Moratuwa
kabilan.19@cse.mrt.ac.lk

Vinojith Gunaratne
Department of
comp. Science and Engineering
University of Moratuwa
gunaratne.19@cse.mrt.ac.lk

Jathavan Mahendrarajah
Department of
comp. Science and Engineering
University of Moratuwa
jathavan.19@cse.mrt.ac.lk

***Abstract*—This research explores how pre-trained neural network models can be used to help with audio classification with the help of AudioMNIST, a dataset of 30,000 spoken recordings of numbers in English. The study investigates the effectiveness of pre-training neural networks to acquire general representations, utilising substantial amounts of labelled or unlabeled data. The wav2vec model is employed to extract features from layers 7 to 12, and the extracted features are used to train and evaluate various models on the AudioMNIST dataset, providing insights into the impact of different depths in the context network on model performance for speech-related tasks with constrained labelled data.**

***Index Terms*—pre-trained neural networks, feature extraction, speech recognition, explainable AI, Deep Neural Network**

I. Introduction

This study explores the application of wav2vec [1], a neural network model pre-trained on extensive unlabeled audio data, for feature extraction in the context of the AudioMNIST dataset [2]. The dataset comprises 30,000 spoken digit recordings in English, with 50 repetitions per digit for each of the 60 different speakers. In response to the challenge of limited labeled data, we investigate the effectiveness of pre-training neural networks to acquire general representations, utilizing substantial amounts of labeled or unlabeled data. The wav2vec model, specifically wav2vec large with a 12-layer context network, is employed to extract features from layers 7, 8, 9, 10, 11, and 12. The extracted features are then used to train and evaluate various models on the AudioMNIST dataset, providing insights into the impact of different depths in the context network on model performance for speech-related tasks with constrained labeled data.

A. Motivation

This study focuses on extracting features from spoken digit recordings in English found in the AudioMNIST dataset using the wav2vec neural network model, which is particularly useful when dealing with limited labeled data for speech recognition. The wav2vec neural network model was pre-trained on a large amount of unlabeled audio data. The study aims to understand how changing the depths of the context network within the wav2vec model affects the performance of models in speech-related tasks with constrained labeled data. The ultimate goal is to explore the efficacy of pre-training neural networks to learn general representations from abundant labeled or unlabeled data sources in order to address the problem of scarce labeled data in speech recognition and related fields. The scope of speech recognition systems' potential applications may increase as a result of this.

B. Literature Review

Speech recognition and the fields related to it are getting popular in recent years. Reddy D.R.'s review on recent developments in speech recognition technology, provides a brief overview of the progress, challenges, and aspects within speech recognition which are not addressed, highlighting system components and It examines the organisation of the system for efficient component interaction, the structure and operation of common systems, and the necessary component subsystems.necessary component subsystems [3]. In order to train acoustic models effectively, Steffen Schneider and his team created an effective model that can learn from audio without a lot of labelled data [1]. With potential applications outside of image analysis, Bach, S. et al. present a general method for recognizing classification decisions through the pixelwise decomposition of nonlinear classifiers The issue of comprehending classification decisions by pixel-wise decomposition of nonlinear classifiers has been addressed generally by Bach and his team. The suggested methodology enables projection down to single pixels and the decomposition of a kernel-based classifier into sums of scores over discrete areas of the image. The paper also extends this methodology to neural network architectures, a number of well-known kernels, and mappings. Experiments perform the experimental assessment of the framework, and a discussion follows. Even outside of the currently discussed field of image analysis, the authors believe that this methodology will play a significant role in knowledge discovery, exploratory analysis, and understanding of complex data in the sciences and industry in the future [4].

Wav2vec 2.0, a framework for self-supervised learning of speech representations, has been described by Schneider, S. et al. It has produced outstanding results in speech recognition with very little labelled data. They have compared their method to earlier semi-supervised speech recognition techniques and describe the contrast task defined over the measurement of latent representations in wav2vec 2.0. Additionally, they go over how this technology might affect speech recognition in languages with limited labelled data. Their test results prove wav2vec 2.0's effectiveness on a variety of benchmark datasets [5].

In order to improve understanding of the audio domain name and feature selection, Soren Becker et al. analyse the ability of neural networks in audio classification. Collectively, these works advance the state of the art in speech recognition and related fields of study. Layer-wise relevance propagation (LRP) has been used to investigate the understanding of neural networks in the audio domain for classification tasks. They have presented a new audio dataset of spoken English digits and generated hypotheses about the feature selection of neural networks. These hypotheses are then put to the test by carefully modifying the input data. The purpose of this paper is to improve the interpretability of neural networks in the audio domain and to shed light on how neural networks function internally [2].

When it comes to deep neural networks (DNNs), convolutional neural networks (CNNs) fare better than deep neural networks (CNNs), according to research by Shawn Hershey et al. They also looked into whether the embeddings from these classifiers can improve Acoustic Event Detection (AED) classification and how the size of the training set and label vocabulary affect audio classification. When compared to more straightforward fully connected networks or older image classification models, their findings show that cutting-edge image networks can produce impressive results in audio classification. Additionally, when evaluating on smaller label sets, training on larger label vocabularies can result in modest performance improvements. Comparatively to using raw audio features, using embedded information from CNN classifiers can improve AED classification. [6] Even for models that are typically regarded as inaccessible or ”black boxes,” David Baehrens and his team have developed a method designed to explain specific reasons for each classification decision. Their method includes recognizing local gradients, which specify the modifications that must be made to a data point in order for it to change its predicted label. They use a variety of datasets, including the well-known Fisher Iris dataset and a challenging drug discovery problem, to validate their method. [7]

II. Materials and Method

A. Layer 7 of the wav2vec context network

In the feature extraction process, 768 features were obtained from wav2vec for each layer. However, not all of these features hold equal significance as standalone entities. Recognizing the importance of feature selection as a preliminary step in model construction, we acknowledge the need to refine and prioritize these features. Given that our dataset encompasses four distinct labels—Speaker ID, Age, Gender, and Accent—feature engineering becomes imperative. To address this, individual models will be constructed for each target label, ensuring a focused and tailored approach to the nuanced characteristics associated with Speaker ID, Age, Gender, and Accent.

1) Speaker ID: Our initial analysis involved scrutinizing the dataset for potential high levels of correlation among individual features. In instances where such correlation was identified, we systematically removed one of the correlated features. This precautionary measure aims to mitigate the risk of model instability, as highly correlated features can amplify the sensitivity of the model to small changes in input data. Such sensitivity may result in pronounced variations in the model's predictions, compromising its reliability and interpretability, and potentially leading to overfitting.

It's noteworthy that while certain features may not exhibit high correlation with the target variable when considered individually, they might collectively contribute significantly to the model's predictive capacity. Consequently, we refrained from discarding features with low correlation with the target, recognizing that their collective impact might be valuable for the overall model performance. This nuanced approach ensures a comprehensive exploration of feature relevance and enhances the robustness of the subsequent modeling process.

In the Speaker ID target, there are 60 classes and the dataset is almost equally distributed among them (Figure 1). Since we do not know much about the dataset, we first built a simple model, a Logistic Regression Classifier with all the features and considered it as a baseline model to build different models with different feature sets and we tried different models with different hyper-parameters. Models, validation scores, and selected hyper-parameters are given below in the table I.

XGBoost is often favored over Support Vector Machine (SVM) in the context of multi-class classification. [8] This preference arises from XGBoost's notable strengths, including its proficiency in handling complex non-linear relationships, the utilization of ensemble learning to mitigate bias and variance, feature importance analysis, built-in regularization, scalability, robustness to missing data, and simplified hyper-parameter tuning. Widely acknowledged for its effectiveness in overcoming these challenges, XGBoost stands as a powerful algorithm in multi-class classification tasks, especially for intricate and high-dimensional datasets, providing a compelling combination of superior predictive performance and user-friendly implementation.

However, it is worth noting that in the specific case at hand, SVM surpasses XGBoost [9]. The linear kernel in SVM, in particular, demonstrates superior accuracy compared to other approaches, leveraging the dataset's linear separability to outperform XGBoost. This observation suggests the presence of both linear and non-linear relationships between features

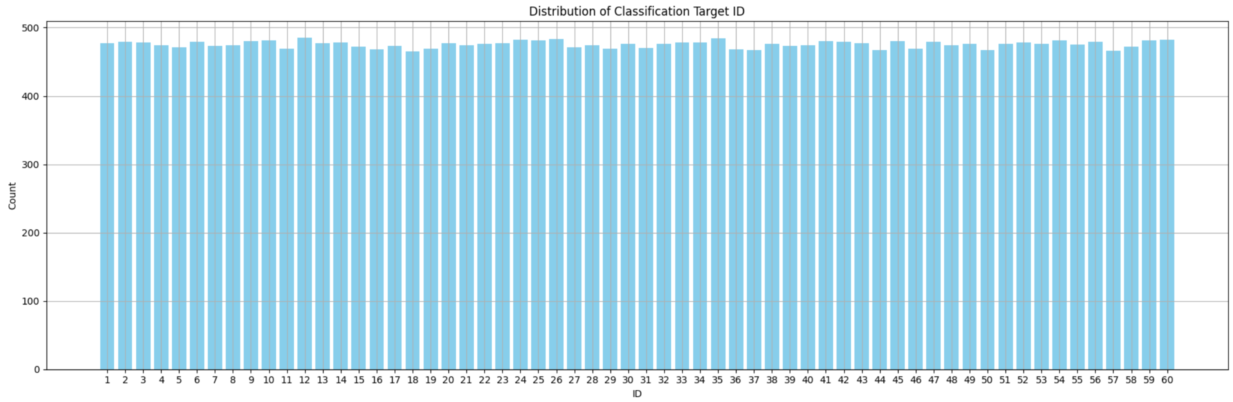

Fig. 1. Distribution of target Speaker ID with different classes of the target

and the target variable within the dataset. Furthermore, the choice of the SVM kernel, along with the polynomial degree, emerges as a critical factor influencing model performance. Notably, the RBF kernel and higher-degree polynomial kernels (degree 4 and above) exhibit particularly strong performance in capturing the underlying patterns in the data.

TABLE I
MULTIPLE MODEL RESULTS ON LAYER 7 FOR SPEAKER ID

Model	hyper-parameter	Validation Score
Logi. Reg. Clas.	1000 maximum iterations	97.33%
Random Forest	n_estimator=100	90.66%
SVM	kernel=RBF	95.47%
SVM	kernel=linear	97.47%
SVM	kernel=sigmoid	89.73%
SVM	kernel=poly, degree=2	95.60%
SVM	kernel=poly, degree=3	97.67%
SVM	kernel=poly, degree=4	97.20%
SVM	kernel=poly, degree=5,6,7,8	97.33%
SVM	kernel=poly, degree=9	96.80%
XGBoost	softmax objective function	91.20%

Prior to model training, we employed lasso regression as a feature selection technique to discard unimportant features. This process resulted in a streamlined feature set, reducing the initial count from 768 to 648. Lasso regression effectively identifies and eliminates redundant or less influential features, enhancing the efficiency and interpretability of the subsequent model while preserving the essential information needed for accurate predictions.

2) Age: Similarly to the approach taken for the Speaker ID target, given the limited initial understanding of the dataset, we initiated our modeling process for the Age label by constructing a simple baseline model. This baseline model utilized a Logistic Regression Classifier incorporating all available features. The intention was to establish a foundational reference point that would serve as a benchmark for subsequent model iterations. For the Age label, the task was treated as a classification problem rather than a regression task, aligning with the dataset's characteristics and the nature of the target variable. This systematic and iterative model-building strategy allows for the exploration of different feature sets to optimize model performance for the Age classification task.

To refine the feature set for the Age label, we employed Lasso Regression once again, targeting features associated with coefficients effectively reduced to zero through L1 regularization. The use of Lasso Cross Validation facilitated the determination of an optimal alpha value for Lasso regression during hyperparameter tuning. Subsequently, a Lasso regression model was created, selecting only the most important features based on their contribution to the target variable.

Despite the similarities in accuracies between Logistic Regression and RandomForest models, both exhibited lower accuracy compared to SVM models. This observation hints at the dataset lacking clear, easily distinguishable boundaries between classes. Moreover, the relationships between features and the target appear to be non-linear or non-trivial. Notably, the SVM with the radial basis function (RBF) kernel outperformed both Logistic Regression and RandomForest, even surpassing the accuracy achieved by SVM with the linear kernel. This disparity underscores the dataset's likely inclusion of intricate non-linear patterns, with the RBF kernel proving particularly adept at capturing and leveraging these complex relationships for improved model performance.

To assess the performance of a neural network on the task at hand, we constructed a model with multiple hidden layers, each defined as follows:

- Hidden Layer 1: Comprising 1024 neurons with Rectified Linear Unit (ReLU) activation and an associated dropout layer featuring a dropout rate of 0.3. The dropout layer serves to mitigate overfitting concerns.
- Hidden Layer 2: 512 neurons with ReLU activation and a dropout rate of 0.2.
- Hidden Layer 3: 256 neurons with ReLU activation and a dropout rate of 0.2.
- Hidden Layer 4: 128 neurons with ReLU activation and a dropout rate of 0.2.
- Hidden Layer 5: 64 neurons with ReLU activation and a

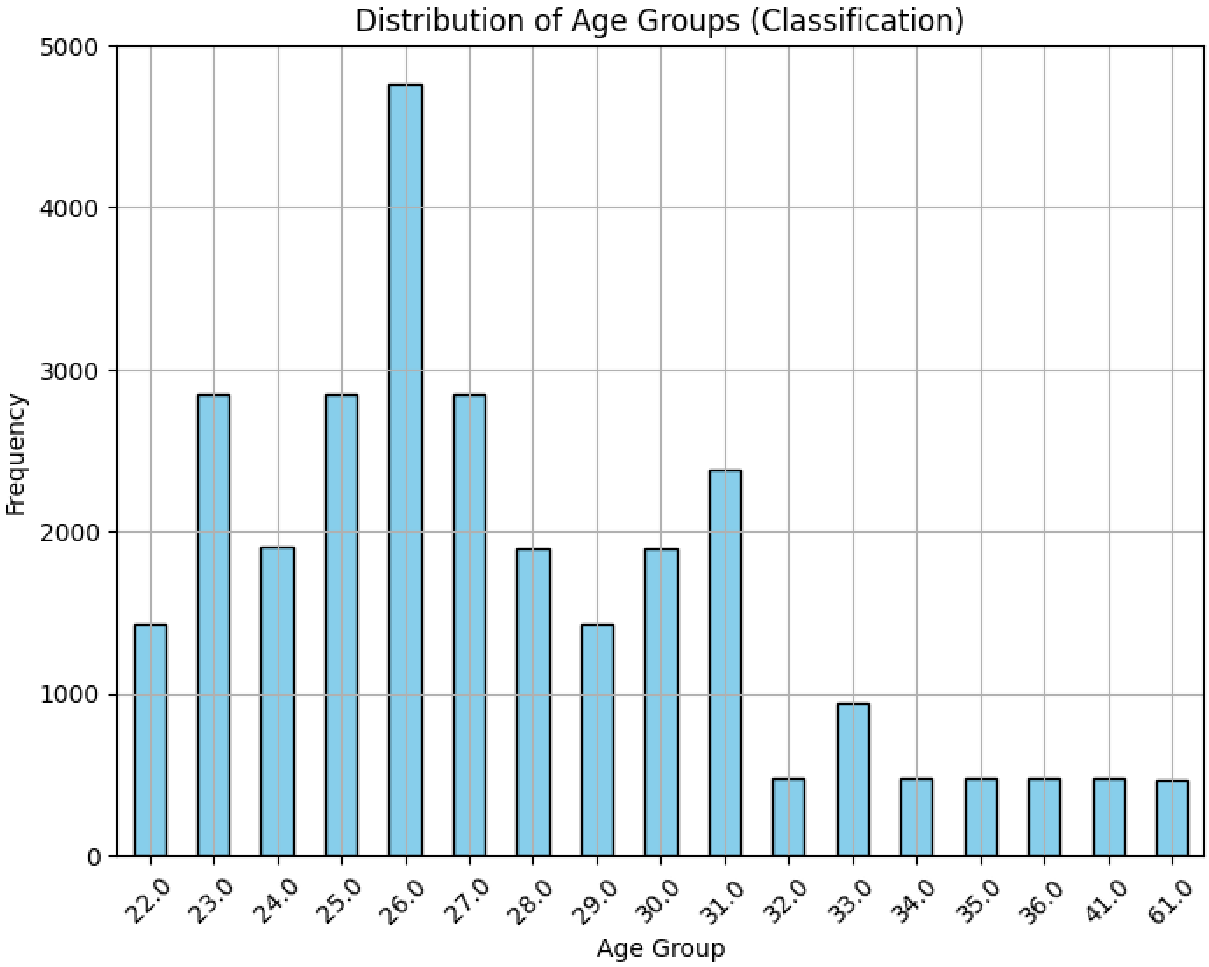

Fig. 2. Distribution of target Age with different classes of the target

TABLE II
MULTIPLE MODEL RESULTS ON LAYER 7 FOR AGE

Model	hyper-parameter	Validation Score
Logi. Reg. Clas.	1000 maximum iterations	78.94%
Random Forest	n_estimator=100	78.53%
SVM	kernel=RBF	92.26%
SVM	kernel=linear	85.46%
SVM	kernel=sigmoid	48.50%
SVM	kernel=poly, degree=2	88.04%
SVM	kernel=poly, degree=3	91.85%
SVM	kernel=poly, degree=4	92.80%
SVM	kernel=poly, degree=5	93.34%
SVM	kernel=poly, degree=6	93.75%
SVM	kernel=poly, degree=7	94.02%
SVM	kernel=poly, degree=8	93.48%
SVM	kernel=poly, degree=9	93.07%
XGBoost	softmax objective function	87.90%
Neural Network	-	81.66%

dropout rate of 0.1.

The output layer of the neural network comprises 17 neurons, aligning with the multi-class classification nature of the problem, where there are 17 possible classes. The 'softmax' activation function is employed in the output layer to compute class probabilities for each of the 17 classes. The choice of 'adam' optimizer, a widely-used optimization algorithm, facilitates the iterative updating of the model's weights during training, contributing to the model's overall learning process. SVM outperforms the Neural Network we built. Some fine tuning is needed to make the NN to perform even better.

3) Gender: Recognizing the imbalance in gender representation within the dataset, we initially trained the model with the provided data without modification. Subsequently, to address the imbalance and potentially enhance accuracy, we implemented oversampling techniques.

Our approach involved maintaining consistency in feature elimination procedures and subsequently constructing various models to glean insightful information about the data while also aiming for improved model performance. This iterative process allowed us to explore different facets of the dataset, ensuring a comprehensive understanding of its characteristics, and enabled the development of models that can better capture the nuanced relationships between features and the target variable, particularly in the context of gender classification.

Since this is a binary classification task we expected the logistic regression to perform better than other models in general.

TABLE III
Multiple Model results on Layer 7 for Age

Model	hyper-parameter	Validation Score
Logi. Reg. Clas.	1000 maximum iterations	99.87%
Random Forest	n_estimator=100	98.00%
SVM	kernel=RBF	100%
SVM	kernel=linear	99.87%
SVM	kernel=sigmoid	93.47%
SVM	kernel=poly, degree=2-7	99.87%
SVM	kernel=poly, degree=8,9	100%
XGBoost	softmax objective function	99.20%

Given the high likelihood of overfitting to the training and validation data when employing SVM with a polynomial kernel of high degree, a more prudent alternative is to explore the SVM RBF kernel. Particularly when coupled with oversampled data and subsequent feature elimination, this approach seeks to achieve optimal performance on validation data. Our feature elimination process, leveraging Lasso Regression, culminated in a refined set of 364 features from the initial 768. This meticulous feature selection aims to enhance the model's generalization capabilities and contribute to achieving a robust and accurate validation score with the SVM RBF kernel. The strategic combination of oversampling, feature elimination, and kernel selection reflects a nuanced and effective strategy for navigating the complexities of the dataset and optimizing model performance.

4) Accent: Drawing parallels to the methodology employed in gender classification, a comparable approach was undertaken for accent classification. The process involved oversampling, followed by feature elimination through Lasso Regression. The refined feature set was then utilized in conjunction with SVM RBF kernel, resulting in an outstanding validation score of 100%. This meticulous combination of oversampling, feature selection, and model choice demonstrates a systematic strategy for achieving optimal performance in accent classification, ensuring robustness and accuracy in capturing the intricacies of the dataset.

B. Final Prediction Model

1) Methodology: The ensemble technique, incorporating both Bagging and Boosting, proves to be a potent strategy for enhancing predictions [10]. Leveraging the distinct prediction sets from all six layers and the corresponding accuracy scores obtained for each layer-label combination, we harness this information to refine our final classification predictions for the test data.

Approaching each label as a standalone classification task, we utilize the six predictions generated from the different layers. Assigning scores to each predicted class based on the accuracy scores obtained during validation, we accumulate these scores for each class within a label. After considering all six predictions, we aggregate the scores for each class associated with a particular label. The final prediction for a given label is determined by selecting the class with the highest accumulated score. This meticulous scoring and selection process ensures a balanced and effective final prediction, leveraging the strengths of multiple layers to enhance the overall predictive performance.

Let $\mathcal{L}$ be the set of labels, $\mathcal{C}$ be the set of classes within a label, and $P_{k,j}$ be the prediction score for class j in label i obtained from layer k.

$$AccumulatedScore_{i,j} = \sum_{k=1}^{6} P_{k,j} \cdot Accuracy_{k,i}$$

Here:

- i represents the label.
- j represents the class within the label.
- $P_{k,j}$ is the prediction score for class j in label i obtained from layer k.
- $Accuracy_{k,i}$ is the accuracy score for predicting label i using layer k.

The final prediction for each label is determined by selecting the class with the highest accumulated score:

$$FinalPrediction_i = \arg\max_{j \in \mathcal{C}} AccumulatedScore_{i,j}$$

2) Evaluation: To assess our approach, we will compute the accuracy score for each label individually from the test data using models trained on each layer. Additionally, we will calculate the accuracy score of our ensembled prediction. Table IV depicts the accuracy for each label from an individual layer's model is determined by comparing the predicted labels with the true labels in the test data. Table V represents the overall accuracies of the ensembled prediction labels which computed by evaluating the correctness of the predicted labels with the labels of test data provided. We have excluded layer 12 during ensembling. We can observe that for the label *Age*, the ensembled model's performance is reduced than the best performing layer 7 and for the rest of the labels, its performance is very close to that of layer 7 model.

TABLE IV
Accuracies of single layer models

Label	layer 7	layer 8	layer 9	layer 10	layer 11
Speaker ID	96.67%	91.86%	94.67%	94.93%	91.2%
Age	93.82%	91.87%	86.56%	86.02%	74.73%
Gender	99.73%	99.87%	99.33%	99.46%	99.06%
Accent	98.13%	91.6%	96.80%	95.86%	87.47%

III. Conclusion

This study demonstrates the effectiveness of using pre-trained transformer neural networks, specifically the wav2vec model, in acquiring general representations for speech-related tasks with limited labeled data. By extracting layer-wise features and utilizing an ensemble technique, the proposed

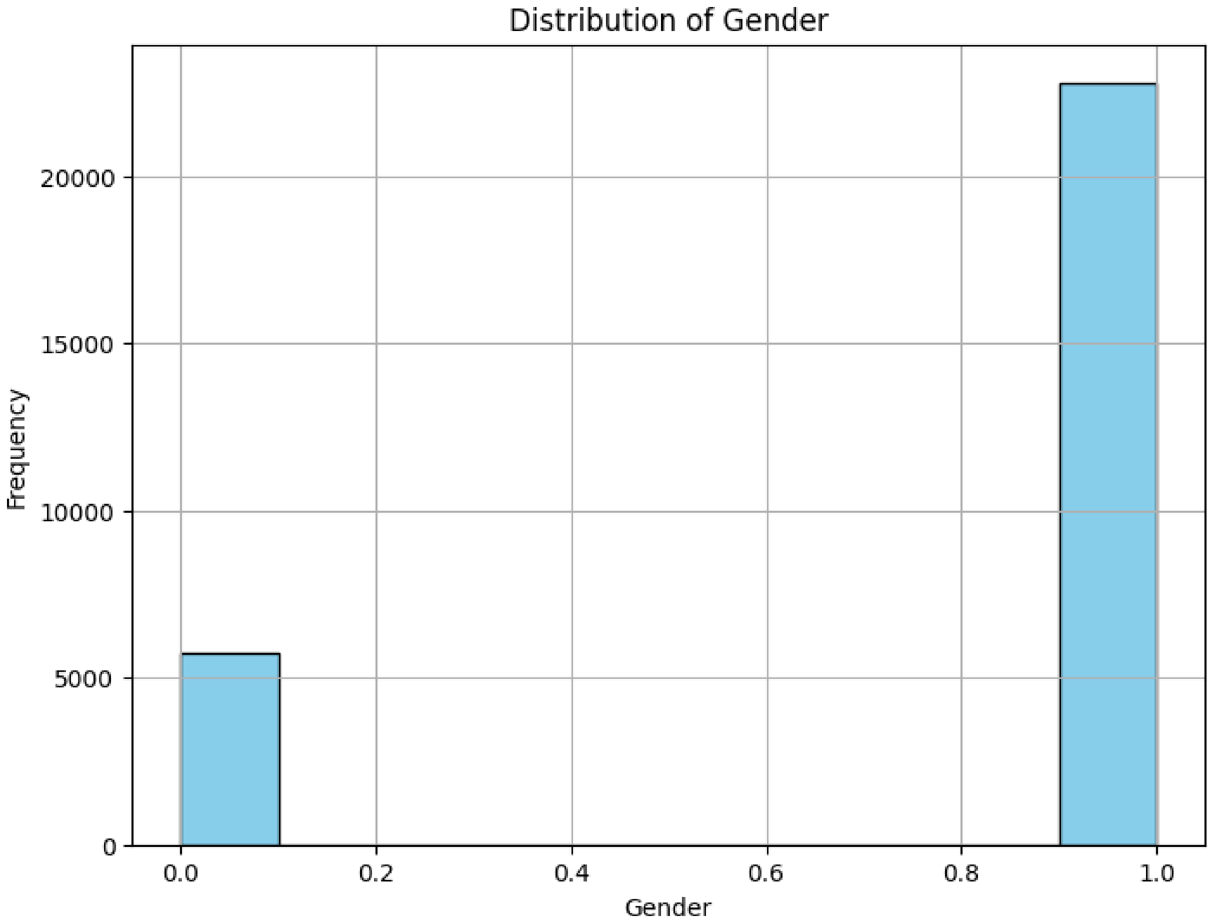

Fig. 3. Distribution of target Gender with target

TABLE V
ACCURACIES OF ENSEMBLED MODEL

Label	Ensempled model
Speaker ID	96.67%
Age	91.26%
Gender	99.73%
Accent	98.27%

approach achieved high accuracy in classifying age, gender, and speaker ID on the AudioMNIST dataset. The results suggest that this approach can be applied to other speech-related tasks with limited labeled data, providing a promising direction for future research

REFERENCES

[1] Schneider, S., Baevski, A., Collobert, R., and Auli, M., "wav2vec: Unsupervised Pre-training for Speech Recognition," *arXiv*, Apr. 2019.

[2] Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W., "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals," *arXiv*, Jul. 2018.

[3] REDDY, D. R., "Speech recognition by machine: A review," in *Readings in Speech Recognition*. Elsevier, 1990, pp. 8–38. [Online]. Available: https://doi.org/10.1016/b978-0-08-051584-7.50006-1

[4] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140

[5] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] Hershey, S., Chaudhuri, S., Ellis, D. P., 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)*. IEEE, 2017, pp. 131–135.

[7] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R., "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[8] Wang, Z., Lou, S., Liang, S., and Sheng, X., "Multi-class disturbance events recognition based on emd and xgboost in -otdr," *IEEE Access*, vol. 8, pp. 63 551–63 558, 2020.

[9] Liu, S., "Sentiment analysis of yelp reviews: a comparison of techniques and models," *arXiv preprint arXiv:2004.13851*, 2020.

[10] Sagi, O. and Rokach, L., "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.