Smart HomeGuard: Real-time Anomaly Detection for Domestic Safety

1st Santoshkumar Tongli
Dept. of Computer Science
Colorado State University
Fort Collins, USA.
tskumar@colostate.edu

2nd Shraddha Patil
Dept. of Electrical Engineering
Colorado State University
Fort Collins, USA.
Paatil98@colostate.edu

***Abstract*—Anomaly detection, a fundamental method for identifying outliers, encompasses manual selection, empirical, and the increasingly influential role of neural networks. Leveraging their ability to learn from experiments and predict based on probability distributions, neural networks have become a powerful tool in this context. Despite extensive applications in manufacturing, public safety, and healthcare, there is a discernible gap in research concerning anomalies within domestic home environments, a critical aspect of resident safety.**

This project addresses this gap through a targeted focus on anomaly detection in and around domestic homes, utilizing both indoor and outdoor camera feeds. The primary goal is to proactively enhance resident safety by promptly alerting homeowners and authorities to potential hazards, including unauthorized entries, vandalism, theft, and other anomalous events. The proposed system employs Convolution Neural Network architectures to identify anomalies and swiftly communicates them to homeowners through the dispatch of event images and warning messages to their mobile applications. In essence, this project strives to significantly elevate residential security by deploying anomaly detection within the specific context of domestic homes.

***Index Terms*—Video based Anomaly, Convolution Neural networks, MLP, Bert Classifier.**

I. Introduction

In the realm of home security, the project at hand addresses the critical need for real-time video anomaly detection, representing a pivotal advancement in fortifying residential safety. With a keen focus on both indoor and outdoor camera feeds, the system endeavors to swiftly identify and alert homeowners and authorities to potential threats, thereby serving as a vigilant safeguard for resident safety. The overarching objective is to establish a robust preemptive shield, ensuring the continuous safety and well-being of residents.

This proposed solution unfolds as a comprehensive pipeline intricately designed for detecting anomalous activities within domestic settings. Leveraging advanced technologies such as Convolution Neural Network (CNN)-based feature extractors and Long Short-Term Memory (LSTM) models, the system engages in a sophisticated analysis of video data, capturing temporal sequences to discern anomalies effectively. The categorization into Anomalous and Normal classes employs a strategic blend of supervised and unsupervised methods, coupled with ongoing experimentation for optimization. Within the anomaly class, detailed classification reaches a new level through the integration of dense layers or Multi-Layer Perceptron (MLP) architectures, enhancing the granularity of anomaly detection.

This holistic pipeline represents a collaborative effort aimed at creating an efficient and adaptive solution for proactive anomaly detection in home security. The anticipated results encompass a heightened level of safety for residents, achieved through timely identification and response to potential threats.

In the broader context of video-based anomaly detection, industry leader Ring exemplifies recent advancements in technology. Utilizing CNNs and LSTM models, Ring has elevated its capability to analyze video feeds from diverse devices, integrating sophisticated anomaly detection algorithms that distinguish between routine and potential security threats. The commitment to ongoing experimentation, including the incorporation of dense layers and MLP architectures, underscores the continuous evolution of these technologies. Despite the effectiveness of video anomaly detection, a notable drawback lies in the potential for false positives, necessitating a delicate balance between sensitivity and specificity. Ongoing advancements, informed by user feedback and real-world data, contribute significantly to the precision and effectiveness of home security systems. In essence, this project, positioned at the forefront of the video anomaly detection landscape, emerges not only as a solution to core security concerns but also as a substantial contributor to the continued evolution of this transformative technology, laying the foundation for more robust and intelligent home security systems.

II. Related Work

It encompasses a diverse array of methodologies and advancements in the realm of anomaly detection within video surveillance. Notably, the ”Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection” paper introduces an innovative strategy leveraging unlabeled data for enhanced model understanding. In ”Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection,” a probabilistic approach using Bayesian nonparametric methods is emphasized for robust anomaly detection, addressing

uncertainty modeling. The "Deep Anomaly Discovery from Unlabeled Videos" paper explores deep learning methods, incorporating normality advantage and self-paced refinement, indicating a nuanced strategy for progressively refining anomaly detection capabilities. "Generative Cooperative Learning for Unsupervised Video Anomaly Detection" hints at collaborative generative-discriminative model use, possibly GANs, for unsupervised anomaly detection. The paper "UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection" contributes to benchmarking by introducing the "UBnormal" dataset for standardized evaluation. "Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection" addresses the challenge of utilizing limited annotated data in weakly supervised learning. "EVAL: Explainable Video Anomaly Localization" stands out for its focus on providing model decision explanations, enhancing interpretability. "Prompt-Guided Zero-Shot Anomaly Action Recognition using Pretrained Deep Skeleton Features" innovatively explores zero-shot anomaly recognition guided by prompts. "Look Around for Anomalies: Weakly-supervised Anomaly Detection via Context-Motion Relational Learning" introduces a strategy exploiting relationships between context and motion cues for weakly supervised anomaly detection. "Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping" suggests a method for artificially creating anomalous instances, beneficial for dataset augmentation. "A New Comprehensive Benchmark for Semi-supervised Video Anomaly Detection and Anticipation" contributes to benchmark development in semi-supervised video anomaly detection. The focus on hierarchical semantic contrast in "Hierarchical Semantic Contrast for Scene-aware Video Anomaly Detection" underscores capturing nuanced semantic information. "Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection" addresses unbiased learning challenges in weakly supervised scenarios. "Win CLIP: Zero-/Few-Shot Anomaly Classification and Segmentation" excels in classifying anomalies leveraging CLIP, a vision-language model. "Normalizing Flows for Human Pose Anomaly Detection" applies normalizing flows for sophisticated human pose anomaly detection. "TeD-SPAD: Temporal Distinctiveness for Self-supervised Privacy-preservation for Video Anomaly Detection" focuses on privacy preservation in video anomaly detection. Survey papers like "Anomalous Instance Detection in Deep Learning: A Survey" and "Self-Supervised Anomaly Detection: A Survey and Outlook" offer comprehensive overviews. Highlighted algorithms cover aspects of unsupervised outlier detection, end-to-end unsupervised outlier detection, deep anomaly detection, outlier exposure, autoencoding Gaussian Mixture Models, active anomaly detection, and Meta-AAD with deep reinforcement learning, contributing to diverse strategies in anomaly detection advancement. Collectively, this literature provides a valuable repository for researchers and practitioners aiming to enhance anomaly detection capabilities in video surveillance applications.

III. Dataset and Pre-Processing

To work on this project, to get the right dataset itself was more challenging task. based on our literature survey on dataset and

A. key Anomalies

Our focus revolves around three key cases of anomaly detection:

1) Human Activity Recognition
    a) Critical Anomalies
        i) Vandalism
        ii) Car damage and theft
        iii) Weapon violence (in front of the main door)
        iv) Arson
    b) Less Critical Anomalies
        i) Fights near your neighborhood
        ii) Kicking someone near the neighborhood
        iii) People running in panic
        iv) Individuals suddenly falling on the ground
        v) Person entering the door

Our initial step was to collect data from various sources to ensure a comprehensive and diverse representation of indoor and outdoor scenarios for anomaly detection in home environment, this task was challenging as it there was no relevant exsisting datasets, thus required to drive from multiple sources. The datasets we have collected are as follows:

1) **Smart-City CCTV Violence Detection Dataset (Kaggle)**: - A dataset specifically designed for violence detection using CCTV footage.
2) **CCTV Surveillance Dataset**: - Handpicked images containing instances of individuals with weapons, fighting scenes, and packages left unattended outside homes. (3000+ images)
3) **NTU CCTV-Fights Dataset**: - A valuable resource obtained from link for detecting physical altercations captured by CCTV cameras.
4) **Human Detection Dataset (Kaggle)**: - A dataset focused on human detection, to enhance normal human activities class.
5) **SPHAR: Surveillance Perspective Human Action Recognition Dataset**: - Utilized from link for human action recognition in surveillance settings.
6) **Large Scale Multi-Camera Detection Dataset (Kaggle)**: - High-quality video frames capturing normal activities, a vital addition to normal class.
7) **ChokePoint Dataset**: - Acquired from link for a unique perspective (people entering the main door) on surveillance scenarios.
8) **UCF-Crime Dataset**: - An essential dataset for crime detection, accessed from link.

After completing the data collection phase, our next challenge was to intricate the process of sorting and selecting the data based on our specific use case. This task demanded manual efforts to be done significantly due to the diverse

nature of the datasets we have collected. Despite having some prior information about the data sources, we encountered the need to fine-tune our focus and align the datasets with the nuances of our anomaly detection objectives.

To address this challenge, we undertook a careful examination of the gathered data, seeking to identify patterns and characteristics relevant to our defined anomalies. This involved refining our criteria to detect specific types of anomalies, such as unauthorized entry, vandalism, or suspicious activities. The process of data segregation ensued, where we systematically organized and categorized the datasets based on the refined criteria. This meticulous categorization allowed us to create subsets of data that were particularly relevant to each type of anomaly we aimed to detect.

Once the data collection and selection process were completed, we determined that there were initially 12 anomaly classes. However, three of these classes had an insufficient number of samples, specifically 7, 12, and 9 samples each. Consequently, we opted to exclude these classes, resulting in the final dataset containing only 9 anomaly classes. The class distribution of the anomaly data is illustrated in the figure below. 1

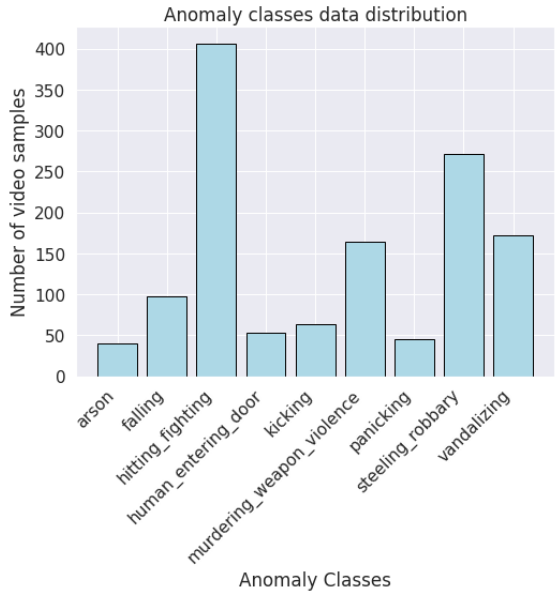

Fig. 1. Collected anomaly classes data distribution

Subsequently, a challenge arose concerning the varying lengths of videos across classes. The dataset comprised approximately 3300 videos, with durations ranging from 15 seconds to 106 minutes. Directly training models on videos of such diverse lengths posed significant difficulties. To address this issue, we performed class-specific data processing, breaking down longer-duration videos into multiple shorter segments. Overall, we limited the maximum video duration to 5 minutes, leading to an increased total number of videos in the dataset, approximately 5700.

Once the dataset was refined, the next step involved training the model. Given the need to input videos into the model, even with the reduced maximum duration of 5 minutes and a frame rate of 30 frames per second (fps), a single video yielded approximately 9000 frames. To streamline the training process, there was a need for pre-processing function to select frames from the videos. A standard approach of frame selection involved selecting frames with a constant frame step (in our case 5), resulting in 1800 frames. However, further refinement was pursued with the frame selection process. Many models commonly employed a random function to select the initial frame and subsequently picked additional frames based on the provided frame step until reaching the desired number of frames. This random approach sometimes led to the inclusion of frames where nothing significant was occurring. To address this, a custom frame selection function was developed. Our frame selection process involved calculating the absolute difference between consecutive frames and counting the number of non-zero pixels. Each frame was scored based on this value, with the frame exhibiting the highest score indicating the maximum number of pixel changes. Frames were then selected based on these scores, without changing the sequence of frames. Typically, 30 to 60 frames were chosen from each video, providing a subset of frames deemed most informative for model training.

In our analysis, we conducted a comparative study of the computational efficiency between a random frame selection method and our proposed method, termed "Smart Select," when extracting frames from a 5-minute video. Remarkably, both methods exhibited comparable processing times, indicating that the overhead incurred by our algorithm was negligible relative to the random selection approach.

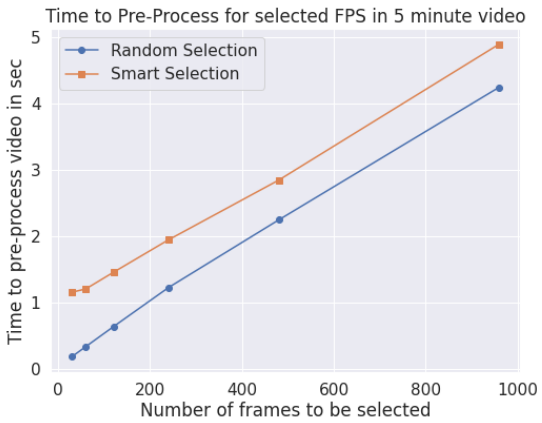

Fig. 2. Plot shows the time taken to select different number of frames from a 5 minutes video

However, we delve further into the temporal dynamics of frame selection by introducing a plot that illustrates the computational time required for selecting frames across videos of varying durations. As observed in the plot, our algorithm's computational time exhibits an exponential increase as the video duration extends. It is imperative to note that this exponential trend may present challenges in scenarios where

video durations are considerably long.

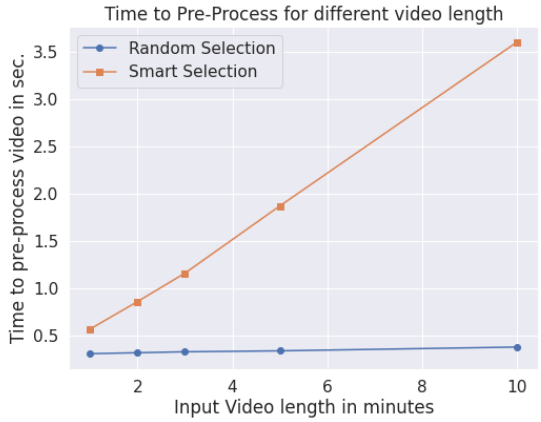

Fig. 3. Plot shows the time taken to select the frames for a given video with different length

Nonetheless, In this case, we have a predetermined constraint on video duration (limited to 5 minutes), the exponential increase in computational time, albeit observed in the analysis, remains a manageable concern. Consequently, the restriction imposed on video duration ensures that the computational overhead associated with our algorithm remains within acceptable bounds for the defined application scope.

IV. Methodology

After the dataset preparation phase, we embarked on the model training process for anomaly detection in smart home security. However, a significant challenge arose during model training due to the substantial size of the dataset, amounting to 25 GB. Working with such a voluminous dataset in real-time during the training process posed considerable difficulties. To overcome this challenge, efficient load functions were essential, capable of handling large data volumes without causing system memory crashes. To address this, we implemented custom preprocess functions specifically designed for video data and leveraged TensorFlow AUTOTUNE to automatically optimize the data loading process.

Once we had a proper and stable data loader, we initiated our training process. In the early stage, we employed pre-trained Convolutional Neural Network (CNN) models to extract features from individual frames within the video sequence. The resulting features, represented as embedding vectors, varied in size depending on the chosen pre-trained models. To capture temporal relationships within these embedded vectors, we utilized Long Short-Term Memory (LSTM) models. Each input video was sampled using 30/60 frames to capture the underlying events effectively. The LSTM models were configured with an input sequence length of 30/60 frames and comprised three layers, each containing 250 LSTM units. This architecture aimed to capture and understand temporal dependencies within the video data, laying the foundation for effective anomaly detection in the context of smart home security.

Throughout the model training process, we experimented with different pre-trained models as feature extractors. In the experiment section, we compared the accuracies achieved by selecting various pre-trained models. Unfortunately with this approach, we did not achieve better performance in identifying anomalous events. This is because the feature extractors used for the embedding generation were originally trained for classification purposes and were not specifically tailored for video-based datasets. Additionally, we were separately learning the lower-level representation of frames and their temporal characteristics. To address this limitation, we transitioned to a unified model using 3D convolution. The 3D convolutional model comprised three layers of convolution, resulting in an improvement in accuracy compared to the previous approach. This shift allowed us to integrate both feature extraction and temporal relation learning within a single model, enhancing the overall performance of anomaly detection in the smart home security system.

Once the dataset was prepared, the model training process for anomaly detection in smart home security was initiated. Initially, pre-trained Convolutional Neural Network (CNN) models were employed on individual frames to extract features from the sequence of images. The resulting features, or embedding vectors, varied in size based on the chosen pre-trained models. To capture temporal relationships within these embedded vectors, Long Short-Term Memory (LSTM) models were utilized. Each input video was sampled using 30/60 frames to understand the events. The LSTM models were configured with an input sequence length of 30/60 frames and had three layers, each containing 250 LSTM units.

Throughout the model training process, different pre-trained models were experimented with as feature extractors. In the experimental section, the accuracies achieved by selecting various pre-trained models were compared. However, this approach did not yield better performance in identifying anomalous events. The limitation arose because the feature extractors used for embedding generation were originally trained for classification purposes and were not specifically tailored for video-based datasets. Furthermore, lower-level representation of frames and their temporal characteristics were separately learned. To address this limitation, a transition was made to a unified model using 3D convolution. The 3D convolutional model comprised three layers of convolution, resulting in an improvement in accuracy compared to the previous approach. This shift allowed for the integration of both feature extraction and temporal relation learning within a single model. In addition to the transition to 3D convolution, a further enhancement of incorporating residual 3D convolution model was introduced. In this model, the Conv3D layers were paired in blocks, with the first having a filter shape (1, k, k) and the subsequent 3D layers having a filter shape of (k, 1, 1). Leveraging a (2 + 1)D convolution with residual connections, which decomposes spatial and temporal dimensions, contributed to parameter efficiency. This model, combined with our weighted loss, achieved a testing accuracy of 65.33 %.

After successfully training the model, the next step involves deploying it on an embedded system. For deployment purposes, we have opted to utilize Raspberry Pi, a popular and versatile embedded platform. The inference process on the Raspberry Pi executes the model, producing anomaly events. These events are subsequently updated in the cloud infrastructure by providing a keyframe extracted from the anomaly video, serving as a visual representation of the event, along with accompanying textual information. The Firebase Cloud platform is employed for this purpose, providing a robust cloud storage and real-time database solution.

To facilitate seamless communication between the embedded system and the cloud, an Android application has been developed. This application maintains continuous interaction with the Firebase Cloud, periodically retrieving data whenever there is a new update in the cloud. The Android application acts as a user interface, ensuring timely and efficient access to the latest information regarding anomaly events. The integration of the Raspberry Pi, Firebase Cloud, and the Android application form a cohesive system for real-time anomaly detection and user notification in a smart home security context.

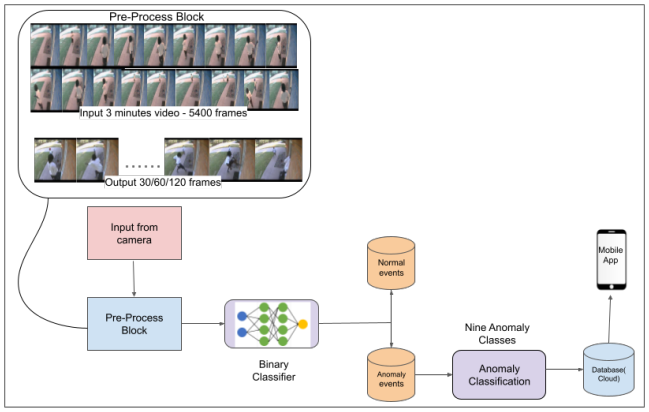

Fig. 4. Block diagram of Proposed Algorithm

V. Experiments

In the initial phases of our project, which revolved around video input, our primary focus was on developing effective frame selection techniques. Once this foundational step was accomplished, we embarked on training a binary classifier to classify between anomaly and normal videos. Our approach involved utilizing a Convolutional Neural Network (CNN)-based feature extractor to transform the video frames into a latent space. The sequential embedding vectors were then passed through Long Short-Term Memory (LSTM) models to capture temporal relationships. In the pursuit of optimizing our model, we explored different feature extractors, incorporating pre-trained classification models initially trained on the ImageNet dataset. Noteworthy choices among these were VGG16, MobileNetV2, InceptionV3, and InceptionResNetV2. We extracted features from these models either at last convolution layers or before the output layer. When we performed the tapping into the model output at the convolutional layer, we employed global average pooling to obtain the embedded vectors. These embeddings served as input to LSTM blocks, with a sequence length of 30/60. The designed model consisted of three layers of LSTM, each with 256 LSTM units.

Despite our efforts, these models did not perform satisfactorily in terms of classification accuracy. This may be attributed to the fact that the feature extractors used were originally trained for classification and not specifically tailored for video-based datasets. Moreover, our approach involved separately learning the lower-dimensional representation of frames and their temporal characteristics.

To address this limitation, we shifted our focus to models capable of simultaneously learning spatial and temporal content, leading us to 3D convolutions. These layers excel in capturing spatiotemporal features from the input data. Our designed model featured three Conv3D layers interleaved with three 3D max-pooling layers. The number of filters increased across 3 conv layers (16, 32, and 64), culminating to one dense layer with 128 neurons and further an output layer with 2 neurons. This Conv3D model achieved a 75% accuracy in anomaly classification. A variant of the model, trained with a weighted loss function (categorical focal cross-entropy), achieved a 79% test accuracy. This model has 11.14 million parameters and a 45 MB file size.

Upon establishing this binary classifier, our next challenge was to localize identified anomalies. We employed the same Conv3D-based model to train a 9-class anomaly event classifier, achieving an accuracy of 52.49%. The complexity arose from the disparate video qualities across classes and the occurrence of similar backgrounds in some anomaly events. To enhance model performance, we conducted numerous experiments, with a notable improvement observed when employing a residual 3D convolution-based model. The skip connections in this model facilitated better spatial representation learning. Notably, the Conv3D layers were paired in blocks, with the first having filter shape (1, k, k) and the subsequent 3D layers having a filter shape of (k, 1, 1). Leveraging a (2 + 1)D convolution with residual connections, which decomposes spatial and temporal dimensions, contributed to parameter efficiency. This model, combined with our weighted loss, achieved a testing accuracy of 65.33% with 440,000 parameters. Once this training of the model is completed, the next step is deployment.

We deploy our trained model on the Raspberry Pi. The Raspberry pi we used to deploy the model has 8 GB RAM. The inference process takes place locally on the Raspberry Pi, leveraging image processing and machine learning techniques. The identified anomalies are subsequently transmitted to the cloud infrastructure for further processing and user notification, utilizing Firebase Cloud as the chosen cloud service. Upon the identification of an anomaly event, a keyframe is extracted from the anomaly video to represent the event visually. This selected image, along with relevant textual

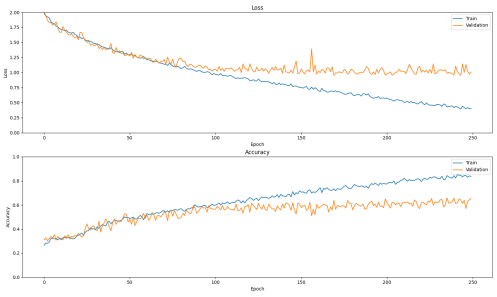

Fig. 5. Nine class anomaly classifier training and validation curve.

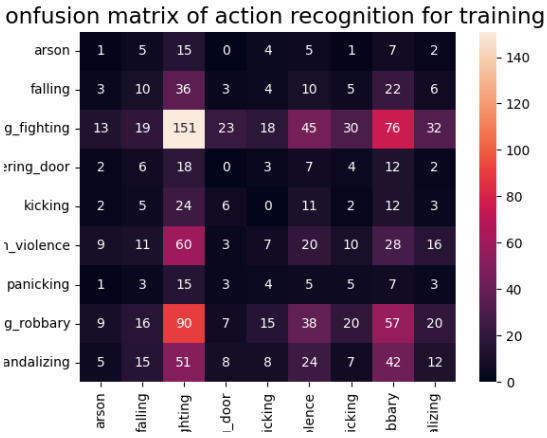

Fig. 6. The table presents the Confusion Matrix for nine class anomaly classifier training data

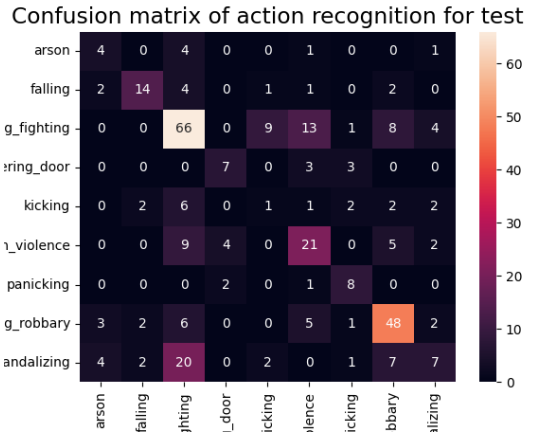

Fig. 7. The table presents the Confusion Matrix for nine class anomaly classifier testing data

SL. No.	No of Classes	Weighted Loss	Model Type	Train Accuracy	Val Accuracy	Feature Model Params	LSTM Model Params
1	2	no	vgg + lstm	48.19	49.66	138.9M	2.65M
2	2	yes	vgg + lstm	59	53	138.9M	2.65M
3	2	no	mobilenetv2 + lstm	56	48	3.5M	2.65M
4	2	yes	mobilenetv2 + lstm	53	51	3.5M	2.65M
5	2	no	Inceptionv3 + lstm	61	58	23.9M	2.65M
6	2	yes	Inceptionv3 + lstm	61	56.7	23.9M	2.65M
7	2	no	InceptionResnetv2 + lstm	60	58	55.9M	2.65M
8	2	yes	InceptionResnetv2 + lstm	63.6	55.8	55.9M	2.65M
9	9	yes	InceptionResnetv2 + lstm	48.6	50	55.9M	2.65M

Fig. 8. The table presents the accuracy results obtained through the CNN + LSTM approach.

SL. No.	No of Classes	Weighted Loss	Model Type	Train Accuracy	Val Accuracy	Model Params
1	2	yes	3d conv	83.6	79	11.14M
2	2	yes	3d conv	84.7	75.3	11.14M
3	9	no	3d conv	60.3	52.69	11.14M
4	9	yes	3d conv + focal loss with single alpha	90.71	61.45	11.14M
5	9	yes	3d conv + focal loss with multiple alpha	93.8	59.6	11.14M
6	9	yes	Residual + Conv2D+1D using Conv3D	83.6	65.33	0.44 M
7	9	yes	Residual + Conv2D+1D using Conv3D double the above model size	89.9	54.2	0.44 M

Fig. 9. Block diagram of Proposed Algorithm

information, is then uploaded to Firebase Cloud's storage class and real-time database, respectively. The storage class handles the visual representation of the anomaly event, while the real-time database stores associated metadata.

As the data propagates to Firebase Cloud, an Android application developed for this purpose interacts with the real-time database to check for the latest updates. The real-time database serves as a dynamic repository of information about anomaly events. Based on the most recent update in the real-time database, the application identifies that a new anomaly has occurred. Subsequently, the application initiates a data retrieval process from Firebase Storage to obtain additional information about the anomaly, specifically retrieving images associated with the anomaly event.

In the user application's interface, the first page provides users with a visual representation of the identified anomaly. Users can view the selected anomaly image along with the timestamp indicating when the event occurred. This user-friendly presentation ensures that relevant information about anomaly events is easily accessible to users through the developed Android application. The integration of image processing, machine learning, and cloud services facilitates a seamless flow of information from local inference to cloud-based processing, enhancing the overall effectiveness of the anomaly detection system.

VI. Conclusion

In conclusion, this project the "Smart HomeGuard: Real-time Anomaly Detection for Domestic Safety" attempts to bridge the existing void in anomaly detection research within domestic home environments, an imperative aspect of resident safety. Spearheaded by Santoshkumar Tongli and Shraddha Patil from Colorado State University, we worked to Through the integration of advanced machine learning techniques, specifically leveraging 3D Convolutional Neural Networks (CNNs), our project contributes to the proactive enhancement of residential security. By inspecting indoor and outdoor camera feeds, our system identifies anomalies, ranging from individual entering house to potential hazards, such as vandalism and theft.

The significance of this work extends to the meticulous handling and pre-processing of video-based data, emphasizing the critical role of dataset quality, especially in addressing class imbalances. Our proposed pipeline integrates two distinct models: a binary classifier, employing 3D convolution, to detect the presence of anomalies, and a residual 3D convolution model to categorize the type of anomaly identified. This approach of identifing the anomaly enhances the depth of understanding of the data.

Furthermore, our deployment strategy involves running the trained models on a Raspberry Pi for real-time video stream analysis. Upon anomaly detection, pertinent information is promptly transmitted to the cloud, ensuring swift notification to the user. This holistic approach not only elevates the technical prowess of anomaly detection but also underscores its practical application in safeguarding domestic homes. As the realms of artificial intelligence and residential security converge, our project serves as a testament to the potential of cutting-edge machine learning methodologies in addressing real-world safety concerns.

REFERENCES

[1] Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection

[2] Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection - link

[3] Deep Anomaly Discovery from Unlabeled Videos via Normality Advantage and Self-Paced Refinement

[4] Generative Cooperative Learning for Unsupervised Video Anomaly Detection

[5] UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection

[6] Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video