REAL-TIME QUARREL DETECTION IN SURVEILLANCE VIDEOS USING DEEP LEARNING-BASED OBJECT AND ACTION RECOGNITION

Aditya Kulkarni[1], Satyam Modi[2], Onkar Vyawahare[3], Aditya Dubberwar[4], Mrs. Nitu L. Pariyal[5]

[1- 4] Department of Computer Science and Engineering,
MGM's College of Engineering, Nanded
Emails: {kulkarniaditya262, satyamodi551, Onkarvyawahare04, Aditya.dubbewar01}@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering,
MGM's College of Engineering, Nanded
Email: pariyal_nitu@mgmcen.ac.in

ABSTRACT

Automatic detection of violent and quarrel-related activities in surveillance videos is essential for ensuring public safety. This paper proposes a real-time quarrel detection framework that integrates YOLOv8 for efficient localization of human interactions and MobileNetV2 for lightweight behavior classification. The hybrid architecture effectively distinguishes quarrel and non-quarrel scenes while maintaining low computational overhead. Experimental evaluation on benchmark datasets, including RLVS and SCFD, demonstrates high detection accuracy and real-time inference performance, making the proposed system suitable for practical surveillance applications.

KEYWORDS: Violence Detection, Quarrel Recognition, YOLOv8, MobileNetV2, Deep Learning, Surveillance

1. INTRODUCTION

The increasing deployment of surveillance systems in public and private environments has intensified the demand for intelligent video analytics capable of automatically identifying violent and quarrel-related activities. Public spaces such as transportation hubs, educational institutions, commercial complexes, and residential areas generate vast volumes of surveillance footage, making continuous human monitoring inefficient, error-prone, and economically impractical.

As a result, critical incidents such as physical altercations or aggressive behavior may go unnoticed or be detected too late for effective intervention.Traditional surveillance systems primarily depend on manual observation or simple motion-based triggers, which often fail to distinguish between normal human interactions and genuinely aggressive behavior. Variations in lighting conditions, crowd density, camera angles, and background clutter further complicate accurate detection. These challenges have motivated researchers to explore deep learning-based approaches that can automatically learn discriminative visual patterns associated with violent and non-violent activities directly from video data.

2. RELATED WORK

Early research on automated violence detection primarily relied on convolutional neural network (CNN) architectures applied to surveillance footage. Evany et al. [1] and Dayes Joseph et al. [5] proposed real-time CNN-based systems that utilized frame-level visual cues to identify violent activities in CCTV videos. These approaches demonstrated the feasibility of automated violence detection but showed limitations in handling complex interactions and background noise. Akter and Islam [4] further enhanced surveillance systems by integrating deep learning models for crime monitoring, emphasizing real-time alert generation in smart surveillance environments.

Subsequent studies focused on improving detection robustness by incorporating motion-aware and temporal modeling techniques. Sreelakshmi and Srividya [2] improved violence recognition accuracy by combining motion segmentation with deep learning classifiers, enabling better differentiation between aggressive and non-aggressive actions. Karthikeyan and Priya [3] evaluated transformer-based architectures against state-of-the-art CNN models, demonstrating improved temporal attention and contextual understanding. However, transformer models incur high computational costs, making them less suitable for real-time deployment in resource-constrained surveillance systems.

Publicly available datasets such as the Fight Detection Surveillance Dataset (SCFD) [6], Real-Life Violence Situations (RLVS) dataset [7], Movies Fight dataset [8], and curated violence datasets from Roboflow Universe [9] have played a critical role in advancing research in this domain. These datasets provide diverse real-world and synthetic scenarios essential for training and evaluating violence detection models. Despite these advancements, there remains a trade-off between detection accuracy and real-time performance. This study addresses this gap by proposing a hybrid, lightweight architecture that combines fast object detection with efficient behavior classification to achieve accurate quarrel detection under real-time constraints.

3. MOTIVATION

Violence and quarrels in public and private spaces pose serious threats to human safety, property, and social order. Incidents occurring in locations such as streets, public transport systems, educational institutions, workplaces, and residential complexes often escalate rapidly, requiring immediate intervention. Relying solely on manual monitoring of closed-circuit television (CCTV) feeds is highly impractical due to the large volume of video data, limited human attention span, and the potential for delayed or inconsistent responses. Moreover, conventional surveillance systems lack the intelligence to autonomously interpret complex human behaviors and differentiate between normal interactions and aggressive confrontations. As a result, critical incidents may go undetected until after significant harm has occurred. These limitations highlight the necessity for automated quarrel detection systems that can continuously analyze surveillance footage and provide timely alerts to security personnel or authorities.

4. PROBLEM DEFINITION

The primary challenge addressed in this research is the accurate identification of quarrel-related activities from continuous surveillance video streams captured in real-world environments. Surveillance footage is often affected by varying lighting conditions, dynamic backgrounds, camera motion, occlusion, and crowd density, which makes reliable detection of aggressive behavior a complex task. Quarrel scenes frequently involve subtle motion patterns and close human interactions that are difficult to distinguish from normal activities such as casual conversations or playful gestures. Traditional frame-based and motion-thresholding techniques rely on low-level visual features and fail to capture the contextual and semantic information required to differentiate between violent and non-violent interactions. These methods are particularly sensitive to background clutter and partial occlusions, leading to high false-positive and false-negative rates. Additionally, variations in camera viewpoints and environmental noise further degrade detection performance.

Therefore, there is a need for an intelligent and robust system capable of continuously analyzing video streams, accurately localizing human interactions, and classifying quarrel behavior in real time. The problem is formally defined as the development of a computationally efficient and accurate quarrel detection framework that can operate under diverse surveillance conditions while maintaining low latency and high reliability suitable for real-world deployment.

$$\boldsymbol{P_q = w_v P_v + w_m P_m + w_a P_a}$$

where P_q denotes the final quarrel probability score, P_v, P_m, and P_a represents the probability outputs obtained from visual, motion, and audio analysis modules, respectively, and w_v, w_m, and w_a are their corresponding fusion weights such that w_v + w_m + w_a = 1.

5. PROPOSED SYSTEM

The proposed quarrel detection framework adopts a hybrid deep learning architecture that combines fast object detection with efficient behavior classification to enable real-time surveillance analysis. The system integrates YOLOv8 for detecting and localizing human interactions within video frames and MobileNetV2 for classifying the detected interactions into quarrel and non-quarrel categories. This separation of spatial localization and behavioral classification allows the system to process complex scenes efficiently while maintaining high detection accuracy.

Initially, continuous video streams acquired from surveillance cameras are segmented into individual frames and preprocessed to normalize resolution and illumination variations. YOLOv8 is then employed to perform real-time human detection by generating bounding boxes around individuals involved in potential interactions. By focusing only on regions of interest containing human activity, the system significantly reduces background noise and computational overhead.

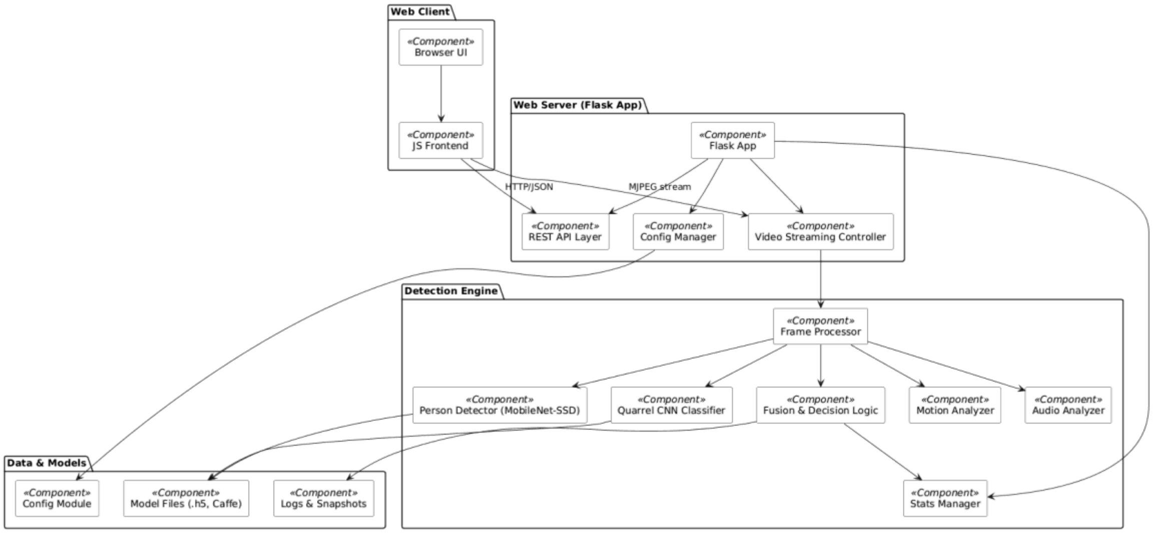
Web Client
«Component» Browser UI
«Component» JS Frontend
Web Server (Flask App)
«Component» Flask App
HTTP/JSON
MJPEG stream
«Component» REST API Layer
«Component» Config Manager
«Component» Video Streaming Controller
Detection Engine
«Component» Frame Processor
«Component» Person Detector (MobileNet-SSD)
«Component» Quarrel CNN Classifier
«Component» Fusion & Decision Logic
«Component» Motion Analyzer
«Component» Audio Analyzer
«Component» Stats Manager
Data & Models
«Component» Config Module
«Component» Model Files (.h5, Caffe)
«Component» Logs & Snapshots

Figure 5.1: System Architecture Diagram

The cropped regions corresponding to detected human interactions are subsequently passed to the MobileNetV2 network, which extracts discriminative spatial features using lightweight convolutional layers optimized for low-latency inference. A classification head then assigns each interaction to either a quarrel or non-quarrel class based on learned behavioral patterns. This dual-stage processing pipeline improves both accuracy and inference speed compared to monolithic deep learning models.

Figure 5.1 illustrates the overall architecture of the proposed quarrel detection system, highlighting the interaction between video acquisition, deep learning modules, and the final decision logic.

6. METHODOLOGY

The proposed quarrel detection system is trained and evaluated using multiple publicly available datasets to ensure robustness across diverse surveillance environments. The Fight Detection Surveillance Dataset (SCFD), Real-Life Violence Situations (RLVS) dataset, and Movies Fight dataset are utilized, as they contain a wide range of violent and non-violent interactions captured under varying lighting conditions, camera viewpoints, and crowd densities. The inclusion of multiple datasets improves the generalization capability of the model and reduces dataset-specific bias.

Model training is conducted using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16, which provides a stable balance between learning efficiency and computational overhead. To mitigate overfitting and improve model robustness, data augmentation techniques including horizontal flipping, random cropping, and minor rotational transformations are applied during training. Additionally, early stopping is incorporated by monitoring validation loss, ensuring that training halts once performance improvements plateau.

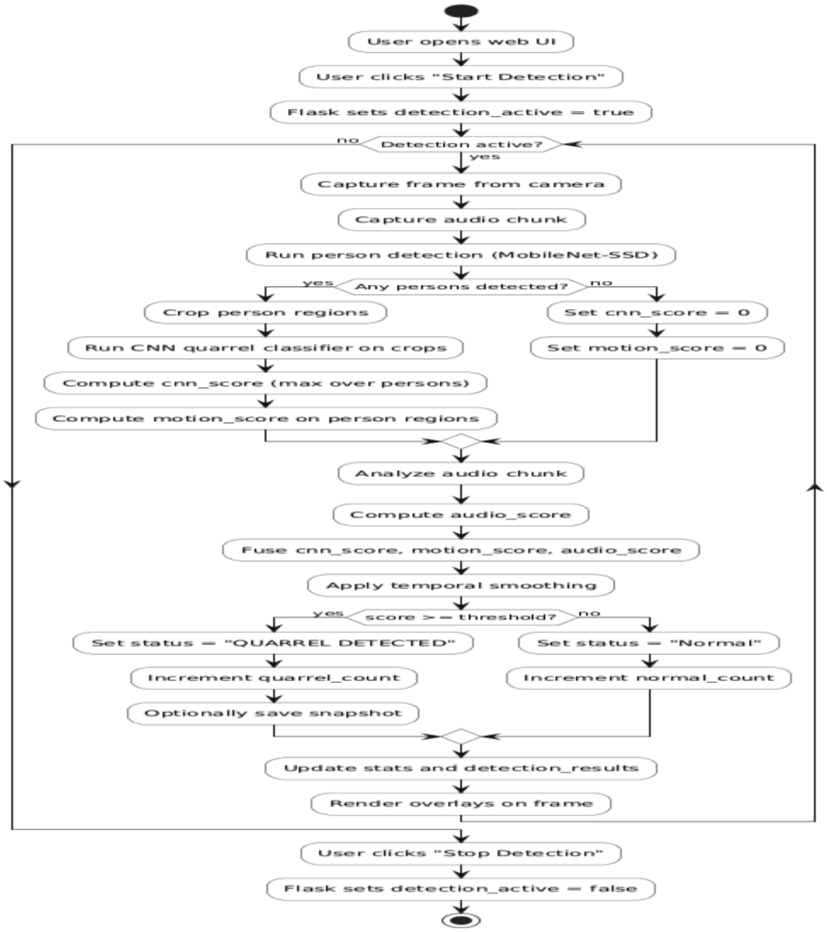
User opens web UI
User clicks "Start Detection"
Flask sets detection_active = true
no
Detection active?
yes
Capture frame from camera
Capture audio chunk
Run person detection (MobileNet-SSD)
yes
Any persons detected?
no
Crop person regions
Set cnn_score = 0
Run CNN quarrel classifier on crops
Set motion_score = 0
Compute cnn_score (max over persons)
Compute motion_score on person regions
Analyze audio chunk
Compute audio_score
Fuse cnn_score, motion_score, audio_score
Apply temporal smoothing
yes
score >= threshold?
no
Set status = "QUARREL DETECTED"
Set status = "Normal"
Increment quarrel_count
Increment normal_count
Optionally save snapshot
Update stats and detection_results
Render overlays on frame
User clicks "Stop Detection"
Flask sets detection_active = false

Figure 6.1: Activity workflow for real-time quarrel detection

The operational sequence of the proposed system is illustrated in **Figure 6.1**, which presents the activity workflow involved in real-time quarrel detection. This figure outlines the sequential stages starting from video acquisition and frame preprocessing, followed by human interaction detection, behavior classification, and final decision making. The activity workflow clarifies how each component contributes to the overall detection process and ensures continuous real-time monitoring.

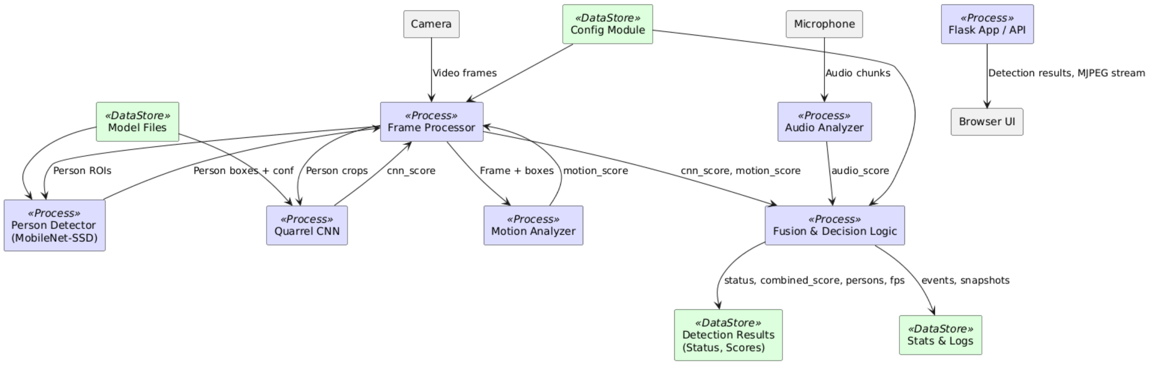

Figure 6.2: Data flow in the quarrel detection pipeline

The internal movement and transformation of data within the system are depicted in **Figure 6.2**, which shows the data flow in the quarrel detection pipeline. This figure highlights how raw video frames are progressively transformed into structured feature representations and classification outputs through successive processing modules.

7. RESULTS

The performance of the proposed quarrel detection framework was evaluated using multiple benchmark datasets to assess both detection accuracy and real-time efficiency. The hybrid YOLOv8 and MobileNetV2 model achieved an accuracy of **96.1%** on the Movies Fight dataset, **95.8%** on the Real-Life Violence Situations (RLVS) dataset, and **94.2%** on the Fight Detection Surveillance Dataset (SCFD). These results indicate consistent performance across diverse data sources, including both controlled and real-world surveillance scenarios.

In addition to classification accuracy, inference speed was evaluated to determine the suitability of the proposed system for real-time deployment. The model demonstrated an average inference time of **38 milliseconds per frame**, enabling processing at approximately **26 frames per second (FPS)** on standard hardware. This performance satisfies real-time constraints required for continuous surveillance monitoring, ensuring timely detection of quarrel-related activities without perceptible delay.

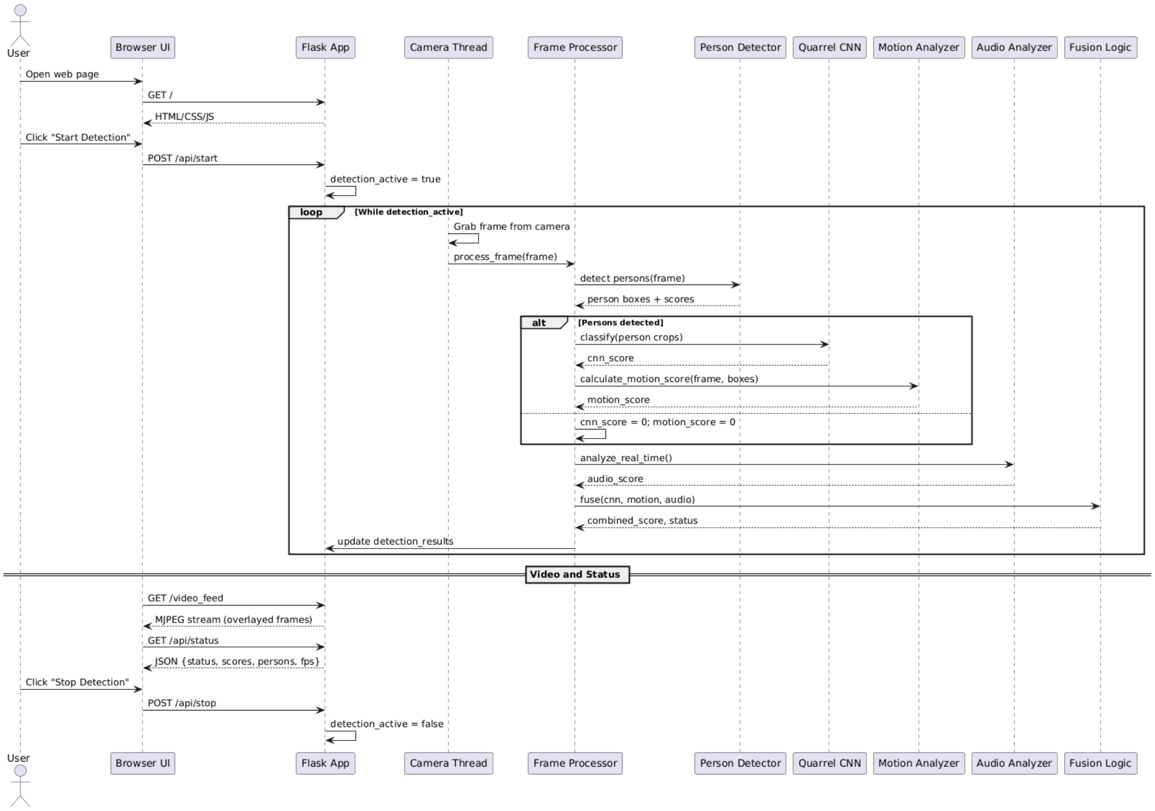

Figure 7.1: Sequence of Operations during real-time quarrel detection

The real-time execution flow of the proposed system is illustrated in **Figure 7.1**, which depicts the sequence of operations involved in quarrel detection. The sequence diagram highlights the interaction between video acquisition, object detection, feature extraction, classification, and alert generation modules. By visualizing the order and dependency of operations, Figure 4 demonstrates how the system efficiently processes incoming video frames while maintaining low latency.

Overall, the experimental results confirm that the proposed hybrid architecture effectively balances detection accuracy and computational efficiency. The combination of fast object localization and lightweight behavior classification enables reliable real-time performance, making the system suitable for deployment in practical surveillance environments such as public spaces, campuses, and transportation facilities.

8. COMPARISON

The proposed hybrid quarrel detection model is compared with existing deep learning-based approaches to evaluate its effectiveness in terms of detection accuracy and real-time performance. Earlier CNN–LSTM-based methods focus on modeling temporal dependencies but often suffer from high inference latency due to sequential processing. Transformer-based architectures improve contextual understanding and temporal attention; however, their high computational complexity limits their applicability in real-time surveillance systems, especially in resource-constrained environments.

In contrast, the proposed approach leverages the strengths of fast object detection and lightweight feature extraction by integrating YOLOv8 with MobileNetV2. This design enables efficient spatial localization of human interactions while maintaining low computational overhead during behavior classification. As a result, the system achieves superior performance in both accuracy and inference speed when compared to existing approaches.

Model	Technique Used	Accuracy (%)	Inference Time
CNN–LSTM Model	CNN + LSTM	89.3	120 ms
Transformer-Based Model	Attention-based	93.6	95 ms
Proposed Hybrid Model	YOLOv8 + MobileNetV2	**96.1**	**38 ms**

Table 1: Comparison of Proposed Model with Existing Approaches

The results presented in Table 1 clearly demonstrate that the proposed model outperforms existing CNN–LSTM and Transformer-based approaches in terms of accuracy while achieving significantly lower latency. This improvement highlights the effectiveness of the hybrid architecture in achieving real-time performance without compromising detection reliability.

9. CONCLUSION

This research presents an effective and computationally efficient framework for real-time quarrel detection in surveillance videos. By integrating YOLOv8 for rapid human interaction localization with MobileNetV2 for lightweight behavior classification, the proposed system successfully addresses the trade-off between detection accuracy and real-time processing speed. Experimental evaluation across multiple benchmark datasets confirms the robustness, generalization capability, and practical applicability of the proposed approach.

The results demonstrate that the hybrid architecture achieves high accuracy while maintaining low inference latency, making it suitable for continuous surveillance in real-world environments. The proposed system can serve as a reliable decision-support tool for security personnel by enabling early detection and timely intervention in quarrel and violence-related incidents.

10. FUTURE WORK

While the proposed system demonstrates strong performance, several avenues for future research remain. Temporal sequence modeling techniques such as LSTM or temporal convolutional networks can be integrated to further enhance the understanding of long-term interaction patterns. Additionally, deploying the system on embedded and edge devices such as NVIDIA Jetson platforms can improve scalability and reduce dependency on centralized computing resources.

Future enhancements may also include the incorporation of real-time alerting mechanisms, multi-camera fusion, and integration with smart city infrastructure to enable large-scale, intelligent surveillance systems. Expanding the model to handle multi-person interactions and incorporating audio-based cues could further improve detection accuracy and situational awareness.

ACKNOWLEDGEMENTS

The author would like to acknowledge the developers and maintainers of TensorFlow and Keras for providing robust deep learning frameworks. The author also thanks the contributors of the open-source datasets, including RLVS, SCFD, and Movies Fight, which enabled effective training and evaluation of the proposed quarrel detection system.

REFERENCES

[1] Evany, M. P., Joseph, D., and J. R. Jenitta, “Violence Detection in Real-Time for Surveillance,” *IRJET*, Vol. 7, Issue 6, 2020.

[2] Sreelakshmi, S., and M. Srividya, “Enhancing Violence Detection in Surveillance Video,” *IJISAE*, Vol. 11, Issue 5, 2023.

[3] Karthikeyan, M., and K. Priya, “Advanced Detection of Violence from Video: Performance Evaluation of Transformer and State-of-the-Art CNN Models,” *Procedia Computer Science*, Elsevier, Vol. 262, 2025.

[4] Akter, S. and Islam, S., “Intelligent Crime Surveillance Video System Using Deep Learning,” *ARASET Journal*, Vol. 57, No. 1, 2021.

[5] Dayes Joseph, D., et al., “Violence Detection in Real-Time for Surveillance,” *IJNRD*, Vol. 8, Issue 7, 2023.

[6] Karadeniz, S., and Akti, S., “Fight Detection Surveillance Dataset (SCFD),” *GitHub Repository*, 2022.

[7] Mustafa, M., “Real-Life Violence Situations Dataset (RLVS),” *Kaggle Dataset*, 2021.

[8] Naveen, K., “Movies Fight Detection Dataset,” *Kaggle Dataset*, 2020.