2023 Seventh International Conference on Image Information Processing (ICIIP) | 979-8-3503-7140-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIIP61524.2023.10537679

Advancements in Environmental Sound Classification: Evaluating Machine Learning and Deep Learning Approaches on the UrbanSound8k

Priyanshu Malaviya
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, Gujarat 382007, India
priyanshumalaviya9210@gmail.com

Yogesh Kumar
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, Gujarat 382007, India
yogesh.kumar@sot.pdpu.ac.in

Nandini Modi
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, Gujarat 382007, India
Nandini.modi@sot.pdpu.ac.in

***Abstract*—In the rapidly evolving domain of audio classification, the quest for optimal model performance remains paramount. This research embarked on a comprehensive journey, juxtaposing traditional machine learning algorithms with state-of-the-art deep learning architectures, all benchmarked on the UrbanSound8k dataset. The study meticulously evaluated models based on pivotal metrics: Accuracy, Precision, Recall, and F1-score. Traditional models, including RandomForest and KNeighbors Classifier, showcased promising results, with the latter achieving a remarkable accuracy surge post hyperparameter tuning. However, the deep learning models, particularly the Artificial Neural Network (ANN), emerged as the zenith, registering an astounding accuracy of 97.59% after optimization. This paper not only underscores the prowess of deep learning in audio classification but also emphasizes the significance of hyperparameter refinement. The findings presented herein offer invaluable insights, setting the stage for future endeavors in the realm of audio data analysis.**

***Index Terms*—Audio classification, feature extraction, Hyperparameter Tuning, Comparative Analysis, Deep Neural Network, Machine learning**

I. Introduction

The domain of audio signal processing has undergone a notable shift towards environmental sound classification, with researchers aiming to construct models that are both efficient and accurate1. A significant method in this regard is the usage of Fully Convolutional Networks that employ parallel branching for feature extraction, combining the benefits of machine and deep learning [1]. Another notable step is the exploration of Transformer-encoder-based models, inspired by sequence classification in Natural Language Processing (NLP), which have exhibited amazing accuracy on datasets like UrbanSound8k [2]. These recent attempts show the promise of new architectures and approaches in increasing audio classification.

Autonomous vehicles (AVs) have further underlined the relevance of audio categorization, with auditory perception functioning as a complimentary tool to visual sensors like cameras, lidars, and radars [3]. The capacity of AVs to distinguish sounds such as sirens or car horns, even when these sources are not within the line-of-sight, is crucial for safety and navigation3. The UrbanSound8k dataset, with its various urban environment sounds, has been essential in training models specialized for such applications, stressing the real-world significance of breakthroughs in audio classification [3].

The fast growth of urban environments has led to an increasing interest in the classification of urban noises, with applications ranging from environmental monitoring to smart city planning. The UrbanSound8k dataset, a curated collection of urban sound recordings, has become a benchmark for academics in this subject. While various models have been presented for audio classification tasks, a full evaluation, especially with rigorous hyperparameter tuning, remains a key activity.

In this paper, we describe a comprehensive analysis of a range of machine learning and deep learning models applied to the UrbanSound8k dataset. On the traditional machine learning front, we dig into models such as the RandomForestClassifier, XGBClassifier, MLPClassifier, KNeighborsClassifier, SVC, GradientBoostingClassifier, and DecisionTreeClassifier. These models, noted for their versatility in handling multiple classification problems, are subjected to rigorous hyperparameter tuning using GridSearchCV paired with Stratified K-Fold cross-validation. This guarantees that each model is optimized for performance, taking into account the intricacies of the dataset.

On the deep learning spectrum, we investigate the capabilities of Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), Convolution 1D, and Artificial Neural Networks (ANN). Deep learning methods, with their inherent capacity to acquire hierarchical features, have showed promise in audio categorization problems. To refine these models, we apply the Keras Tuner, a powerful tool for hyperparameter tuning designed for deep learning architectures.

Feature extraction, a critical stage in audio categorization, is addressed utilizing two unique strategies. The first is the Mel-frequency cepstral coefficients (MFCC), a widely established

approach in audio and voice processing. The second, and probably more innovative, is the use of Yamnet for feature extraction, specifically when linked with LSTM and CNN models. Yamnet, a deep net trained on a massive dataset of YouTube videos, gives a novel perspective on feature extraction, potentially capturing subtle patterns in urban sounds.

The ultimate purpose of this research is to provide insights into the interplay between multiple models, hyperparameter tuning techniques, and feature extraction methods. By offering a holistic view of the audio classification challenge on the UrbanSound8k dataset, we seek to guide future research endeavors and contribute to the greater understanding of urban sound classification.

II. Related Work

The authors, Li, Song, and Hu, propose a method that combines the Waveform Similarity Overlap-Add (WSOLA) technique with Convolutional Neural Networks (CNNs). This combination is used to enhance audio data in categories with fewer samples, thereby improving classification accuracy and stability. The WSOLA technique is particularly effective in preventing the model from being biased towards categories with larger data volumes. The experimental results on the UrbanSound8K dataset show significant improvements in classification accuracy and stability, demonstrating the effectiveness of this approach in audio classification tasks [4].

Liu, Lu, Yuan, and Li introduce the Causal Audio Transformer (CAT), a novel framework for audio classification. CAT integrates a Multi-Resolution Multi-Feature (MRMF) feature extraction with an acoustic attention block, tailored for audio signals. The paper also proposes a causal module that enhances the model's interpretability and alleviates overfitting. CAT achieves state-of-the-art performance on several datasets, including UrbanSound8K. Its ability to generalize to other Transformer-based models is a significant advancement in the field of audio classification [5].

Tan et al. explored an alternative to the widely used Mel-Frequency Cepstral Coefficients (MFCC) for sound classification. They proposed the use of sound spectrum features, which retain audio features from the original audio file. Their experiments, particularly with Convolutional Neural Networks (CNN), showed that sound spectrum achieved comparable, if not better, results than its MFCC counterpart [6].

Kang, He, Wang, Peng, Qu, and Xiao present a novel approach in the realm of audio classification, focusing on Data-Free Knowledge Distillation (DFKD). Their work introduces the Feature-Rich Audio Model Inversion (FRAMI) framework, specifically designed for general sound classification tasks. This framework generates high-quality Mel-spectrograms through a feature-invariant contrastive loss and utilizes the hidden states before and after the statistics pooling layer for knowledge distillation. The experimental results on datasets like Urbansound8k demonstrate that FRAMI not only generates feature-rich samples but also significantly improves the accuracy of the student model compared to baseline methods. This advancement in DFKD showcases a promising direction for efficient and effective audio classification [7].

A content - Morsali et al. (2023) introduced a context-aware framework for feature selection and classification to achieve fast and accurate audio event annotation and classification. Their context-aware design explored various feature extraction techniques to determine an optimal combination for classification accuracy with minimal computational effort. The study also investigated the audio Tempo representation, a feature extraction method overlooked in previous environmental audio classification research. Their proposed algorithm for sound classification achieved an average prediction accuracy of 98.05% on the UrbanSound8K dataset [8].

Yunhao Chen and colleagues (2022) proposed an efficient network structure called Paired Inverse Pyramid Structure (PIP) and a network named Paired Inverse Pyramid Structure MLP Network (PIPMN). This lightweight architecture was designed to leverage the inherent characteristics of audio for efficient classification. The PIPMN achieved an Environmental Sound Classification (ESC) accuracy of 96% on the UrbanSound8K dataset and 93.2% Music Genre Classification (MGC) on the GTAZN dataset, all without the need for data augmentation or model transfer [9].

Ana Filipa Rodrigues Nogueira et al. (2022) carried performed a systematic study of the literature on sound classification and processing in urban settings. In the domain of audio classification, their review underscored the importance of Deep Learning (DL) architectures, attention mechanisms, data augmentation methodologies, and pretraining. The review stressed the efficacy of models such as DenseNet-161 with ImageNet pretrained weights, which achieved accuracies of 97.98%, 98.52%, and 99.22% for the UrbanSound8K, ESC-50, and ESC-10 datasets, respectively [10].

III. Problem Statement

The noises that surround us in our daily lives have developed in today's quickly urbanizing world, forming a mosaic of varied auditory experiences. These environmental noises create an audio portrayal of our surrounds, from the distant hum of traffic to the nearby chirping of birds. Accurately identifying these noises is more than simply an intellectual exercise; it has the ability to influence urban planning, health research, and even public safety. The UrbanSound8k dataset, a collection of urban ambient noises, offers a view into this world. The challenge, however, is in creating a robust model capable of detecting and categorizing these sounds with high precision, especially given the inherent complexities and variability in real-world audio data. The purpose is to navigate these complications and construct an effective audio classification system for the UrbanSound8k dataset that is both accurate and adaptive to the changing nature of urban soundscapes.

IV. Research Gap

Despite the plethora of studies dedicated to data classification techniques, there exists a noticeable void in comprehensive research that juxtaposes these methodologies under

uniform evaluation metrics. The majority of existing literature tends to zoom in on a singular method or a confined set of methods, often overlooking the potential benefits of integrating diverse techniques or embracing emerging ones. This scenario underscores the pressing need for holistic research endeavors that encompass a wide spectrum of classification tools, particularly in real-world contexts where data might not always be pristine or evenly distributed. Such expansive research would not only shed light on the optimal tools for varied scenarios but also highlight areas demanding innovation and refinement.

V. Research Hypothesis

Given the increasing complexity and diversity of urban sounds, it is postulated that a combination of traditional machine learning techniques and advanced deep learning methodologies can significantly enhance the accuracy and efficiency of urban sound classification. Specifically, when audio features are extracted using Mel-frequency cepstral coefficients (MFCC), algorithms such as Random Forest Classifier, XGB Classifier, MLP Classifier, among others, are anticipated to exhibit superior performance. Moreover, when leveraging Yamnet as a feature extractor, the convolutional and recurrent neural network models are hypothesized to capture intricate temporal and spectral patterns inherent in urban sounds.

A. Machine Learning algorithms used for audio classification

1) Random Forest: An ensemble learning technique, Random Forest Classifier constructs a 'forest' of decision trees during training. By aggregating the outputs of individual trees, it offers robustness against overfitting and often yields high classification accuracy, especially in datasets with diverse features like UrbanSound8k.

2) XGB Classifier: A gradient boosting framework that employs decision trees as base learners. XGBoost is renowned for its computational efficiency and capability to handle datasets with missing values, making it a suitable choice for complex audio datasets.

3) MLP Classifier: Multi-layer Perceptrons (MLP) are a class of feedforward artificial neural network. The MLP relies on an underlying Neural Network to perform the task of classification.

4) K-Nearest Neighbors Classifier: An instance-based learning or non-generalizing learning method. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

5) Support Vector Classifier: Support Vector Machines (SVC) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. They are effective in high dimensional spaces and are versatile as different Kernel functions can be specified for the decision function.

6) Gradient Boosting Classifier: An ensemble machine learning technique that builds on weak learners. It optimizes a loss function by adding trees sequentially, where each tree corrects the errors of its predecessor. Given its iterative nature, it's adept at minimizing errors and often yields high accuracy in diverse datasets.

7) Decision Tree Classifier: A non-parametric supervised learning method, DecisionTreeClassifier splits the dataset into subsets based on the most significant features, resulting in a tree-like model of decisions. Its transparent nature allows for easy visualization and interpretation of the decision-making process.

B. Deep Learning algorithms used for audio classification

1) LSTM: A specialized form of Recurrent Neural Networks (RNN), LSTM is designed to recognize patterns over time intervals. Given the temporal nature of audio signals, LSTM can effectively.

2) CNN: Primarily designed for image classification, Convolutional Neural Networks (CNN) have found utility in audio classification due to their ability to recognize spectral patterns in audio signals.

3) Convolution 1D: A variant of traditional CNNs, 1D CNNs are tailored for sequential data. By convolving over time steps, they can identify temporal patterns, making them apt for audio classification tasks.

4) ANN: Representing the cornerstone of deep learning, Artificial Neural Networks (ANN) consist of interconnected nodes or 'neurons'. Their ability to learn from vast amounts of data makes them particularly effective for tasks like audio classification.

By meticulously integrating these models with the feature extraction capabilities of MFCC and Yamnet, this research endeavors to derive a comprehensive understanding of the UrbanSound8k dataset. The ultimate objective is to elucidate the underlying patterns in urban sounds and establish a benchmark in audio classification methodologies.

VI. Research Methodology

The research is carried out in five main steps: dataset collection, feature engineering, preprocessing, classifier implementation, and result evaluation.

A. Dataset Collection

The primary source for this study was the UrbanSound8K dataset, which encompasses a diverse range of urban sounds, amounting to a total of 8,732 audio samples. These samples are systematically categorized into ten distinct classes, each representing a unique urban sound. The dataset comprises approximately 1,000 samples for each of the following categories: 'dog bark', 'children playing', 'air conditioner', 'street music', 'engine idling', 'jackhammer', and 'drilling'. However, some categories have a slightly different number of samples: 'siren' with 929 samples, 'car horn' with 429 samples, and 'gun shot' with 374 samples. Each audio sample is accompanied by two fundamental attributes: the audio file itself and its corresponding class label. Figure 1 illustrates the distribution of these sound categories in terms of their sample counts.

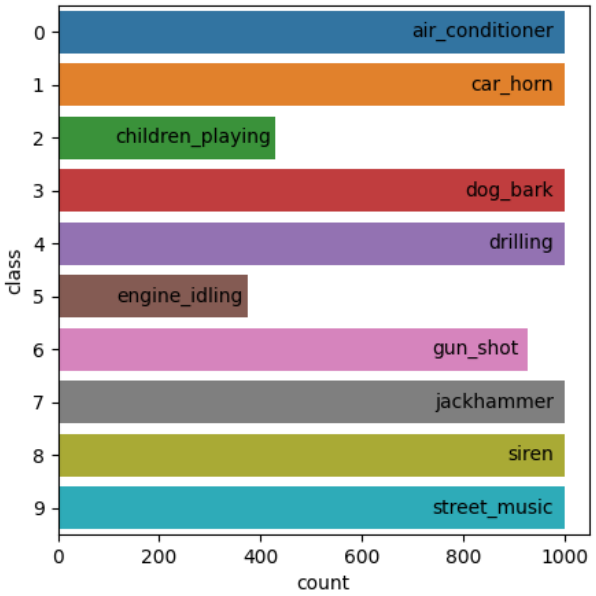

Fig. 1: Distribution of Audio samples

B. Audio Signal Processing

The librosa library, a python package for music and audio analysis, is employed to load and process each audio file. This step is crucial for subsequent feature extraction and analysis.

C. Data preprocessing and Feature Extraction from Audio

In the realm of audio signal processing, the quest for effective feature extraction techniques is perpetual. Two methodologies that have garnered attention are the Mel-Frequency Cepstral Coefficients (MFCCs) and the deep learning model, YAMNet. This paper seeks to elucidate these techniques, shedding light on their intricacies and applications.

1) Mel-Frequency Cepstral Coefficients (MFCCs): At the heart of many audio processing tasks lies the MFCC, a representation that captures the spectral properties of sound. Its significance stems from its ability to mirror the human ear's perception of frequency, making it a favorite among researchers and practitioners alike [11].

a) Extraction Process: The journey from a raw audio signal to MFCCs is a multi-step process [12]:

- Pre-emphasis: A filter amplifies the high frequencies, addressing the inherent imbalance in the frequency spectrum.
- Framing: The continuous signal is segmented into short frames, capitalizing on the stationary nature of frequencies over brief periods.
- Windowing: A Hamming window ensures smooth transitions at frame boundaries.
- Fourier Transformation: Each frame undergoes a transformation, revealing its frequency components.
- Mel Filter Bank Processing: The power spectrum is transformed, aligning with the Mel scale's human-centric frequency perception.
- Discrete Cosine Transformation: A return journey to the time domain yields the coveted MFCCs.

2) YAMNet: A Foray into Deep Learning for Audio: YAMNet is a model trained to discern audio events. Its prowess isn't limited to its primary task; it doubles as a feature extractor, making it a valuable tool in the audio processing toolkit [13]. YAMNet's foundation is a convolutional neural network, a deep architecture adept at discerning patterns in data. The model's modus operandi is as follows:

- Input Processing: Raw audio is the starting point.
- Convolutional Layers: These layers tease out patterns, moving from simple to complex.
- Log-Mel Spectrogram Creation: A time-frequency representation emerges, offering insights into the audio's characteristics.
- Embedding Derivation: The model's depth and training enable the extraction of embeddings, encapsulating the essence of the audio.

D. Classifiers

Within the machine learning discipline, foundational algorithms have paved the way for the development of advanced models. For instance, K-Nearest Neighbors (KNN) operates on the concept of similarity, classifying an unlabeled sample based on the predominant class among its 'k' closest data points. Contrarily, Decision Trees utilize a structured decision-making process to categorize data. Enhancing this approach, Random Forests amalgamate multiple decision trees to derive a collective decision, enhancing prediction accuracy and curbing overfitting tendencies. Support Vector Machines (SVM) are designed to pinpoint the most suitable hyperplane that distinctly segregates classes in a dataset. Evaluative metrics such as accuracy, precision, and recall are instrumental in determining the efficacy of these classifiers. Accuracy reflects the ratio of accurate predictions, precision underscores the validity of positive predictions, while recall emphasizes the classifier's proficiency in recognizing all pertinent samples.

Deep learning, an advanced branch of machine learning, focuses on intricate data representation. Specifically, Convolutional Neural Networks (CNN) excel in analyzing visual data, leveraging convolutional layers for local pattern identification and pooling layers for spatial dimensionality reduction. In contrast, Artificial Neural Networks (ANN) are structured with interconnected nodes, often termed "neurons," which facilitate data processing through layered architectures. Another noteworthy architecture is the Long Short-Term Memory (LSTM) network, a specialized variant of Recurrent Neural Networks (RNN). LSTMs are tailored to discern patterns across temporal sequences, rendering them apt for tasks involving sequential datasets, such as time series analysis or linguistic processing.

E. Model validation

Model validation is a pivotal step in the machine learning pipeline. It assesses the model's performance on unseen data, ensuring its generalizability. Stratified K-Fold cross-validation is a preferred technique, especially when dealing with imbalanced datasets. By maintaining the original class distribution in each fold, it ensures that every class is adequately represented, leading to a more reliable and unbiased evaluation.

F. Hyperparameter Optimization

Hyperparameter optimization stands as a cornerstone in the realm of machine learning. It's the art and science of refining model parameters to extract the best performance for a specific task. Let's delve into two prominent techniques: GridSearchCV combined with Stratified Cross-Validation and the Keras Tuner.

1) GridSearchCV: GridSearchCV is an optimization tool in machine learning that systematically traverses through a specified set of hyperparameters to determine the best combination for a given model. It operates by partitioning the data multiple times into training and validation sets, a process known as cross-validation. This iterative evaluation ensures that the model's performance is assessed comprehensively across different data splits. A specialized form, Stratified Cross-Validation, is employed when dealing with imbalanced datasets. It ensures that each fold retains the original class distribution, thereby providing a more representative evaluation. By leveraging GridSearchCV with stratified sampling, researchers can achieve a more accurate and unbiased hyperparameter tuning, enhancing the model's generalization capabilities [14].

2) Keras Tuner: Keras Tuner represents a dedicated solution for hyperparameter optimization within the Keras deep learning framework. This tool alleviates the challenges of manual hyperparameter selection by automating the exploration of optimal configurations. Through sophisticated methodologies, including Bayesian Optimization, Hyperband, and Random Search, Keras Tuner systematically evaluates various hyperparameter combinations. Each technique provides a distinct strategy for navigating the hyperparameter landscape, striking a balance between broad exploration and focused exploitation. By incorporating Keras Tuner, researchers can enhance the efficiency of their model development, ensuring optimal neural network configurations with reduced manual oversight [15].

VII. Results and Discussion

To rigorously assess the performance of classification models, a set of evaluative criteria was established:

- True Positive (TP): Instances where the positive classification was accurate.
- False Positive (FP): Instances where the classification was positive, but inaccurately so.
- True Negative (TN): Instances where the negative classification was accurate.
- False Negative (FN): Instances where the classification was negative, but inaccurately so.
- Accuracy: Denotes the fraction of all predictions that are correct.
- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
- Precision: Represents the fraction of correctly predicted positive observations to the total predicted positives.
- Precision = $\frac{TP}{TP+FP}$
- Recall (or Sensitivity): Measures the fraction of actual positives that were correctly classified.
- Recall = $\frac{TP}{TP+FN}$
- F1-score: A metric that harmonizes precision and recall, providing a comprehensive view of model performance.
- F1 = $2 \times \frac{Precision \times Recall}{Precision+Recall}$

Using these metrics as benchmarks, a range of models were trained and subsequently evaluated to determine their performance capabilities.

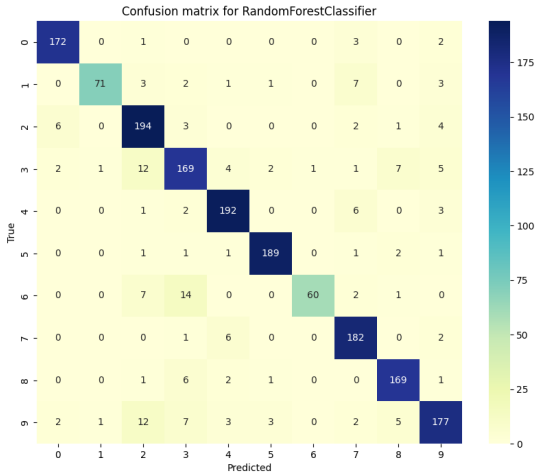

Fig. 2: Confusion matrix for Random Forest

Model Name	Accuracy Score	Precision	Recall	F1-Score
RandomForest Classifier	90.04%	0.91	0.89	0.90
XGBClassifier	89.64%	0.91	0.89	0.90
MLPClassifier	88.55%	0.89	0.88	0.89
KNeighbors Classifier	88.09%	0.89	0.88	0.88
SVC	86.43%	0.88	0.86	0.87
GradientBoostingClassifier	80.88%	0.83	0.80	0.81

TABLE I: Classification Report of Machine Learning Models

Model Name	Accuracy Score	Precision	Recall	F1-Score
KNeighbors Classifier	91.87%	0.92	0.92	0.92
DecisionTreeClassifier	69.16%	0.69	0.69	0.69
RandomForestClassifier	90.04%	0.92	0.89	0.90
SVC	92.76%	0.93	0.93	0.93
MLPClassifier	88.19%	0.89	0.89	0.89
XGBClassifier	89.48%	0.91	0.89	0.90

TABLE II: Classification Report of Machine Learning models after Hyperparameter Tuning

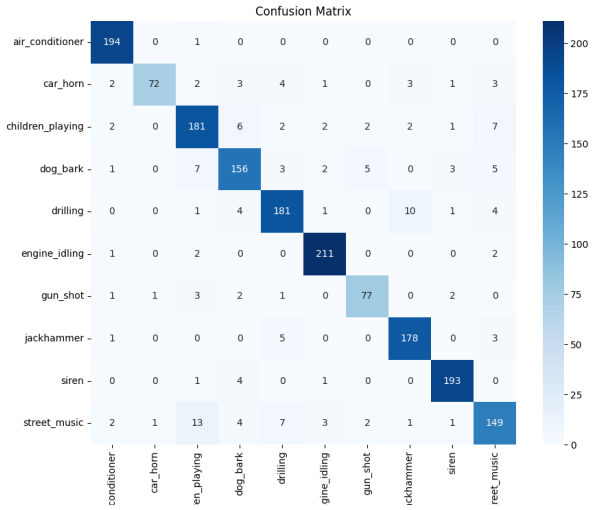

Fig. 3: Confusion matrix for ANN

Model Name	Accuracy Score	Precision	Recall	F1-Score
CNN	87.98%	0.90	0.86	0.88
LSTM	83.57%	0.85	0.83	0.84
ANN	86.92%	0.95	0.95	0.95
Convolution 1D	87.92%	0.90	0.87	0.89

TABLE III: Classification Report of Deep Learning Models

Model Name	Accuracy Score	Precision	Recall	F1-Score
CNN	96.69%	0.98	0.96	0.97
LSTM	96.18%	0.96	0.96	0.96
ANN	97.59%	0.98	0.98	0.98
Convolution 1D	96.99%	0.98	0.96	0.97

TABLE IV: Classification Report of Deep Learning Models after Hyperparameter Tuning

A. Models with Yamnet as a feature extraction

Model Name	Accuracy Score	Precision	Recall	F1-Score
CNN	91.52%	0.91	0.91	0.91
LSTM	97.25%	0.98	0.97	0.97

TABLE V: Classification Report of Deep Learning Models with YAMNET as a Feature Extractor

VIII. Conclusion

In our comprehensive exploration of audio classification, we delved into both traditional machine learning and advanced deep learning techniques. The evaluation was grounded on four key metrics: Accuracy, Precision, Recall, and F1-score. Among traditional models, the RandomForest Classifier stood out with a commendable accuracy of 90.04%. Interestingly, after fine-tuning the parameters, the KNeighbors Classifier exhibited a marked improvement, achieving an accuracy of 91.87%. However, the DecisionTreeClassifier remained consistent, hovering around the 69% mark both pre and post-tuning. Transitioning to the deep learning realm, the initial results were promising with the ANN model registering an accuracy of 86.92% and the CNN closely tailing at 87.98%. The true prowess of deep learning was unveiled post hyperparameter optimization. The ANN model's accuracy soared to 97.59%, making it the crown jewel of our study. In conclusion, this research illuminates the capabilities of both traditional and deep learning models in audio classification. The standout performance of the ANN model, especially after meticulous tuning, underscores the significance of refining model parameters. As the journey in audio classification continues, the findings from this study will undoubtedly serve as a beacon, directing future research and innovations in the field.

References

[1] M. Neri, L. Pallotta, and M. Carli, "Low-complexity environmental sound classification using cadence frequency diagram and chebychev moments," in *2023 International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2023, pp. 1–6.

[2] S. Karam, S.-J. Ruan, and Q. M. u. Haq, "Task incremental learning with static memory for audio classification without catastrophic interference," *IEEE Consumer Electronics Magazine*, vol. 11, no. 5, pp. 101–108, 2022.

[3] F. Walden, S. Dasgupta, M. Rahman, and M. Islam, "Improving the environmental perception of autonomous vehicles using deep learning-based audio classification," 2022.

[4] P. Li, T. Song, and J. Hu, "Audio classification based on audio wsola and cnn algorithm," in *Second International Conference on Electronic Information Technology (EIT 2023)*, vol. 12719. SPIE, 2023, pp. 1145–1149.

[5] X. Liu, H. Lu, J. Yuan, and X. Li, "Cat: Causal audio transformer for audio classification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] K. In Tan, S. Yean, and B. Sung Lee, "Sound classification using sound spectrum features and convolutional neural networks," in *2022 3rd International Conference on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 2022, pp. 94–99.

[7] Z. Kang, Y. He, J. Wang, J. Peng, X. Qu, and J. Xiao, "Feature-rich audio model inversion for data-free knowledge distillation towards general sound classification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] M. M. Morsali, H. Mohammadzade, and S. B. Shouraki, "Face: Fast, accurate and context-aware audio annotation and classification," 2023.

[9] Y. Chen, Y. Zhu, Z. Yan, Y. Huang, Z. Ren, J. Shen, and L. Chen, "Effective audio classification network based on paired inverse pyramid structure and dense mlp block," 2023.

[10] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Sound classification and processing of urban environments: A systematic literature review," *Sensors*, vol. 22, no. 22, p. 8608, 2022.

[11] L. Pallotta, M. Neri, M. Buongiorno, A. Neri, and G. Giunta, "A machine learning-based approach for audio signals classification using chebychev moments and mel-coefficients," in *2022 7th International Conference on Frontiers of Signal Processing (ICFSP)*, 2022, pp. 120–124.

[12] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2020.

[13] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, 2021. [Online]. Available: https://www.mdpi.com/2224-2708/10/4/72

[14] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, and Imdadullah, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv," *IEEE Access*, vol. 10, pp. 80 151–80 173, 2022.

[15] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "Kerastuner," https://github.com/keras-team/keras-tuner, 2019.