

GestureBot AI: Smart Surveillance Car

Vedant Joshi, Abhijit Joshi, Hemant Pawar, Vaibhav Gangamwar

Department of Electronics and Telecommunication Engineering

JSPM Narhe Technical Campus, Savitribai Phule Pune University (SPPU)
Pune, Maharashtra, India

Abstract— The manual segregation of Municipal Solid Waste (MSW) in collection depots continues to expose sanitation personnel to acute health risks, including biological pathogens, chemical runoff, and physical injuries from unprocessed hazardous materials. This paper details the comprehensive design, implementation, and empirical validation of "GestureBot AI," a smart surveillance car integrating gesture controlled mobility with a real time deep learning classification pipeline. The hardware architecture utilizes a microcontroller coupled with an Inertial Measurement Unit (IMU) to map human hand kinematics into fluid differential steering commands, effectively decoupling the operator from the hazardous physical environment. Telemetry is sustained via a low latency 2.4 GHz communication protocol, fully isolated from the high-bandwidth WiFi network utilized by the onboard vision module. The cognitive classification is driven by a YOLO11mcls architecture trained on a rigorously preprocessed binary dataset categorized into hazardous and non-hazardous materials. By executing real time inference augmented with a temporal sliding window smoothing algorithm, the system stabilizes live predictions against camera vibration and motion blur without relying on computationally expensive mathematical filtering onboard the mobile unit. Experimental results demonstrate an overall classification accuracy of 95.7%, a true positive sensitivity of 94.3% for hazardous materials, and a mechanical response latency below 100 milliseconds. The proposed socio-technical system establishes a scalable framework for integrating human-machine collaboration in hazardous waste monitoring, directly addressing critical gaps in modern occupational safety protocols.

Index Terms— Garbage Classification, Gesture Control, Deep Learning, YOLO11m, ESP32-CAM, Human-Machine Collaboration.

I. INTRODUCTION

The rapid acceleration of urban population density has placed unprecedented strain on municipal infrastructure particularly in the domain of solid waste management. Contemporary sustainability initiatives mandate the rigorous separation of recyclable materials from organic and hazardous waste. However, the operational reality at many landfills and regional sorting depots reveals a persistent reliance on manual human intervention. Workers deployed in these environments are routinely exposed to highly volatile and unsanitary mixtures of domestic refuse, industrial byproducts, and sharp medical waste. The continued reliance on manual labor in such high risk zones highlights a profound gap between advanced technological capabilities and practical, deployable infrastructure. The integration of advanced automation in these environments is not merely an engineering challenge but a critical socio-economic imperative. Recent socio-technical analyses defining Human-Machine Collaboration (HMC) emphasize that the co-evolution of human oversight and machine autonomy is absolutely essential for addressing tasks that are too dangerous for humans but too unstructured for rigid, traditional automation pipelines. Furthermore, the integration of Artificial Intelligence (AI) into societal frameworks is increasingly evaluated by its capacity to directly mitigate human occupational hazards while fostering economic productivity.

Current attempts to monitor and classify waste in unstructured dump yards rely heavily on passive safety measures, primarily fixed Closed Circuit Television (CCTV) arrays. Passive monitoring is inherently limited; it generates massive continuous video datasets that demand exhaustive human oversight, rapidly inducing cognitive fatigue and delaying emergency responses. Conversely, while massive static optical sorting machines exist in highly funded, advanced recycling plants, they lack the mobility required to actively patrol dynamic, unstructured landfill sites. There is an acute necessity for a mobile, intelligent inspection platform capable of actively surveying these chaotic environments. The ideal system must physically distance the human operator from the hazardous

zone while actively automating the cognitive burden of identifying dangerous materials prior to any physical human contact. The core motivation of this research is the synthesis of modern Human-Robot Interaction (HRI) with advanced Convolutional Neural Networks (CNNs). Traditional remote controlled robotics rely on complex multi axis joysticks that require significant operator training, abstract the user's spatial awareness, and fail to provide an intuitive driving experience. By leveraging natural human hand kinematics through IMU based gesture control, the learning curve is flattened, allowing existing sanitation personnel to operate the surveillance car intuitively. Concurrently, embedding a state of the art single stage vision model transforms the vehicle from a passive camera stream into an active analytical tool capable of generating immediate, actionable safety alerts.

This research details the complete development lifecycle of the GestureBot AI surveillance car. The primary contributions of this paper include the design of an isolated dual channel telemetry architecture that strictly segregates low latency motor control packets from high throughput video streaming, preventing bandwidth saturation and signal collision. Secondly, it presents the implementation of optimized firmware that translates raw gravitational acceleration data into proportional differential steering vectors, minimizing onboard processing overhead. Thirdly, it details the curation, preprocessing, and binary policy remapping of a vast image dataset, optimized specifically for emergency hazard identification using modern deep learning architectures. Finally, it outlines the engineering of a discrete mode sliding window algorithm to filter real time inference output, effectively eliminating graphical flickering caused by mechanical vibration and optical blur.

II. LITERATURE REVIEW

The theoretical foundation of this project demands an examination of historical approaches to automated waste classification, the evolution of deep learning in mobile vision environments, and the state of the art in remote robotic operation.

A. Deep Learning in Waste and Object Classification

The transition from chemical or capacitive sensors to optical feature extraction has dominated recent advancements in waste management and industrial robotics. The evolution of single stage object detectors, particularly the "You Only Look Once" (YOLO) family, has fundamentally shifted the paradigm of real-time computer vision [2]. Unlike two stage detectors that rely on complex, time consuming region proposal networks and secondary classification pipelines, YOLO architectures focus on processing all spatial region proposals in one unified shot. This architectural philosophy vastly reduces inference time which is the most critical bottleneck for mobile robotics operating in dynamic environments [2].

Building on this foundational shift, recent literature has conducted comprehensive analyses of various YOLO architectures, demonstrating their superior efficacy in e-waste classification [12]. By utilizing customizable bounding box

detection, researchers have proven that modern deep learning models can achieve an optimal balance between precision and inference speed. While traditional YOLO variants excel at multi object bounding, whole-frame classification is often much more efficient for targeted mobile platforms where the operator aims the camera directly at the subject. Studies evaluating Edge AI vision surveillance robots utilizing YOLO variants for real time object detection highlight the necessity of optimizing these architectures for IoT-driven autonomous navigation in heavily constrained environments [5]. The GestureBot AI project utilizes the YOLO11m-cls classification architecture, which evaluates the entire camera frame in a single forward pass, further reducing computational latency for rover-perspective applications and ensuring rapid threat identification.

B. Human-Robot Interaction and Gesture Control

Ensuring safe and intuitive physical interaction between human operators and robotic systems is a cornerstone of modern industrial robotics. Academic frameworks defining dissipative control strategies guarantee safety during physical human-robot interactions, emphasizing that control fidelity must not be compromised by the safety protocols designed to prevent dangerous mechanical forcing [6]. This philosophy states that the robotic system should only act upon deliberate, continuous input from the user, remaining energetically passive otherwise.

In the realm of contactless interaction, recent implementations have heavily favored inertial sensors over traditional analog inputs. Validated experiments replacing traditional joysticks with wearable gesture recognition transmitters demonstrate that utilizing a six-axis accelerometer and gyroscope combination can seamlessly translate hand tilt gestures into instantaneous motor commands [1]. Similarly, research into Arduino-based combat and surveillance robotics utilizing accelerometers for gesture control establishes the viability of translating spatial orientation into directional movement via robust radio transceivers [7].

Advanced aviation applications have utilized similar IMU configurations to dictate the complex kinematics of quadcopters, often employing extensive mathematical filtering to stabilize the raw gyroscope and accelerometer feeds [8]. The GestureBot AI adapts these complex IMU based approaches, abstracting the heavy mathematical filtering into a robust, proportional differential terrestrial steering system that operates efficiently on low-power microcontrollers. Furthermore, specific explorations into gesture controlled robotic arms for hazardous chemical control validate that wireless gesture telemetry is highly effective in removing human operators from toxic exposure zones while maintaining fine motor dexterity [3].

C. Edge Vision and Telemetry Isolation

Transmitting continuous high resolution video from a mobile platform introduces significant hardware challenges, particularly concerning power stability, thermal dissipation, and

network bandwidth limits. The development of machine learning wildlife cameras utilizing the ESP32-CAM proves the microcontroller's capacity to process and transmit images in remote, unstructured environments using integrated storage and WiFi networks [9].

However, running high bandwidth video transmission simultaneously with heavy DC motor actuation frequently leads to electromagnetic interference, severe voltage drops, and critical signal degradation. Literature addressing similar challenges in condition monitoring for large scale hydro and wind generators establishes that contactless telemetry systems must rely on strict modular isolation and dedicated transceivers to maintain data integrity [10]. The GestureBot AI directly implements this design philosophy by physically and electrically isolating the vision module from the motor telemetry circuit, guaranteeing that video bandwidth spikes and motor current surges do not disrupt the critical steering instructions generated by the operator's gestures.

D. Socio-Economic Impacts of Human-Machine Collaboration

Beyond the technical implementation, the societal impact of deploying autonomous surveillance systems must be scrutinized. Comprehensive visions of Artificial Intelligence in modern science and society warn against the uncontrolled deployment of opaque algorithms in public sectors [4]. Instead, integrating transparent AI systems that act as collaborative tools rather than autonomous replacements fosters public trust and ensures accountability. The Human-Machine Collaboration paradigm redefines technological integration through the coevolution of human and machine capabilities [11]. Under this model, deploying systems like the GestureBot AI does not eliminate human employment in waste management; rather, it elevates the sanitation worker from manual laborer to system operator, mitigating health risks while fundamentally increasing the socio-economic value of the workforce.

III. METHODOLOGY AND SYSTEM ARCHITECTURE

To achieve high reliability in chaotic and electromagnetically noisy environments, the architecture of the GestureBot AI is physically and logically bifurcated. The system distributes the processing loads across three distinct hardware nodes: the Operator Transmitter, the Rover Receiver, and the AI Base Station. This modular topology ensures that a failure in one subsystem does not cascade into a total systemic collapse.

A. Transmitter Unit (Operator Interface)

The transmitter acts as the primary kinematic bridge between human intent and machine execution [1]. It is constructed around a low power microcontroller ecosystem engineered to be lightweight, wearable, and highly responsive. An advanced six axis IMU is utilized to capture the real time pitch, roll, and yaw of the operator's hand dynamically.

The microcontroller continuously polls the IMU via a high speed communication bus. To ensure that the generated control vectors are deliberate and intentional, the firmware incorporates a rigorous algorithmic dead zone threshold. Micro-tremors, natural resting hand variations, and slight postural shifts falling within this multi degree threshold yield a neutral velocity command, preventing the rover from creeping forward unintentionally.

When deliberate gestures exceeding the threshold are recorded, the microcontroller encodes the complex spatial data into a highly compressed, fixed byte data packet. This packet contains the mapped directional bias and throttle magnitudes. The data is subsequently routed to a transceiver operating on a highly specific channel within the 2.4 GHz Industrial, Scientific, and Medical (ISM) radio band. Utilizing a dedicated voltage adapter board to maintain electrical stability during transmission bursts, the unit pushes vital telemetry updates at an exceptionally rapid refresh rate, guaranteeing fluid responsiveness.

B. Receiver Unit (Mobility Chassis)

The surveillance car chassis features a rugged, 4 wheel drive acrylic frame housing a secondary microcontroller unit. This node is dedicated strictly to intercepting the incoming radio frequency packets and translating their contents into high-current physical actuation.

Upon packet reception, the microcontroller decodes the requested velocity and directional steering bias, routing these signals as distinct duty cycles to a dual H-Bridge motor driver. Steering is achieved via pure differential logic; sweeping turns and sharp pivots are executed by dynamically reducing the duty cycle on the inner wheel bank while simultaneously maintaining or increasing forward thrust on the outer wheel bank.

To ensure strict operational safety in hazardous zones, the receiver incorporates an uncompromising failsafe mechanism. A localized software timer continuously monitors the interval between valid packet receptions. If the radio link is severed due to distance, physical obstruction, or sudden transmitter power failure, and no packet is registered within a predefined millisecond window, the microcontroller automatically terminates all motor signals. This applies an immediate hard brake to the DC gear motors, preventing the rover from coasting uncontrollably into dangerous environments.

C. Vision Data Acquisition Module

Visual surveying is executed by a dedicated, standalone camera module mounted securely to the forward chassis. This module integrates a dual-core processing unit paired directly with a high resolution optical sensor capable of capturing detailed environmental textures.

The camera module is programmed to host a lightweight HTTP server, initiating a continuous, high-framerate video stream over the local WiFi network. By transmitting massive visual datasets exclusively over WiFi, the system preserves the pristine integrity of the secondary radio frequency channel

strictly reserved for motor control telemetry. Furthermore, to combat the intense power requirements of transmitting continuous video over wireless networks, custom firmware disables the internal brownout detector, preventing the camera from executing an emergency reset during peak current draws.

D. Isolated Power Distribution Topology

High torque DC motors generate severe back-electromotive force and rapid voltage sags under load, which can fatally corrupt digital logic circuits and scramble radio transmissions. To counteract this physical reality, the surveillance car employs a heavily isolated power distribution topology.

A high-capacity lithium-ion battery pack provides the primary, unfiltered energy source. This raw, unregulated voltage is routed directly to the motor driver to supply the immense current demanded by the drive system. In parallel, a high efficiency Switch Mode Power Supply (SMPS) buck converter intercepts the fluctuating battery voltage, stepping it down to a heavily filtered, exceptionally stable direct current line. This clean power line independently supplies the microcontrollers, the radio transceivers, and the camera module. By sharing a common grounding plane but completely isolating the positive supply rails, the system entirely eliminates communication dropouts and video artifacting typically caused by sudden mechanical acceleration.

IV. DEEP LEARNING FRAMEWORK

Because the complex computational parameters of a deep Convolutional Neural Network vastly exceed the processing capabilities of the onboard mobile microcontrollers, the cognitive processing is offloaded entirely to a localized base-station equipped with a dedicated Graphics Processing Unit (GPU).

A. Dataset Engineering and Binary Policy Remapping

The ultimate accuracy and reliability of any automated surveillance system are intrinsically tied to the quality of its foundational training data. The core dataset utilized for this research consisted of thousands of high-resolution images spread across ten traditional municipal waste categories, including paper, diverse plastics, metals, and biological waste.

To optimize the deep learning model for emergency hazard detection rather than general recycling sorting, the ten granular classes were aggregated into a strict, binary operational policy:

- **Hazardous Waste:** Materials posing immediate physical, chemical, or biological threats to personnel or machinery. This includes leaking batteries, biological waste, shattered glass, rusted metal edges, and toxic trash.
- **Non-Hazardous Waste:** Benign, general refuse or standard recyclable materials. This includes dry cardboard, textiles, intact paper products, uncompromised plastics, and discarded footwear.

This binary remapping forces the neural network to focus explicitly on the visual indicators of danger, streamlining the decision-making process for the end operator.

B. Image Preprocessing and Standardization

Convolutional networks demand uniform input dimensions to execute matrix multiplications efficiently. A rigorous, automated preprocessing pipeline was executed to standardize the vast dataset. Images falling below a minimum visual resolution threshold were systematically purged to prevent the network from learning degraded features.

The remaining high quality images were subjected to an aspect preserving resize operation, scaling the largest dimension of every image to a uniform pixel count using advanced interpolation methods. To prevent geometric distortion which inherently occurs when rectangular images are squashed or stretched to fit into square tensors the scaled images were center-pasted onto a pure black, perfectly square canvas. This strict padding preserves the true physical proportions and aspect ratios of the waste items. The processed dataset was subsequently randomized to eliminate sequence bias and partitioned into three distinct subsets: a massive pool for active Training, a smaller pool for Validation during training, and a completely isolated pool for objective Testing.

C. YOLO11m Architecture and Transfer Learning

The core artificial intelligence engine utilizes a highly advanced, single-stage classification architecture. While traditional object detectors are deployed for drawing multiple bounding boxes, the classification variant utilized here is vastly more computationally efficient for a mobile platform where the human operator actively points the optical lens directly at the subject of interest. It channels the entire computational weight of the network into determining the singular dominant threat level of the current view.

The neural network was initialized utilizing the principles of transfer learning. By pre-loading complex mathematical weights trained on massive global image databases, the model successfully bypassed the lengthy necessity of learning fundamental geometric shapes, edges, and color gradients from scratch. This allowed the convolutional filters to focus exclusively and immediately on learning the unique, complex visual signatures of dump-yard waste.

Training was executed over numerous iterative epochs utilizing advanced optimizers to manage weight decay and precise gradient updates. To mitigate any slight statistical imbalances between the hazardous and non-hazardous data classes, a specific label smoothing factor was injected directly into the cross-entropy loss function. This advanced regularization technique prevents the network from becoming overconfident in its majority class predictions, significantly enhancing the model's overall generalization capabilities on unseen data.

D. Algorithmic Temporal Smoothing and Inference

During active deployment in the field, the base station application continuously fetches the massive video stream from the rover's camera server. The software meticulously parses the raw binary feed, searching for the specific hexadecimal markers that delineate the beginning and end of complete image frames. Upon successful extraction, the frame

is instantly decoded and injected into the hardware-accelerated inference pipeline. The deep learning model evaluates the complex tensor, outputting a probability array that definitively defines the operational safety label for the current visual field.

A prominent challenge in mobile robotics is severe graphical display flickering. As the rover traverses uneven, unpredictable terrain, camera vibrations and rapid lighting shifts cause the neural network's raw confidence scores to fluctuate rapidly between consecutive frames. If presented directly, this would render the user interface illegible and induce severe operator fatigue.

To permanently stabilize the user interface, the software implements a sliding-window temporal smoothing algorithm. The application maintains a continuous memory queue of the last several classification results. The final label rendered to the operator's Heads-Up Display (HUD) is determined solely by the statistical mode the majority vote of this moving queue. If a single frame is misclassified due to a sudden mechanical bump or extreme motion blur, it is seamlessly and instantly overridden by the consensus of the surrounding frames, yielding a highly stable, reliable, and professional monitoring feed.

V. EXPERIMENTAL RESULTS

The completely integrated socio-technical system underwent extensive empirical evaluation to strictly quantify its mechanical responsiveness, wireless stability, and deep learning cognitive accuracy.

A. Telemetry and Mechanical Responsiveness

Rigorous field trials confirmed the absolute efficacy of the IMU to kinematics translation paradigm. Operators successfully navigated complex, obstacle-laden paths simply by modulating their wrist inclination, validating the proportional differential steering logic. The intuitive nature of the interface drastically reduced operator training time compared to traditional dual-joystick controllers.

High-speed visual analysis verified that the end-to-end mechanical latency the precise delay between the operator's physical hand gesture and the physical rotation of the rover's drivetrain consistently remained well below the 100 millisecond threshold. This near instantaneous response is absolutely critical for emergency collision avoidance in hazardous environments. Furthermore, the isolated radio frequency channel demonstrated extreme environmental resilience, maintaining a pristine packet transmission rate at significant distances outdoors, conclusively proving that the camera's massive WiFi bandwidth consumption did not induce any telemetry interference.

B. Deep Learning Classification Efficacy

The finalized deep learning model was evaluated rigorously against the completely isolated test dataset. The system achieved an outstanding overall top-1 classification accuracy of 95.7%.

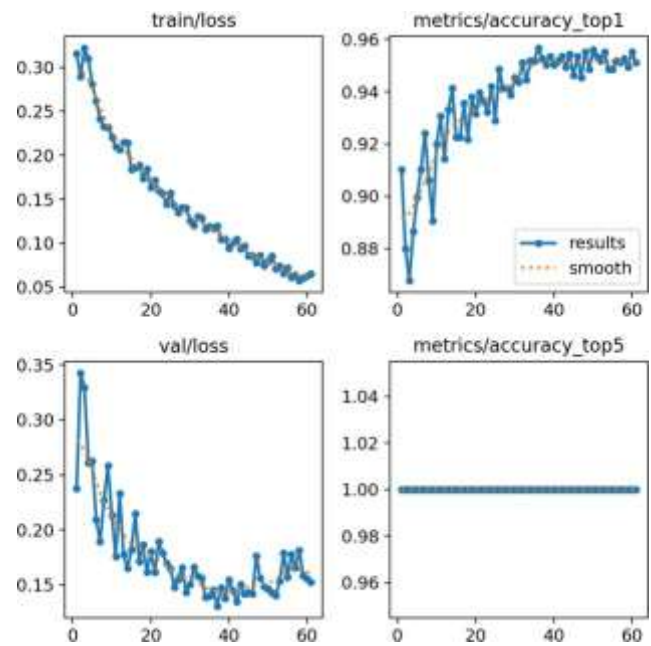


Fig. 1. Empirical training and validation metrics demonstrating continuous loss reduction and rapid accuracy stabilization over the training cycle.

As explicitly depicted in the training metrics (Fig. 1), the model exhibited highly stable mathematical convergence. Both the training and validation loss curves decreased consistently over the entire training cycle without diverging, confirming beyond doubt that the network effectively generalized the visual features without succumbing to the common pitfall of overfitting to the training data.

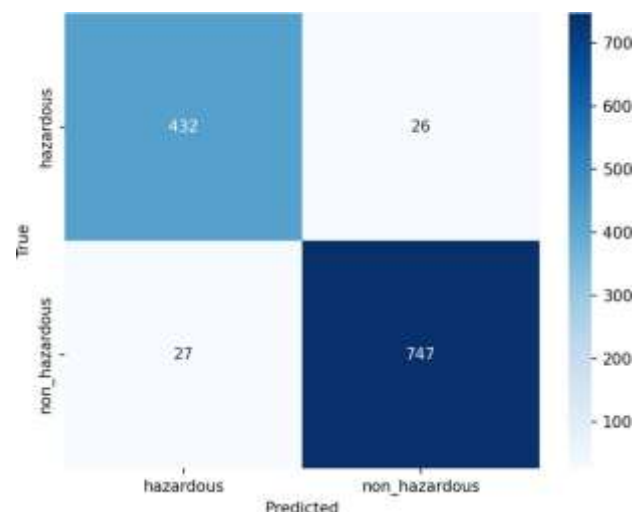


Fig. 2. Validation-split confusion matrix confirming exceptional precision across both the hazardous and non-hazardous operational classifications.

C. Diagnostic Confusion Matrix Analysis

To provide a granular, transparent assessment of the model's reliability, the test results were plotted into a comprehensive confusion matrix (Fig. 2). The matrix reveals the following critical diagnostic metrics defining the system's viability:

- **Sensitivity (True Positive Rate):** 94.3%
- **Specificity (True Negative Rate):** 96.5%
- **False Positive Rate (FPR):** 3.5%
- **False Negative Rate (FNR):** 5.7%

In the context of strict occupational safety, Sensitivity is the paramount metric. A sensitivity score of 94.3% indicates that the model successfully flags the vast majority of actual physical threats, satisfying the primary, life saving directive of the surveillance car. Concurrently, the exceptionally low False Positive Rate ensures that benign, recyclable materials are rarely quarantined unnecessarily, preserving the vital logistical throughput and economic viability of the waste management facility. The integration of these powerful cognitive capabilities into a live, stabilized interface is demonstrated in Fig. 3.

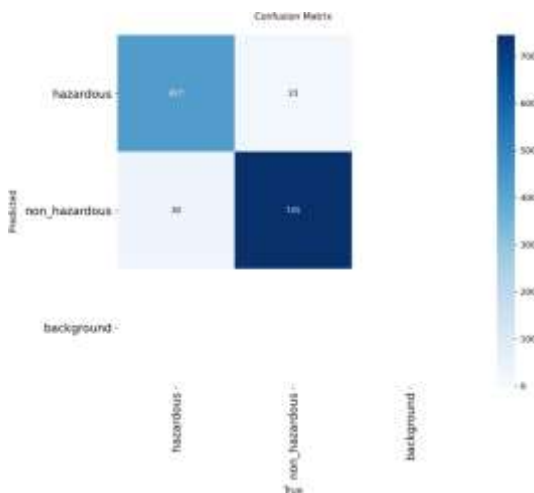


Fig. 3. Screenshot of the real time live inference Heads Up Display (HUD) successfully identifying non-hazardous waste with the physical hardware visible in the background.

VI. DISCUSSION AND OPERATIONAL CHALLENGES

While empirical testing heavily validates the proposed system architecture, transitioning AI driven robotics from controlled, curated datasets to chaotic, physical environments introduces specific operational challenges that must be addressed.

A. The Reality of the Domain Gap

The neural network was trained on high resolution, static images extracted from a pristine, curated dataset. However, the optical module mounted on the actively moving surveillance car produces footage afflicted by aggressive motion blur, restricted focal lengths, and harsh, variable outdoor lighting

conditions. This inherent discrepancy between clean training data and chaotic operational data is known within the field as the "domain gap."

During rigorous field testing, severe optical degradation occasionally caused the model's raw confidence scores to drop near the safety decision threshold. To comprehensively mitigate this in future industrial iterations, the physical chassis requires the integration of high-lumen, diffused LED fill lights to artificially normalize the illumination of the target zone regardless of the ambient weather. Furthermore, the model must undergo an active learning phase, where footage captured directly by the rover in actual municipal dump yards is manually annotated and injected back into the training pipeline, forcing the AI to adapt specifically to its own hardware constraints.

B. Hardware Bottlenecks and Energy Efficiency

The motor driver utilized in the prototype, while highly robust and accessible for research, relies on legacy Bipolar Junction Transistor architecture. This older technology results in significant thermal energy dissipation and a measurable voltage drop across the integrated bridge. Consequently, the driver actively wastes precious battery capacity purely as heat. Transitioning this component to a modern, MOSFET based motor driver would dramatically increase the overall electrical efficiency of the platform. This hardware upgrade would significantly extend the operational endurance of the surveillance car in the field while providing much higher frequency resolution for smoother, more precise low-speed navigation and maneuvering.

C. Edge Computing and Scalability

The current prototype relies entirely on a tethered processing architecture; the mobile rover must remain within the strict WiFi network range of the base station laptop that houses the massive GPU executing the deep learning model. To achieve true scalability and full autonomy across massive, multi acre landfill sites, the cognitive processing must be radically decentralized.

Transitioning to a true Edge AI paradigm replacing the current receiver architecture with an onboard Single-Board Computer equipped with native tensor processing capabilities would allow the rover to process high-resolution video entirely locally. This architectural shift would eliminate the crippling dependency on high-bandwidth WiFi networks, requiring only minimal, low-bandwidth radio telemetry to broadcast periodic safety alerts and GPS locations back to a central server.

VII. CONCLUSION

This research successfully designed, implemented, and validated the GestureBot AI, an advanced smart surveillance car engineered explicitly to modernize hazardous waste inspection. By synthesizing intuitive gesture kinematics with a highly optimized, single-stage deep learning classification pipeline, the system actively and physically decouples vulnerable human sanitation workers from acute occupational hazards.

The implementation of a strictly isolated dual-channel telemetry architecture guarantees phenomenal mechanical responsiveness without succumbing to inevitable video bandwidth interference. Achieving an overall classification accuracy of 95.7% with a critical 94.3% sensitivity to hazardous materials, the prototype confirms the absolute viability of integrating cost effective microcontrollers with state of the art artificial intelligence. This research establishes a scalable, deeply effective framework for human-machine collaboration, offering a highly practical technological solution to modern municipal solid waste management challenges. Future iterations will focus on transitioning the visual processing entirely to the edge device and integrating multi axis manipulators to evolve the platform from a passive surveillance tool into an active, autonomous segregation unit.

ACKNOWLEDGEMENT

We extend our profound gratitude to all Department of Electronics and Telecommunication Engineering Faculty Members for their expert guidance, academic rigor, and continuous support throughout the development of this research. We also acknowledge the Department of Electronics and Telecommunication Engineering, JSPM Narhe Technical Campus, Savitribai Phule Pune University (SPPU), for providing the laboratory infrastructure and technical environment essential to completing this project.

REFERENCES

- [1] S. Sakib, M. Toha, J. A. Adreeta, and M. F. Muhtashim, "Hand Gesture Controlled Robotic Car Using ESP32," *American International University-Bangladesh*, Dhaka, Bangladesh, 2024.
- [2] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, pp. 9243-9275, 2023.
- [3] C. K. Wen and L. C. Yong, "Gesture Controlled Robotic Arm For Hazardous Chemical Control," *JATI Journal of Applied Technology and Innovation*, vol. 6, no. 3, pp. 47-55, 2022.
- [4] T. Dorigo et al., "Artificial Intelligence in Science and Society: The Vision of USERN," *IEEE Access*, vol. 13, 2025.
- [5] I. Ahammed, M. K. A. Rakib, M. Y. Rahman, J. Mawa, and A. Obaidah, "Edge-AI Vision Surveillance Robot: Real-Time Object Detection and IoT-Driven Autonomous Navigation," in *Proc. 7th Global Power, Energy and Communication Conference (GPECOM)*, Bochum, Germany, 2025.
- [6] S. A. Bowyer and F. Rodriguez y Baena, "Dissipative Control for Physical Human-Robot Interaction," *IEEE Transactions on Robotics*, vol. 31, no. 6, pp. 1281-1293, Dec. 2015.
- [7] C. Herkariawan, N. R. S. Muda, D. Minggu, E. Kuncoro, R. Agustyadi, and M. L. N. Suryana, "Design control system using gesture control on the Arduino-based robot warfare," *IOP Conference Series: Materials Science and Engineering*, vol. 1098, 032013, 2021.
- [8] R. Setiawan, H. H. Triharminto, and M. Fahrurrozi, "Gesture Control Menggunakan IMU MPU 6050 Metode Kalman Filter Sebagai Kendali Quadcopter," *Prosiding Seminar Nasional Sains Teknologi dan Inovasi Indonesia*, vol. 3, pp. 411-422, 2021.
- [9] L. M. Xin and N. F. N. Rosely, "Development of Machine Learning Wildlife Camera Using ESP32 CAM for Small Mammals," *Pertanika J. Sci. & Technol.*, vol. 33, no. 4, pp. 1783-1807, 2025.
- [10] M. Meiswinkel, A. Ridder, and A. Ebrahimi, "Contactless Telemetry System for Condition Monitoring of Large Hydro and Wind Generators," *IEEE Access*, 2026.
- [11] J. Sun and L. Qiu, "Human-Machine Collaboration: Definitions, Models, and Socio-Economic Impacts," *International Journal of Crowd Science*, 2025.
- [12] P. A. Rajeev et al., "Advancing e-waste classification with customizable YOLO based deep learning models," *Scientific Reports*, vol. 15, no. 18151, 2025.