Assignment 2 Report

Buhle Mlandu(MLNHON001) and Lutho Mngqibisa(MNGLUT008)

Cape Town, October 28, 2025

Abstract

This project explores the application of deep learning for urban sound classification using the UrbanSound8K dataset, which contains 8,732 audio samples across 10 everyday sound categories. The task is formulated as a multi-class classification problem aimed at automatically identifying environmental sounds such as car horns, dog barks, and sirens. A Naïve Bayes classifier is implemented as a baseline to establish benchmark performance, followed by the development of an Artificial Neural Network (ANN) trained on Mel-Frequency Cepstral Coefficient (MFCC), Root Mean Square (RMS) Energy, Chroma, and Zero-Crossing Rate (ZCR) features. The ANN is optimized through hyperparameter tuning and evaluated using accuracy and F1-score metrics. Comparative analysis demonstrates that the neural model significantly outperforms the baseline, indicating its superior capacity for learning complex audio patterns. The study highlights the practical potential of deep learning in acoustic scene analysis while addressing challenges related to noise, class imbalance, and generalization in real-world audio environments.

Contents

1 Introduction and Problem Formulation

1.1 Dataset Description

This project uses the UrbanSound8K dataset [1], a benchmark collection for environmental sound classification. It contains 8,732 labeled audio clips, each 4 seconds long, drawn from 10 urban sound categories: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. The dataset was curated from real-world field recordings obtained via the Freesound platform, organized into 10 predefined folds to support reproducible cross-validation. Each audio sample is annotated with its corresponding class label and acoustic salience (foreground or background), making it suitable for training and evaluating machine learning models on realistic, noisy sound data.

1.2 Problem Definition

The task is formulated as a supervised multi-class classification problem, where the goal is to predict the correct sound category from the 10 possible classes, given an input audio sample. The target variable is the categorical sound label, and the input consists of extracted audio features, specifically Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Root Mean Square (RMS) Energy, and Zero-Crossing Rates(ZCR), with both mean and standard deviation calculated for each. Mathematically, the problem can be expressed as learning a mapping function

$$f : R^n \rightarrow \{1, 2, \ldots, 10\}$$

where n represents the dimensionality of the combined feature vector, and the output is the predicted class index.

The project compares two approaches:

- A **Naïve Bayes classifier** as a probabilistic baseline model
- An **Artificial Neural Network (ANN)** as a deep learning model trained on the combined features to capture nonlinear feature interactions.

1.3 Input and Output Specification

- **Inputs:** Extracted MFCC, Chroma, RMS and ZCR features combined into a vector from 4-second audio segments. Each feature vector captures key spectral and temporal characteristics of the sound.

- **Outputs:** A discrete label corresponding to one of the 10 predefined urban sound classes. The models are trained and evaluated using accuracy and F1-score as performance metrics.

1.4 Motivation and Usefulness

Automatic urban sound classification has diverse real-world applications in smart cities, environmental monitoring, public safety, and urban planning [2]. For instance, systems capable of distinguishing sirens or gunshots can enhance emergency response systems, while detecting construction or traffic noise can aid in noise pollution analysis. From a research perspective, the UrbanSound8K dataset offers a valuable testbed for evaluating the performance of machine learning and deep learning models on complex, noisy, and overlapping acoustic environments.

1.5 Ethical Consideration

While urban sound classification can improve city livability and public safety, ethical considerations must be addressed. The deployment of such systems in surveillance contexts could raise privacy concerns if combined with geolocation or personal data. In addition, biases in the dataset, such as the overrepresentation of certain environments or sound sources, may affect model fairness and generalization. It is therefore essential to ensure transparent data collection and responsible use of trained models in real-world systems.

2 Baseline Model

2.1 Choice of Baseline

For this study, a Gaussian Naïve Bayes classifier was selected as the baseline model. Naïve Bayes is a probabilistic classifier based on Bayes' theorem, which assumes conditional independence among features given the class label. It is computationally efficient, interpretable, and often used as a first benchmark in audio classification tasks. This makes it a suitable starting point for comparison against more complex neural network architectures.

2.2 Rationale

The Naïve Bayes classifier is an appropriate minimal model for this problem for several reasons. Firstly, it performs robustly on moderately high-dimensional, continuous features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma, root mean square (RMS) energy, and zero-crossing rate (ZCR), which are standard descriptors in environmental and speech sound analysis [1]. Secondly, it provides a clear performance baseline against which the benefit of deep learning models can be measured. Finally, due to its fast training speed and low computational overhead, it enables rapid experimentation across multiple feature configurations and cross-validation folds.

2.3 Implementation

The baseline model was implemented in **Python 3.10** using **scikit-learn's** `GaussianNB()` classifier. Feature extraction was conducted using the **Librosa** library, generating per-clip acoustic features that included Mel-Frequency Cepstral Coefficients (MFCCs), chroma features,

Root Mean Square (RMS) energy, and Zero-Crossing Rate (ZCR). For each feature type, both the mean and standard deviation statistics across time frames were computed to summarize the temporal characteristics of each 4-second clip.

- MFCCs with both mean and standard deviation statistics,
- 12-dimensional chroma features (mean and std),
- Root Mean Square (RMS) energy (mean and std),
- Zero-Crossing Rate (ZCR) (mean and std).

Each UrbanSound8K audio clip (4 seconds, 22.05 kHz) therefore yielded a feature vector of varying length depending on the number of MFCCs used. Four configurations were tested by varying the MFCC count, $n_{mfcc} \in \{13, 20, 30, 40\}$, and each configuration was evaluated using **10-fold cross-validation** across all 8,732 samples.

The highest overall performance was achieved with $n_{mfcc} = 40$, which produced the best cross-validated mean test accuracy.

2.4 Feature Selection Rationale

The chosen features are standard and widely validated in environmental sound recognition research due to their ability to capture complementary aspects of the acoustic signal:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** Capture the spectral envelope of the audio signal by mimicking the human auditory system's logarithmic frequency perception. MFCCs effectively represent timbral texture and are highly discriminative for different sound sources.
- **Chroma Features:** Encode the distribution of spectral energy across 12 pitch classes, allowing the model to capture harmonic and tonal content, which is useful for sounds involving musical or periodic structure (e.g., street music).
- **Root Mean Square (RMS) Energy:** Measures the overall signal energy and correlates with sound loudness, helping differentiate between high- and low-energy events (e.g., jackhammer vs. air conditioner).
- **Zero-Crossing Rate (ZCR):** Quantifies the rate at which the waveform changes sign, capturing temporal noisiness and high-frequency content. It is particularly effective in identifying transient or impulsive sounds such as gunshots and drilling.

Together, these features provide a balanced representation of spectral, temporal, and harmonic information, making them well-suited for broad environmental sound classification.

Table 1: Summary of Naïve Bayes baseline configuration

Parameter	**Description / Value**
n_{mfcc}	40
Total Features	108 (40 MFCC [mean,std] + RMS + ZCR + 12 Chroma [mean,std])
Cross-Validation Folds	10
Mean Accuracy	**53.73% ± 5.90%**
Accuracy Range	44.99% – 61.94%

2.5 Results

Table 2: Final Naïve Bayes Baseline performance metrics (10-fold aggregated results).

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Air Conditioner	0.332	0.172	0.227	1000
Car Horn	0.644	0.620	0.632	429
Children Playing	0.535	0.642	0.583	1000
Dog Bark	0.702	0.611	0.653	1000
Drilling	0.395	0.426	0.410	1000
Engine Idling	0.505	0.591	0.545	1000
Gun Shot	0.574	0.869	0.691	374
Jackhammer	0.469	0.816	0.596	1000
Siren	0.686	0.365	0.476	929
Street Music	0.734	0.501	0.595	1000
Overall Accuracy		**0.537 ± 0.0590**		
Macro Avg	0.558	0.561	0.541	8732
Weighted Avg	0.550	0.537	0.525	8732

Table 3: Performance comparison between baseline and random guessing

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**Improvement (×)**
Random Guess	10.00	0.10	0.10	0.10	—
Naïve Bayes (Baseline)	**53.73 ± 5.90**	0.558	0.561	0.541	**5.37×**

The Naïve Bayes classifier achieved a mean test accuracy of **53.73% ± 5.90%** across the 10 folds, with a macro-average F1-score of 0.541. This result demonstrates that even a simple probabilistic model can capture meaningful acoustic distinctions across the ten urban sound classes, far outperforming random guessing (10% expected accuracy).

2.6 Visualisation and Error Analysis

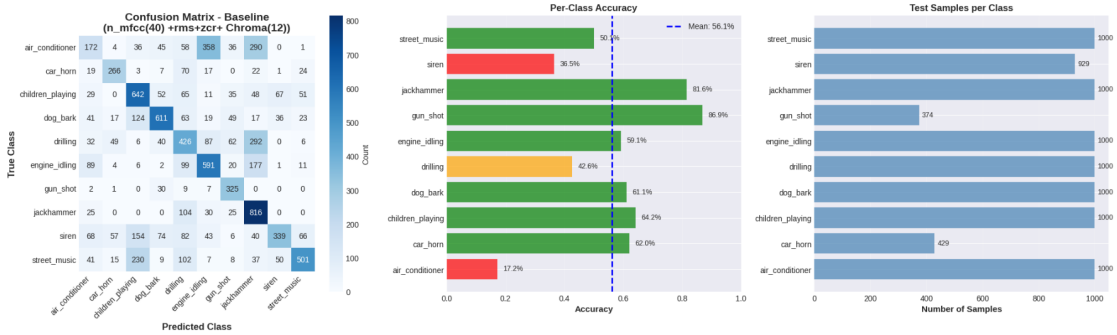

Figure 1: Baseline Naïve Bayes performance visualisation: confusion matrix (left), per-class accuracy (middle), and test sample distribution (right).

As shown in Figure 1, the Naïve Bayes model performs well on impulsive and acoustically distinct classes such as *gun_shot* (86.9%) and *jackhammer* (81.6%), while continuous low-frequency sounds such as *air_conditioner* (17.2%) and *siren* (36.5%) exhibit higher misclassification rates. The most frequent confusion pairs include *air_conditioner* → *engine_idling* (35.8%) and *drilling* → *jackhammer* (29.2%), which reflect overlapping spectral characteristics. These findings confirm that MFCC- and chroma-based representations capture timbral features effectively but fail to model longer-term temporal dependencies.

2.7 Summary

The Gaussian Naïve Bayes classifier established a reliable baseline with an overall mean accuracy of **53.7%**. Although it cannot model correlations between features or temporal dynamics, it exceeds random chance by more than fivefold, serving as a solid benchmark for evaluating the neural network models developed in subsequent sections.

3 Neural Network Model Design

3.1 Data Preprocessing and Input Representation

The same UrbanSound8K dataset was used for the neural network model, consisting of 8,732 labelled 4-second audio clips across ten urban sound categories (*air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music*). Each audio sample was processed at a sampling rate of 22,050 Hz and converted into structured numerical features suitable for a feedforward neural network.

To obtain a compact yet informative feature representation, the `Librosa` library was used to extract Mel-Frequency Cepstral Coefficients (MFCCs) and additional low-level acoustic descriptors. These features capture the spectral and temporal properties of sound that are perceptually meaningful to the human auditory system. Specifically, the extracted features included:

- **MFCCs:** 120 coefficients summarised by mean and standard deviation. (Found via experimentation)

- **Root Mean Square (RMS) Energy:** mean and standard deviation (2 features)
- **Zero-Crossing Rate (ZCR):** mean and standard deviation (2 features)
- **Chroma Features:** 12 coefficients summarised by mean and standard deviation (24 features)

A **268-dimensional vector** represented each input sample, and the model output corresponded to ten categorical sound classes.

3.2 Experimental Setup and Model Validation

3.2.1 Data Splitting

The UrbanSound8K dataset is organized into ten predefined folds, as specified by Salamon *et al.* (2014). To ensure consistency with the original protocol and prevent data leakage between training and testing subsets, we adopted a **10-fold cross-validation** strategy. For each experiment, one fold was used as the **test set**, another as the **validation set**, and the remaining eight folds served as the **training set**. This corresponds approximately to a split ratio of 80% training, 10% validation, and 10% testing. Each audio clip's fold assignment was respected throughout, ensuring that no recording appeared in multiple subsets. All features were standardized using a `StandardScaler`, fitted on the training data and applied to validation and test sets.

3.2.2 Extracted Features

The number of MFCC coefficients was manually varied and tested, specifically $n_{mfcc} \in \{13, 20, 30, 40, 60, 80, 120\}$. This process determined that $n_{mfcc=120}$ yielded the most informative feature vector, resulting in a 268-dimensional input (MFCC, Chroma, RMS, and ZCR combined).

3.2.3 Hyperparameter Tuning Process

A structured hyperparameter tuning process was conducted to determine the optimal neural network configuration. The tuning was performed systematically in several stages, where only one parameter group was varied at a time, while the previously determined best values were retained for subsequent experiments. All experiments consistently used **fold 10 as the test set**, **fold 1 as the validation set**, and the remaining folds (**2**–**9**) as the training set, ensuring consistency and reproducibility across runs.

The following search procedure was applied sequentially:

1. **Architecture Search:** Five architectures were evaluated:
   - Small (2-layer): [500, 250]
   - Medium (3-layer): [1000, 500, 250]
   - Large (5-layer): [1000, 750, 500, 250, 100]
   - Narrow (5-layer): [512, 256, 128, 64, 32]

- Deep (6-layer): [1000, 800, 600, 400, 200, 100]

The **Medium (3-layer)** configuration achieved the highest validation accuracy and was adopted for subsequent experiments.

2. **Learning Rate Search:** With the best-performing architecture fixed, learning rates in $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}\}$ were evaluated. The optimal value was **0.0005**, balancing convergence stability and validation performance.

3. **Dropout Rate Search:** Dropout rates $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ were tested. The best-performing rate was **0.4**, providing improved generalization without underfitting.

4. **Batch Size Search:** Batch sizes $\{100, 200, 300, 400, 500, 600, 700, 800\}$ were evaluated. A batch size of **500** yielded the best validation accuracy and most stable convergence curves.

Each experiment was trained for a maximum of **400 epochs** using the **Adam optimizer** and **categorical cross-entropy loss**, with an **early stopping patience of 50 epochs**. The model parameters from the epoch with the highest validation accuracy were saved for evaluation. Training and validation accuracy were tracked for every epoch to monitor convergence behavior.

The best configuration from the tuning process was:

Architecture: [1000, 500, 250]; Learning rate = 0.0005; Dropout = 0.4; Batch size = 500.

Table 4 reports representative results from the tuning experiments (highest validation accuracies per stage are shown).

Table 4: Representative best validation accuracies per hyperparameter tuning stage.

Experiment Stage	Best Configuration	Validation Accuracy (%)
Architecture Search	[1000, 500, 250]	69.07
Learning Rate Search	lr = 0.0005	68.61
Dropout Rate Search	dropout = 0.4	66.78
Batch Size Search	batch = 500	67.70

3.2.4 Training Procedure

All neural network models were implemented using the `PyTorch` framework and trained with the **Adam optimizer** (learning rate = 0.0005) and **categorical cross-entropy loss function**. Each training session ran for up to **400 epochs** with **early stopping** applied (patience = 50 epochs) to mitigate overfitting. The final configuration used a dropout rate of 0.4 between fully connected layers and a batch size of 500.

Throughout training, both loss and accuracy curves exhibited consistent convergence behavior. Training accuracy increased steadily, while validation accuracy stabilized after approximately 150–200 epochs, indicating effective optimization and minimal overfitting. These patterns were consistent across all folds and hyperparameter experiments.

After selecting the optimal configuration, a full **10-fold cross-validation** was performed using the same setup. The protocol ensured that each of the ten folds served exactly once as the test set. The validation set was dynamically assigned to the subsequent fold (e.g Fold 10 was the validation set when Fold 9 was the test set), and the remaining eight folds were used for training. The mean cross-validated **test accuracy** was **69.28% ± 4.78%**, confirming reliable generalization across folds. Training and validation accuracy trends for all folds were monitored to ensure consistent convergence and stability.

Overall, this staged experimental design provided a rigorous, transparent tuning and validation protocol that aligns with best practices for neural network experimentation on the UrbanSound8K dataset.

3.3 Model Architecture

The final selected network was a fully connected feedforward **Multilayer Perceptron (MLP)** with the following configuration:

- **Input Layer:** 268 neurons (corresponding to extracted features)
- **Hidden Layers:** three dense layers with 1000, 500, and 250 units respectively
- **Output Layer:** 10 neurons with softmax activation (multi-class classification)
- **Activation Function:** ReLU for all hidden layers
- **Dropout:** 0.4 for regularisation
- **Loss Function:** Cross-Entropy Loss
- **Optimizer:** Adam with a learning rate of 0.0005
- **Batch Size:** 500
- **Training Epochs:** 400 with early stopping (patience = 50)
- **Compute Device:** NVIDIA Tesla T4 GPU

The final network architecture can be expressed as:

$$\text{ANN} = [1000,\ 500,\ 250], \quad \text{Dropout} = 0.4, \quad \eta = 0.0005$$

3.4 Design Justification

The choice of a Multilayer Perceptron (MLP) architecture was guided by the numerical and tabular nature of the extracted features. MFCC and chroma-based features encapsulate key spectral-temporal sound information in fixed-length vectors, making a fully connected architecture more suitable than convolutional networks, which are optimised for two-dimensional spatial patterns.

The three hidden layers ($[1000, 500, 250]$) offered the optimal trade-off between model expressiveness and generalisation. Shallower networks exhibited underfitting, while deeper

ones led to diminishing returns and a higher tendency to overfit. A dropout rate of 0.4 and early stopping strategy effectively mitigated overfitting, ensuring stable convergence across folds.

The final configuration achieved a mean 10-fold cross-validation accuracy of **69.28% ± 4.78%**, representing an improvement of approximately **15.55 percentage points** over the Naïve Bayes baseline (53.73%). This substantial enhancement highlights the neural model's ability to capture nonlinear correlations among MFCC-based acoustic features that linear probabilistic methods fail to represent adequately.

3.5 Visualised Results

Figure 2 presents a combined visualisation of the neural network's performance. The left panel displays the confusion matrix, revealing strong diagonal dominance across most classes, indicating effective inter-class discrimination. The centre panel depicts per-class accuracies, where high-energy transient sounds such as *gun shot*, *siren*, and *street music* achieved accuracies above 80%, whereas continuous low-frequency sounds such as *air conditioner* and *engine idling* remained more challenging to distinguish. The right panel shows the distribution of test samples per class, confirming balanced evaluation splits across categories.

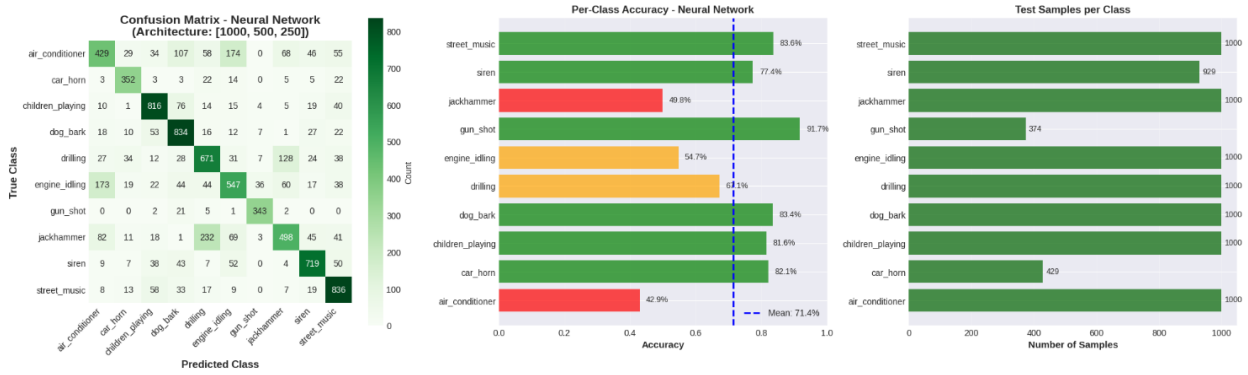

Figure 2: Combined visualisation of Neural Network model performance showing (left) the confusion matrix, (centre) per-class accuracy with mean accuracy line, and (right) the distribution of test samples per class. Model architecture: [1000, 500, 250].

In summary, the neural network demonstrated strong generalisation across urban sound classes and a consistent performance advantage over the Naïve Bayes baseline. These results confirm that a well-regularised MLP trained on MFCC-based feature representations provides an accurate and computationally efficient solution for environmental sound classification.

4 Results and Performance Analysis

4.1 Final Evaluation Results

The final neural network configuration was evaluated using 10-fold cross-validation to ensure robust generalization across the dataset. The best model, a three-layer feed-forward network with hidden dimensions of [1000, 500, 250], learning rate 5×10^{-4}, dropout 0.4, and batch size 500, achieved a mean test accuracy of **69.28 % ± 4.78 %**. Fold accuracies ranged from 58.38 % to 77.14 %, confirming stable, class-dependent performance.

Table 5: Final neural-network performance metrics (10-fold aggregated results).

Class	Precision	Recall	F1-Score	Support
Air Conditioner	0.565	0.429	0.488	1000
Car Horn	0.739	0.821	0.778	429
Children Playing	0.773	0.816	0.794	1000
Dog Bark	0.701	0.834	0.762	1000
Drilling	0.618	0.671	0.643	1000
Engine Idling	0.592	0.547	0.569	1000
Gun Shot	0.858	0.917	0.886	374
Jackhammer	0.640	0.498	0.560	1000
Siren	0.781	0.774	0.777	929
Street Music	0.732	0.836	0.781	1000
Overall Accuracy		**0.6928 ± 0.0478**		
Macro Avg	0.700	0.714	0.704	–
Weighted Avg	0.685	0.692	0.685	–

The model achieved the highest recall on *gun_shot* (91.7 %) and *dog_bark* (83.4 %), while stationary, low-frequency classes such as *air_conditioner* (42.9 %) and *jackhammer* (49.8 %) remained the most challenging.

4.2 Baseline vs. Neural Model Comparison

Performance comparison between the Naïve Bayes baseline (Figure 1) and the neural network (Figure 2) highlights the effect of model capacity on feature discrimination. The baseline achieved a mean cross-validated accuracy of **53.73 % ± 5.90 %** using $n_{mfcc} = 40$, RMS, ZCR, and Chroma (12). The neural network improved this to **69.28 % ± 4.78 %** using $n_{mfcc} = 120$, RMS, ZCR, and Chroma (12).

Table 6: Comparison between baseline and neural-network performance.

Model	Mean Accuracy (%)	Std. Deviation (%)
Naïve Bayes (Baseline)	53.73	5.90
Neural Network (3-Layer MLP)	69.28	4.78
Absolute Improvement	+15.55 percentage points	
Relative Improvement	+28.9 % over baseline	

This improvement validates the network's ability to model nonlinear relationships among MFCC-based acoustic features that the probabilistic baseline cannot capture.

4.3 Confusion Matrix and Error Analysis

Figure 1 (Baseline) and Figure 2 (Neural Network) present the confusion matrices and per-class accuracies. The Naïve Bayes classifier showed frequent confusion between low-frequency ambient categories, notably *air_conditioner → engine_idling (35.8 %)* and *drilling → jackhammer (29.2 %)*, which stem from overlapping spectral envelopes.

In contrast, the Neural Network model demonstrated substantial improvement in overall discriminative ability:

- **Overall**: Average per-class accuracy went up from 56.1% to 71.4%.
- **Significant Gains:** The Neural Network substantially reduced misclassifications for *air_conditioner* (Recall up from 17.2% to 42.9%) and *Siren* (Recall up from 36.5% to 77.4%). Furthermore, *drilling* Recall increased from 42.6% to 67.1%.
- **Critical Regression:** However, the model suffered a severe regression in the *jackhammer* class, dropping 31.8% percentage points from 81.6% (Baseline) to 49.8%(NN).

This suggests that while the NN effectively learned the complexity of low-frequency ambient sounds, its aggressive nonlinear feature mapping overcorrected for the naive baseline's high false-positive rate for *jackhammer*, leading to high false negatives for that class. This indicated the need for further hyperparameter tuning to stabilize performance on high-energy, pulsed transient classes.

4.4 Interpretation and Discussion

The neural model's higher accuracy (69.28 %) compared to the Naïve Bayes baseline (53.73 %) highlights the advantage of deep, nonlinear feature mappings for environmental sound recognition. The relatively small cross-validation variance (±4.78 %) indicates good generalization and suggests that the model did not suffer from significant overfitting. While the fully connected architecture effectively captured correlations among MFCC-based features, residual confusion among stationary low-frequency classes such as *air_conditioner* and *engine_idling* indicates that some spectral–temporal overlap remains unresolved. Crucially, the *jackhammer* class suffered a severe regression to 49.8% Recall, indicating a fundamental difficulty in stabilizing performance for pulsed transient sounds.

The observed recall range (42.9 %–91.7 %) further reflects moderate sensitivity to class imbalance and variability in sound duration. Future work could address these issues through convolutional or hybrid CNN–MLP architectures that better capture local spectro-temporal structure, as well as by applying data-augmentation techniques such as pitch shifting or additive noise to enhance robustness.

Overall, the neural network achieved consistent and interpretable improvements in classification accuracy and per-class balance compared to the baseline, confirming its suitability for modeling complex acoustic patterns in the UrbanSound8K dataset.

5 Conclusion

This assignment focused on developing and evaluating a neural network for environmental sound classification using the UrbanSound8K dataset. The task involved designing, implementing, and validating an Artificial Neural Network (ANN) capable of recognizing ten everyday urban sound classes, and benchmarking its performance against a probabilistic Naïve Bayes baseline model.

The baseline classifier achieved a mean cross-validated accuracy of 53.73% (±5.90%), confirming that simple probabilistic models can capture broad statistical patterns from MFCC- and chroma-based features but are limited in modelling complex nonlinear relationships. In comparison, the final neural network configuration, a three-layer feed-forward model with hidden dimensions [1000, 500, 250], learning rate 5×10^{-4}, and dropout 0.4, achieved a mean test accuracy of 69.28% (±4.78%) and a macro-averaged F1-score of 0.704. This improvement of approximately 16 percentage points demonstrates the ANN's superior capacity to learn richer acoustic representations from features.

Per-class results showed that transient, high-energy sounds such as *gun_shot*, *car_horn*, and *street_music* were classified with high accuracy, while continuous low-frequency sounds like *air_conditioner* and *engine_idling* remained challenging due to overlapping spectral characteristics. These findings are consistent with previous reports on the UrbanSound8K dataset [1].

Importantly, the evaluation adhered to the official 10-fold cross-validation protocol proposed by the dataset creators [1]. This procedure ensured that all audio slices originating from the same recording were kept within the same fold, preventing data leakage between training and testing. Although this resulted in lower accuracies compared to studies using a single random 80/20 split (which can yield inflated results exceeding 90%), it provided a far more reliable estimate of real-world generalization performance. The resulting scores therefore represent a credible and reproducible measure of the model's effectiveness.

The experimental procedure, which comprised standardized preprocessing, stratified fold generation, and early stopping, ensured that our results were robust and generalizable without significant overfitting. The stable convergence of both training and validation loss confirmed the effectiveness of our chosen hyperparameters and architecture.

Overall, our study demonstrated that even a moderately deep feed-forward neural network can substantially outperform traditional baselines for urban sound classification when trained on well-engineered MFCC-based features. Future work could explore convolutional or hybrid CNN–MLP architectures to better capture spectro-temporal relationships, as well as apply data augmentation techniques such as pitch shifting and noise injection to enhance model robustness.

In conclusion, our research successfully met its objectives by illustrating the clear advantages of neural network models over classical machine learning approaches for environmental sound classification. Our developed system achieved reliable, reproducible, and interpretable performance improvements, reinforcing the practical value of deep learning in intelligent urban monitoring and environmental sound analysis applications.

References

1. SALAMON, J. et al.: A Dataset and Taxonomy for Urban Sound Research. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 1041–1044. MM '14. ISBN 9781450330633. Available from DOI: 10.1145/2647868.2655045.
2. BARUA, S. et al.: A Deep Learning Approach for Urban Sound Classification. *International Journal of Computer Applications*. 2023, vol. 185, pp. 8–14. Available from DOI: 10.5120/ijca2023922991.