Topic: Gun Model Classification from Gunshot Audio for Security Intelligence in Katsina Using a Multi-Model Ensemble

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Katsina State, located in the northwestern region of Nigeria, has recently faced escalating security challenges, primarily characterized by widespread banditry and kidnapping. These issues not only threaten the peace and stability of the area but also impede economic development and strain local and national security resources. The geographical and socio-economic landscape of Katsina presents unique challenges in tackling these crimes, as the bandits often exploit the vast and less monitored terrains to carry out their operations. In response to this dire situation, there is an urgent need for innovative and efficient solutions to enhance security measures and safeguard the populace.

Katsina's security predicament is exacerbated by its geographic vastness and the sparse deployment of law enforcement resources across such a broad area. The current strategies for managing these security challenges involve conventional policing methods, community vigilante groups, and intermittent military interventions, which have been only partially effective due to logistical and coordination challenges (Smith, 2021). Moreover, the state's proximity to other volatile regions further complicates the security dynamics, making it a transit route for arms and criminals (Johnson, 2020).

Artificial Intelligence (AI) emerges as a potent tool for revolutionizing security strategies. AI can analyze vast amounts of data quickly and with high accuracy, making it ideal for monitoring and reacting to security breaches in real time. Specifically, AI technologies can be deployed for the digital forensics of gun audio samples, a method that involves the analysis of audio clips to detect and classify firearm discharges. This technology leverages machine learning algorithms to identify the acoustic signatures of gunshots, distinguishing them from other background noises, and can even pinpoint their location and the type of firearm used (Brown & Patel, 2019). The application of AI in this domain could provide law enforcement agencies with critical insights and reaction capabilities. For instance, real-time audio surveillance systems equipped with AI could be deployed across Katsina, particularly in areas known for high criminal activity.

The integration of AI into digital forensics and surveillance systems is well-documented, with numerous scholars emphasizing the effectiveness of machine learning in enhancing security measures. Notable studies include the work of Srikiar et al. (2023), who utilized advanced models like ResNet50, Swin-T, and ConvNeXt-T to classify firearms based on visual data from CCTV footage. Similarly, Junwoo et al. (2022) explored the classification of gun sounds from video games data using a hybrid approach that combines CNNs and Transformers. Furthermore, Najihawan et al. (2022) applied Transformer technology to identify gunshots in efforts to secure public spaces, while Raponi et al. (2022) leveraged CNNs to categorize firearms into shotguns, rifles, and pistols based on their acoustic profiles.

Moreover, integrating AI with existing security infrastructure can enhance the overall effectiveness of security operations. Such systems would not only detect gunshots but also alert nearby authorities instantly, thereby reducing response times and improving the chances of apprehending perpetrators (White, 2022). Also, data collected from AI-based audio forensic systems can be analyzed alongside visual surveillance data, witness reports, and other intelligence inputs to create comprehensive security assessments and predictive models, thereby enabling proactive rather than reactive strategies (Lee & Kim, 2021).

1.2 Statement of the Research Problem

In Katsina, traditional security approaches have predominantly been reactive, struggling with significant challenges in timely intelligence gathering, resource limitations, and the inherent unpredictability of criminal behavior. The prevalent methods for detecting and responding to security threats primarily involve physical patrols and human intelligence. These methods are not only resource-demanding but also subject to considerable risks and delays. In light of the escalating security crisis, there is a crucial need for adopting and integrating advanced technological solutions. Specifically, there is a gap in the use of artificial intelligence (AI) for the auditory detection, surveillance, and analysis of gunshots, which frequently occur during criminal incidents.

While there have been numerous global efforts to classify gunshots, these attempts have generally focused on merely distinguishing gunshots from other ambient sounds to confirm the occurrence of gunfire. This approach, while beneficial for immediate threat recognition, falls short of providing the detailed information about the gun type or model that is essential for comprehensive security forensics. Understanding the specific type of firearm used in an incident can greatly aid law enforcement in their investigations and in crafting more targeted responses to criminal activities.

Park et al. (2022) made strides in this direction by attempting to detect and classify guns based on their models using synthetic data from video games. However, such data may lack the fidelity and complexity of real-world gunshot sounds, thus potentially limiting the applicability of their findings to actual field conditions. Moreover, the sound signatures of firearms in video games are often altered for dramatic effect, which could lead to inaccuracies when applied to real-world scenarios.

Similarly, the work of Raponi et al. (2022) made an important contribution by categorizing firearms into three broad categories—rifles, shotguns, and pistols—using audio data. Nonetheless, their approach relied on a single classification model and provided a very high-level categorization. This methodology may not capture the subtleties and variances among different models and makes of firearms within each category, which are often crucial for forensic analysis. Furthermore, such a generalized model may not adequately account for environmental variables that significantly affect sound propagation, such as urban versus rural settings, indoor versus outdoor environments, and background noise levels.

This research aims to bridge the gaps identified in previous studies by proposing a novel and sophisticated approach: the development of a Multi-Model Ensemble technique. This technique leverages the strengths of multiple predictive models to enhance the accuracy and reliability of gunshot detection and classification systems. By integrating various algorithms, each with unique strengths in handling different aspects of the audio analysis, this ensemble method can robustly discern between numerous gun models and types based on their distinct acoustic signatures.

The proposed system will utilize pre-recorded, real-time audio data, which includes a diverse range of gunshots recorded under various conditions. This rich dataset allows for training the models to recognize and differentiate gun sounds more effectively, accommodating factors such as echo, distance, and interference from environmental noise, which are common in real-world scenarios.

By classifying gunshots not just as generic noises but according to specific models and makes of firearms, this approach can provide critical information more swiftly and accurately. For instance, knowing whether a gunshot originated from a high-caliber rifle as opposed to a handgun can drastically alter the response strategy of law enforcement agencies. Such precise information enhances tactical decision-making and resource allocation, thereby improving the overall efficiency and effectiveness of security operations in Katsina.

Furthermore, the application of a Multi-Model Ensemble approach can adapt over time to new types of gunfire and emerging acoustic patterns through continuous learning processes. This adaptability ensures that the system remains effective even as new firearms enter the market or as criminals modify existing guns to evade detection.

1.3 Aim of the Research

The primary aim of this research is to develop and implement a Multi-Model Ensemble technique for the classification of gun models based on audio recordings of gunshots.

1.4 Research Objectives

The following objectives of the study are:

i. To design a Multi-Model Ensemble architecture that effectively combines different machine learning algorithms for gunshot audio classification.

ii. To evaluate the performance of the ensemble model in accurately classifying different types of firearms based on their acoustic signatures.

iii. To develop a prototype system that can be integrated into existing security infrastructure for real-time gunshot detection and classification in Katsina.

1.5 Scope and Delimitations

1.5.1 Scope of the Study

The scope of this research is explicitly on the technological aspects of security intelligence, focusing on the acoustic analysis of gunfire encompassing:

i. **Model Development:** The core of the research involves creating a sophisticated Multi-Model Ensemble system designed to classify gunshots based on their audio signatures. This includes the selection, tuning, and integration of various machine learning models to optimize performance.

ii. **Data Acquisition and Processing:** The study will utilize a curated dataset of gunshot sounds, recorded under controlled and diverse environmental conditions to mimic real-world scenarios. The dataset includes gunshots from eight types of firearms that are most commonly found in Katsina State. This controlled dataset ensures that the system can handle variations in gunshot audio due to factors such as echo, distance, and ambient noise.

iii. **Model Evaluation:** The models will be rigorously tested for accuracy, robustness, and efficiency. Evaluation metrics will include precision, recall, and the F1-score, among others, to assess the models' capabilities in correctly identifying and classifying the types of firearms based on their gunshot sounds.

iv. **Technological Implementation:** The research aims to create a deployable prototype that can integrate with existing digital surveillance and security systems to provide real-time gunshot detection and classification.

1.5.2 Delimitations

While the study is comprehensive in terms of technological development for audio analysis of gunshots, it has specific delimitations:

i. **Focus on Audio Data:** The research is strictly limited to audio data. It will not explore other sensory data modalities such as visual (CCTV footage), thermal imaging, or seismic sensors, which might also be useful in the broader context of security and forensic analysis.

ii. **Exclusion of Non-Technical Aspects:** The project does not encompass policy-making, community engagement strategies, or the direct physical deployment of security technologies, such as the installation of microphones or sensors. These aspects are considered beyond the technical scope of this study.

iii. **Geographic and Model Limitation:** The investigation is geographically confined to Katsina State, Nigeria, and will only consider eight specific models of guns. These limitations are set to focus the research and ensure the manageability of the study, although they might impact the generalizability of the results to other regions or firearm types.

iv. **Exclusion of Post-Detection Response:** The research will not cover response strategies or operational protocols following the detection and classification of gunshots. The study's focus is on the detection and classification technology itself, not on the tactical or strategic responses by law enforcement or security personnel.

1.6 Definitions of Terms

i. **Gun:** A firearm designed to discharge projectiles (bullets) at high speeds through the confined burning of propellants. In this research, guns refer specifically to the firearms involved in the recorded gunshot sounds.

ii. **Model:** In the context of this research, a model refers to a specific make and configuration of a firearm, distinguishable by unique characteristics such as size, caliber, and mechanism.

iii. **Classification:** The process of identifying and assigning categories to data points. In this case, audio recordings based on specific characteristics or features. Classification helps in distinguishing between different models of guns based on their gunshot sounds.

iv. **Gunshot:** The sound emitted when a gun is fired, which results from the explosion of the gunpowder and the movement of the bullet through the gun barrel. The acoustic properties of a gunshot can vary based on the gun model, firing conditions, and environment.

v. **Audio:** Sound, specifically the recorded or digitally captured sound waves of gunshots. This research utilizes audio data to analyze and classify gunshots by model.

vi. **Security:** Measures and protocols implemented to protect people, property, and institutions from threats and crimes. In this context, enhancing security refers to improving the ability of law enforcement agencies in Katsina to respond to gun-related crimes through better gunshot detection and analysis.

vii. **Intelligence:** Information relevant to crime prevention and security enforcement. In this study, intelligence pertains to the data derived from analyzing gunshot sounds, which can provide actionable insights into the type of firearms used in criminal activities.

viii. **Katsina:** A state in the northwestern region of Nigeria, which serves as the primary area of focus for this research due to its unique security challenges, including banditry and kidnapping.

ix. **Multi-Model:** Pertaining to the use of multiple models in a computational or analytical process. Here, it involves employing various machine learning models within an ensemble to capitalize on their combined strengths in analyzing gunshot audio data.

x. **Ensemble:** A technique in machine learning where multiple models (often diverse in nature) are trained and then combined to solve the same problem more effectively than any single model could alone. This ensemble approach aims to improve prediction accuracy and robustness in classifying gunshots into specific models.

CHAPTER THREE

METHODOLOGY

3.1 Proposed Model

3.1.1 Audio Signal Pre-Processing

The dataset consists of audio recordings of gunshots. Each file is a digital audio file in WAV format, capturing the unique acoustic signature of a firearm discharge. These recordings represent nine different classes of gun sounds. Several audio features are extracted from the gunshot WAV files to capture the essential characteristics of the sounds for machine learning analysis. Firstly, the Mel-Frequency Cepstral Coefficients were obtained by converting the time signal to the frequency domain using the Fourier transform. Then a set of triangular filters were applied (20-40 filters) aligned on the Mel scale to the power spectrum to extract bands of frequencies. Lastly, the discrete cosine transform was applied to decorrelate the log filter bank coefficients and yield a compressed representation of the filter banks as represented in Equation 1.

$$C_k = \sum_{n=1}^{N} \log(S_n) \cos\left[k \left(n - \frac{1}{2}\right)\frac{\pi}{N}\right]$$

where N is the number of Mel-scale filters, and Sn is the output of the n-th filter.

The second feature was spectrogram representing the intensity of frequencies in the sound over a period of time. It was calculated by performing a Fourier transformation on overlapping segments of the audio signal. Equation 2 depict the spectrogram calculation

$$S(t,f) = |FFT(x(t).w(t))|^2$$

where $x(t)$ is the signal, w(t) is a window function, and f denotes frequency.

The sounds was then distinguish between percussive and harmonic using Zero-Crossing-Rate which measures the rate at which the signal changes from positive to negative of back as provided in equation 3.

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} 1((x(t) > 0\ \wedge x(t+1) < 0) \vee (x(t) < 0\ \wedge x(t+1) > 0))$$

where 1 is the indicator function.

The next extracted was ‘Chroma’ which allows the audio energy to be distributed across 12 different pitch classes each corresponding to a musical semitone within an octave. Time-Domain signal was first transform into frequency using Short-Time Fourier Transform and the mapped each frequency to its respective pitch classes. Then sum the energies of all frequencies belonging to each pitch class, across multiple octaves. And, lastly, normalization is applied to each time frame to make the chroma features robust to variations in dynamics as provided equation 4.

$$C(t) = [C_0(t), C_1(t), \ldots \ldots, C_{11}(t)]$$

where $C_i(t)$ is the energy sum for pitch class ii at time frame tt. This is computed by summing up the magnitudes of the STFT coefficients that correspond to frequencies falling into pitch class i.

The last extracted feature is the spectral contrast of the gunshot audio which considers the difference in amplitude between peaks (highest energy) and valleys (lowest energy) in the spectrum of an audio signal. This was computed across several frequency bands to capture different characteristics of the sound spectrum. If P_k and V_k represent the peak and valley amplitude in the k-th frequency band, respectively, the spectral contrast S_k can be calculated as in Equation 5:

$$S_k = \log(P_k) - \log(V_k)$$

This is done for each frequency band and each time frame, providing a measure of spectral contrast over time.

3.1.2 Multi-Model Ensemble

Following the pre-processing of audio signals where key features like Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, Zero-Crossing Rate (ZCR), chroma features, and spectral contrast were extracted, the gunshots sounds classification was done using a Multi-Model Ensemble (MME) approach. These include Support Vector Machines, Random Forest, Gradient Boosting Machine as well as Deep Neural Network. This approach leverages the strengths of multiple machine learning algorithms to improve prediction accuracy and robustness compared to a single model.

The Support Vector Machine are adept at managing high-dimensional spaces, which is essential for effectively interpreting the MFCCs and spectral contrast features in this domain. The SVM classification is represented in equation 6.

$$f(x) = sgn\left(\sum_{i=1}^{n} \sigma iyi\langle xi,x\rangle + b\right)$$

The Random Forest classifiers was also employed which are beneficial for integrating insights from the chroma features and Zero-Crossing Rate as provided in equation 7.

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} Tree_b(x)$$

As presented in equation 8, the Gradient Boosting Machines enhance classification performance by sequentially correcting the mistakes of weak learners, making them suitable for high degree of distinctions among gunshot classes.

$$y = \sigma\left(\sum_{i=1}^{n} W_i x_i + b\right)$$

3.1.3 Ensemble Strategy

The weighted voting ensemble mechanism was employed for this study. Each classifier predicts a class label for the input sample, and the final output is determined by aggregating these predictions, with weights assigned based on the individual classifier's performance on the validation set as presented in Equation 9 respectively.

$$Class_{final} = argmax_k\left(\sum_{j=1}^{j} w_j 1(class_j = k)\right)$$

where J is the number of classifiers, wj is the weight for the j-th classifier, and 1 is the indicator function.

3.2 Prototype Web Application & Model Deployment

An intuitive interface was designed to allow users to upload gunshot audio files and view the classification results. The interface included elements such as file upload buttons, sections for displaying results, and interactive features for detailed analysis. The backend was developed using Django, a high-level Python web framework that encourages rapid development and clean, pragmatic design. It handled requests such as audio file uploads, processing these files through the gunshot sound classification model, and sending the results back to the frontend.

The Multi-Model Ensemble, developed as per the methodology section, was integrated into the backend. This integration involved setting up a pipeline from audio file input, through pre-processing and feature extraction, to classification using the ensemble model

3.3 System Requirements

Item	**Requirements**
Computer Device	HP X360 – Convertible
Operating System	Windows 11 Professional
Programming Language	Python 3.13
Framework	Django

3.4 Performance Evaluation

In assessing the performances of the gunshot sound detection of the model, metrics such as the precision-recall (P-R) curve, average precision (AP), and mean average precision (mAP) were utilized. Precision, a measure of accuracy in information retrieval contexts where precision and recall are often considered together, quantifies the ratio of relevant targets accurately identified among the returned results to the total number of targets returned for a specific query. The evaluation includes terms like true positive (TP), true negative (TN), false positive (FP), and false negative (FN) to describe classification outcomes. TP signifies the correct prediction of positive instances as positive, TN denotes the correct prediction of negative instances as negative, FP indicates negative instances incorrectly predicted as positive (false positives), and FN represents positive instances incorrectly predicted as negative (false negatives). The precision formula is expressed as in equation 10:

$$Precision = \frac{TP}{TP + FP}$$

Additionally, the recall rate, measuring the proportion of relevant targets among all relevant targets, is defined as equation 11:

$$Recall = \frac{TP}{TP + FN}$$

In certain cases, specific values offer a clearer representation of the test model's performance than a graphical representation. Average precision (AP) is commonly used for this purpose, calculated using the formula as presented in equation 12:

$$AP = \int_0^1 p(r)d(r)$$

In this formula, 'p' represents precision, 'r' represents recall, and precision is a function of recall. Therefore, the average precision corresponds to the area under the precision-recall (P-R) curve, and mAP (mean average precision) is the average of the average precision values across all categories.