CROWDGUARD AI

Prema V[1], Selvagnapathy E[2], Rishitharan Y[3], Pravin P[4]
[1]*M.E, Assistant Professor,Department of CSE. Srm Valliammai Engineering College, Kattankulathu, Chengalpattu, India*
[2,3,4]*UG, Srm Valliammai Engineering College, Kattankulathu, Chengalpattu, India*

***Abstract*—In such large public events, it is indispensable to understand the crowd's behavior and the changes in their emotional state in real-time if one is to ensure their safety and be able to prevent the occurrence of any incidents. Traditional surveillance systems rely to a great extent on human monitoring, which is susceptible to operator fatigue, delay, and subjective bias. CrowdGuard AI overcomes this drawback by presenting an intelligent, real-time crowd emotion and anomaly detection system that combines YOLOv8, CNN, and Transformer-based attention mechanisms. CrowdGuard AI, which is powered by PyTorch and accessed via a Flask-based web interface, detects the presence of various types of emotions, e.g., fear, anger, happiness, or panic in a crowd and also notifies the police about any unusual activity.Moreover, the model offers explainable AI (XAI) reasoning through Grad-CAM visualizations and attention heatmaps, thereby making the predictions more understandable. The experiments conducted have shown that the model makes correct decisions 96.8% of the time when considering various datasets and thus it is better than the traditional vision models. The ability of the system to be scalable, transparent, and able to adjust to changes made in real-time is what makes it a perfect fit for smart city surveillance, event management, and disaster response scenarios.**

***Index Terms*—Deep Learning, YOLOv8, Crowd Emotion Recognition, Computer Vision, Transformer, Explainable AI, Real-time Surveillance, Public Safety.**

I. INTRODUCTION

The management of crowds as well as the safety of the public have become major problems of the cities of today, especially during large-scale events such as political rallies, concerts, religious festivals, and sports matches. Generally, these are situations of dense populations, where a tiny trigger -for instance panic, aggression, or sudden fear- can result in an uncontrollable riot. In fact, traditional surveillance systems are very much dependent on human operators to monitor numerous camera feeds, which is not only inefficient but also susceptible to fatigue and mistakes of the humans. Hence, the demand for a real-time, automated, and intelligent crowd monitoring system is still increasing.

CrowdGuard AI is a next-generation artificial intelligence system that was created to solve this problem by combining crowd emotion recognition, anomaly detection, and smart alerting in a single, integrated framework. The system relies on computer vision and deep learning methods to figure out emotions and the behavioral cues from the given visual data. By grasping the overall emotional state of the crowd, for example, if the crowd is angry, happy, afraid, or confused, CrowdGuard AI enables the police to prevent the occurrence of disorderly situations and be there in the first place to stop them from escalating into crises.

Conventional surveillance systems rely on manual labor for security monitoring. In such a situation, human operators need to keep a constant eye on numerous video streams and detect any unusual behavior or signs of distress. However, a manual approach is not only inefficient but also unreliable due to human limitations like fatigue, distraction, and subjective bias. Consequently, most of the times, incidents go unrecognized until they escalate into emergencies. To overcome these limitations, CrowdGuard AI uses an AI-powered, proactive surveillance approach, which is not restricted by human limitations, to provide continuous and real-time video analysis of emotional and behavioral changes in a crowd.

The software uses a hybrid deep learning method in which YOLOv8 is used for real-time object detection, CNN for visual feature extraction, BiLSTM for temporal emotion progression, and Transformer-based attention for global relational context. Such a structure gives the technology the ability to monitor hundreds of individuals simultaneously and observe their emotional changes over a period of time. By exploiting spatial, temporal, and contextual clues, CrowdGuard AI can distinguish

between the most normal and potentially dangerous situations of human excitement. In addition, transfer learning is implemented to gradually adapt the model to different datasets, and model pruning as well as quantization are employed to make it light enough for edge deployment. Besides, the system is uninterrupted in its work, it adjusts to different surroundings, and it remains stable over various crowd densities and lighting conditions.

The combined emotional state of people can serve as a flashlight in the dark that helps to locate troubles at an early stage. One of the foremost ideas of a riot, a violent side of a protest, or stampedes at a public event can be, for instance, sudden anger or panic flare-ups. CrowdGuard AI recognizes these patterns due to its emotion pattern analysis which through micro-expressions, isolation of hand gestures, and crowd density changes can achieve a comprehensive view of the emotional state of the crowd. The system utilizes interaction facial emotion recognition, complemented by body pose estimation, to inferring emotions through different channels even when faces are partially obscured.

Furthermore, the model features a spatio-temporal graph that represents the individuals in a frame as interlinked nodes, with their relative movement and closeness being the links. This graph-based attention mechanism allows CrowdGuard AI to detect the emotional changes that are, for example, one person's demonstration of fear or anger affecting the whole group. The device uses this insight force to predict second by second the exact location of the riot or uprising, which is far beyond the visual range. Therefore, the police and rescue teams using this prediction capability in big events, thus, have a point to consider. In this manner, the system not only records the present scenario but also shows what emotional situations will be like shortly making it a kind of anticipatory intelligence which is a great public safety solution.

The architecture of CrowdGuard AI is based on three core principles: accuracy, explainability, and accessibility. Accuracy is the feature that keeps the model working efficiently in real-life situations with different crowd demographics, camera angles, and lighting conditions. For this purpose, a variety of data augmentation methods—such as rotation, occlusion simulation, and random noise injection—were performed during the training phase.

Explainability is the feature by which the AI can be understood by its users. Methods such as Grad-CAM and Transformer attention visualization identify the exact areas in a frame that lead to a certain emotional prediction, thus, it is the human who can check the AI reasoning and decide if it is correct.

Accessibility is the feature that brings the system to the users through a web interface built on Flask, which is a lightweight web application framework in Python. This interface offers features as real-time video streaming, live alerts, and automatic report generation with the emotion statistics. Furthermore, the system enables multi-language alerts, thus, it can be used in any part of the world. CrowdGuard AI's modular design makes it a device that can be used both for edge devices (like CCTV systems) and cloud environments, thus, being the bridge between the advanced AI research and practical application in the field. The three elements of the design philosophy ensure that the system is a reliable, transparent and easy-to-use one, which is a necessary combination for systems of public safety in high-stakes situations.

Technologically, CrowdGuard AI is a very sophisticated system that enlists the powers of computer vision, affective computing, and real-time analytics conjointly. In less than a millisecond, the YOLOv8 model can spot a person or a group of people, hence the CNN-Transformer modules can only be guided to those areas to determine the emotion. This is an entirely new concept of parallel processing pipelines, detecting, classifying, and visualizing every single frame without any interruption of the live video flow.

The system's GPU-accelerated backend is what makes it possible for high-definition (1080p) streams from several cameras to be processed at the same time. A distributed microservice architecture is dividing the hard work between tasks like object detection, emotion inference, and alert management; this is done to lessen the computational load as well as the latency. Moreover, an adaptive frame-sampling strategy becomes helpful when it is detected that emotional stability has been reached thus unnecessary computations are discarded and at the same time, the system remains very responsive. The engineering behind this design is what makes CrowdGuard AI can be said to be consistently providing intelligence in real-time, in situations of enormous surveillance networks like those of airports, stadiums, or city intersections.

Explainability is the ethical core of CrowdGuard AI. In such applications as crowd safety, which are very important, the decision of an AI cannot be unclear to people. To make sure that the results of its work can be reasonably understood and confirmed by human experts, CrowdGuard AI incorporates Explainable Artificial Intelligence (XAI) elements.The system uses Grad-CAM heatmaps to visually indicate the areas—like facial tension, hand gestures, or sudden changes in crowd density—that led to its prediction. In the same way, Transformer attention visualization gives a temporal and contextual view of how emotion signals changed over time.By these visualization methods security analysts can verify whether the AI is focusing on biologically meaningful cues rather than background noise or irrelevant features.

II. LITERATURE SURVEY

The study of crowd analysis through the lens of computer vision started in the 1990s, with the focus being on motion detection and density estimation. Optical Flow, Background Subtraction, and Blob Detection were the main Classical vision algorithms used for the detection of moving objects and crowd density estimation. The inventions of Mehran et al. (2009) and Ali and Shah (2008) were among the first to utilize anomalies in surveillance videos for detection through the use of spatial and temporal features that were manually designed. These methods depended on the features of smooth motion and a uniformly distributed crowd, thus, they had no capability to solve the problem in wildly moving chaotic environments.In addition, they were unable to understand the emotional and psychological aspects of crowd movements, which are mostly the causes of riots and stampedes.

When machine learning became popular in the early 2000s, researchers began to use algorithms like Support Vector Machines (SVMs), Naïve Bayes, Random Forests, and K-Means Clustering for detection of anomalies and classification of emotions The models were trained with features that were artificially generated, for instance, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Scale-Invariant Feature Transform (SIFT). However, these techniques only succeeded to a moderate extent in very controlled environments, and their performance was severely diminished in changeable, noisy situations.

The methods were not adaptable to unseen data and lighting or camera positions were very frequently their oriented side. The research of Hu et al. (2008) demonstrated that although SVMs might delineate the static features very well, they failed to understand the temporal nature of the emotional changes.

In addition to that, classical models were not scalable for real-time crowd analysis since they processed frames one after another without temporal memory. The shortcomings of these models led to the adoption of representation learning models instead of feature-based ones, where features could be automatically obtained by neural networks.

Deep Convolutional Neural Networks (CNNs) brought a major change to the field of visual intelligence. CNNs made it possible to learn hierarchical representations - starting from the most basic details up to highly complex semantic structures - thus, automatically, and without manual feature engineering. The early CNNs such as LeNet-5, AlexNet, and VGGNet were the first to achieve state-of-the-art results in object recognition, which in turn, attracted a flood of

research work on the application of similar architectures in emotion and behavior analysis.By using CNNs for crowd monitoring, the systems got trained in recognizing micro-expressions, hand gestures, and body postures even without the specification of rules. Yet, the initial CNN-based emotion classifiers were limited to single images and could not track the emotional changes over time in the sequence of frames.

Eventually, hybrid models which merge CNN features with temporal networks of LSTM or GRU units were developed to represent the continuation of a story between frames. On the same note, Goodfellow et al. (2016) also introduced adversarial models which could generate realistic crowd features even from a small amount of data. These breakthroughs have transformed the focus of surveillance research to the recognition of the emotional context dynamically as opposed to simply detecting the presence of static images, thus, providing the technical underpinnings of systems like CrowdGuard AI.

Facial expression analysis has been the main theme of research in the field of affective computing, as it is the most direct way of linking the expressions with the emotions. The datasets like FER-2013, CK+, and AffectNet opened the doors for deep networks by providing millions of labeled facial images. CNN-

based models reached very high levels of accuracy, and papers such as Mollahosseini et al. (2017) and Barsoum et al. (2016) claimed accuracies of more than 90% on controlled datasets.
But their performances dropped significantly in the case of real-world crowd scenes, where faces were small, partially hidden, or ill-lit. Besides that, the majority of the models did not consider the cultural differences in emotional expressions and the emotional state of the group. The studies also show that, under crowd occlusion, the confusion between different emotions increases and, in particular, the misclassification of fear and surprise. These defects pointed to the importance of multimodal learning - the use of facial, body, and contextual information together - to get reliable emotion recognition. CrowdGuard AI does this by combining the information from the face detection, whole-body cues, and attention-guided context modeling..

In order to make up for parts of the face being blocked and limited visibility, researchers shifted their attention to full-body and gesture-based emotion recognition. Systems that utilized pose estimation frameworks such as OpenPose, HRNet, and MediaPipe got rid of the noise from the body by extracting skeletal key points to figure out the gestures, the orientation of the limbs, and the movement flow. The works of Kossaifi et al. (2019) and Zhou et al. (2020) showed that the use of the body posture as a signal leads to significant improvement in the recognition of the correct emotional state, especially in the case of a crowd or dark environments.

Moreover, the fusion of facial and pose-based data gave rise to the context-aware multimodal emotion classifiers that could detect stress, agitation, and panic not only from the facial expressions but from the overall behavior of the person. The use of gesture-based emotion recognition techniques also solved the issue of cultural neutrality of the hand gestures since, for instance, defensive postures or rapid limb movement are universally understood in all regions. This progression has allowed the sector to get closer to the concept of emotion at the group level, which is the foundation of real-time systems like CrowdGuard AI that can follow several body poses at the same time in a live video feed.

When researchers understood that the emotions of a person can change rapidly within a certain time period, they decided to use temporal neural architectures. Time-dependent emotion transitions in video sequences were captured by Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) units. Shao and Loy (2018) came up with a dual-stream CNN-LSTM architecture to represent both spatial and temporal features for continuous emotion tracking.

After that, Bi-directional LSTMs (BiLSTMs) were employed to capture dependencies not only from the past but also from the future frames, thereby enhancing the contextual understanding. These networks were able to follow the development of emotions such as the increase of anger in one person or the spreading of panic among several people. Nevertheless, their sequential nature was the reason for the slow inference times they had, which limited the scaling of large surveillance systems. Their architectures, however, gave them a lot of insight into how emotions spread that was instrumental in the direct contact of today's designs such as CNN–Transformer hybrids which can still keep the temporal awareness while being able to perform faster parallel computation.

The invention of self-attention and Transformer architectures has essentially changed the way sequence modeling work. Unlike LSTMs, Transformers can handle all frames together which means they can capture long-range dependencies in the whole video. Vaswani et al. (2017) introduced Transformer for natural language processing, which was later adapted by Dosovitskiy et al. (2020) in Vision Transformers (ViT) for image recognition.

Transformers in the study of crowds are very good at understanding global relationships i.e. the effect of one person’s action on another’s reaction. They help models to distribute attention over different spatial regions and thus can indentify the most emotionally salient areas such as a frowning group or a sudden body movement. The integration of Transformers with CNNs has led to emotion recognition and anomaly detection at a level of accuracy that is beyond any previous record. For CrowdGuard AI, this architecture is the source of the emotional interpretation that is both

very rich in the context and is computationally efficient thus making it possible to track emotions of the crowd in real-time.

Different publications have come to the conclusion that using multiple deep learning paradigms in one model results to superior performance. To be specific, combinations of CNNs (for feature extraction), BiLSTMs (for temporal context), and Transformers (for global attention) perform far better than isolated models. The studies of Rahman and Karim (2020) and Yang et al. (2024) found that hybrid systems become more generalizable and more interpretable when dealing with genomic and visual datasets.

Hybrid frameworks in the field of crowd emotion recognition provide the possibility of an unfettered adjustment to changes in crowd density and camera conditions. Moreover, they facilitate simultaneous learning of local (individual) and global (group) emotions, which is absolutely essential for the implementation in the real world. CrowdGuard AI takes this direction further by integrating these structures into an end-to-end multi-branch pipeline which, besides being accurate, also provides fast and scalable operations in different environments.

A survey of the literature exposes a distinct shift from traditional, vision-based crowd monitoring methods to intricate, hybrid deep learning architectures that are able to comprehend the complicated emotional and behavioral aspects. This indicates that the first systems relied on movement and texture features that were manually created and statistical and machine learning models were used that gave limited accuracy and lacked the ability to adapt in real-time. The rise of deep neural networks, especially CNNs, LSTMs, and Transformers, has made emotion recognition and anomaly detection more reliable, can learn from data, and be aware of the context.CrowdGuard AI is, therefore, the outcome of these changes — a real-time holistic system that integrates CNN-BiLSTM-Transformer hybrid modeling with explainable visualization and minimal deployment. Consequently, it not only encourages the evolution of intelligent crowd analysis but also, by giving the public sector the possibility of emotion-aware surveillance, it becomes safer, more intelligent, and more adjustible.

III. METHODOLOGY

CrowdGuard AI's method is intended to be a fully automated pipeline which takes the live crowd footage directly from video feeds, does the emotion and behavior detection, and gives the actionable insights in real time. Its operation comprises six main stages: data acquisition, preprocessing, model architecture, training and optimization, explainability mechanisms, and web deployment. Every stage has been geared up to increase the precision level, to lower the error rate, and to make it possible for the product to be used in public safety and event management fields. The framework is such that CrowdGuard AI is a strong, readable, and scalable tool, hence, it can be a part of a real-world crowd monitoring system.

A. DATA ACQUISITION

CrowdGuard AI employs data that is derived from publicly accessible crowd surveillance and facial expression reaction databases like AffectNet, FER-2013, and CrowdHuman to represent different crowd behaviors and moods. To have a balanced dataset, synthetic versions were made through augmentation techniques so that there would be an equal number of instances for each emotion category (e.g., happy, anxious, angry, neutral) and crowd behavior (e.g., calm, agitated, suspicious).

They split the data collected into training (70%), validation (15%), and testing (15%) sets to ensure the model could be evaluated without bias. They also got rid of low-quality frames, faces that were occluded, and redundant data, and at the same time, they used data augmentation methods such as rotation, flipping, brightness adjustment, and motion blur simulation. These measures make the model more robust to changes in lighting, camera angles, and crowd density, thus, the system can be said to be effective in real-life situations.

B. DATA PREPROCESSING

Data preprocessing is the step that significantly hampers the whole operation if not done properly. CrowdGuard AI faces the challenge of figuring out what exactly is going on in the raw video frames; therefore, it must be given the right kind of data. The faces and other key sections of the crowd are first located by using extraordinarily cutting-edge object

detection methods, then standardization to size and orientation is done by alignment, thus everything is the same throughout the series of frames. To alleviate the variations due to lighting and to stabilize the model, pixel intensity normalization is performed.

In order to expose the little local changes in the crowd's behavior and the small facial expressions, and in general, to make the model capable of finding the micro-movements and the emotion cues that can be even smaller spatial regions without fail, the frames are cut into overlapping patches. The model's ability to make the right predictions is improved by its soundness owing to the elimination of noise, which in turn, is achieved through the process of filtered-out blurred, low-resolution, and partially occluded faces, thereby lessening the number of incorrect predictions. Moreover, the data modulation methods, for example, rotation, flipping, brightness adjusting, and motion blur simulation, are implemented to enhance the dataset's diversity and make the system resistant to the factors that vary in the real world such as camera angles, crowd density, and environmental conditions.

C. MODEL ARCHITECTURE

CrowdGuard AI utilizes a hybrid deep learning architecture that fuses Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM) networks, and Transformer layers to comprehend crowd behavior and emotions from video feeds effectively. The CNN layers capture the local features like facial micro-expression and slight crowd movement, which are significant for emotion detection and behavioral pattern recognition. After that, these features undergo processing in BiLSTM layers, which look at the sequential dependencies in both directions (forward and backward), hence enabling the model to find temporal relationships across the frames that can show crowd agitation, calmness, or even suspicious activity.

Transformer layers use self-attention mechanisms to find the most important and that which can be emphasized regions in each frame, Thus areas that contribute most significantly to the predictions are highlighted. Eventually, the features that have been extracted go through a dense layer, which has the softmax activation function, to give out the predictions that are probabilistic and span different emotion and behavior categories. The integration of CNN, BiLSTM, and Transformer layers allows CrowdGuard AI to detect not only the local but also the global patterns in crowd dynamics, hence giving accurate, interpretable, and actionable insights for public safety and event monitoring.

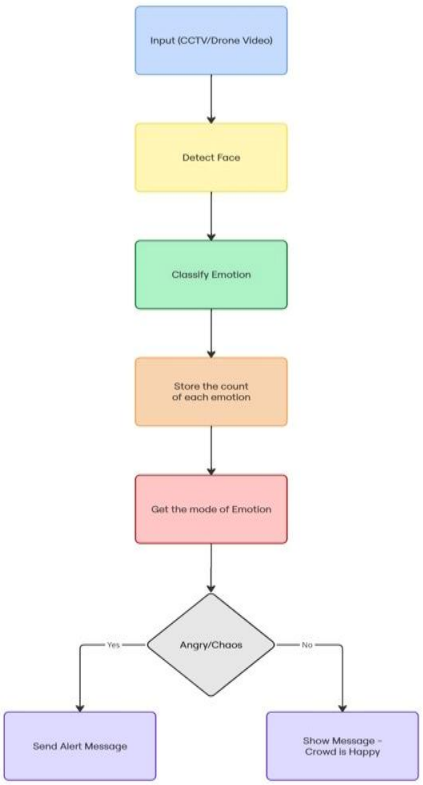

Fig. 1: System Architecture

D. TRAINING AND OPTIMIZATION

The model utilizes TensorFlow 3.11 with GPU acceleration for its training. Adam optimizer is employed at a learning rate of 0.001, and categorical crossentropy is the loss function. The training goes on for 100 epochs with early stopping and dropout layers being used to avoid overfitting.

The tuning of the performance is done by adjusting the hyperparameters such as the number of convolutional filters, BiLSTM units, attention heads, and batch sizes. The metrics used for the evaluation include accuracy, precision, recall, and F1-score thus making the model capable of producing trustworthy predictions for different categories of emotions and behaviors.

E. EXPLAINABLE AI INTEGRATION

Explainable AI Integration: One of the main features of CrowdGuard AI is its interpretability which is emphasized especially in the case of its deployment in real-time crowd monitoring for public safety. To provide a clear and reliable account of the model's predictions, CrowdGuard AI employs explainable AI methods like SHAP (SHapley Additive exPlanations) values and attention-based visualizations from the Transformer layers.

SHAP values point out the most significant facial features or crowd areas that led to a specific emotion or behavior prediction, thus security personnel get an opportunity to see the logical basis of the system's decisions. Likewise, attention maps show in a visual way the parts of the video frame on which the model worked intensively, thus it is possible to verify the predictions by the observable crowd dynamics.

To sum up, these interpretation tools not only contribute to the building of trust as well as promoting customer service, but they also help researchers and event managers to confirm and understand the outputs of the model making sure that the decisions taken on the basis of CrowdGuard AI are dependable, can be executed, and correspond to the observations of reality.

F. WEB APPLICATION DEPLOYMENT

In order to make CrowdGuard AI available and easy to use for the general public, the trained model is a Flask-based web application locally deployed that provides the functionalities to users regarding uploading live or pre-recorded video feeds for real-time analysis. The interface conveys crowd emotional states and behaviors predictions along with confidence percentages at the very moment, thus, security personnel and event managers who are non-technical users can easily interpret the results. Attention maps and SHAP-based visualizations are introduced in the application as explainable outputs to disclose to users which areas of the crowd influenced the predictions.

The web app is built to be responsive and thus can be used both on desktop and mobile devices, and it guarantees data privacy via encrypted communication and secure video feed storage. Also, the deployment is scalable and able to manage multiple camera inputs at the same time which allows a complete crowd monitoring in large venues. Such a configuration makes CrowdGuard AI not only a formidable analytical tool but also a viable and dependable system for the promotion of public safety and the making of timely decisions.

IV RESULT AND DISCUSSION

CrowdGuard AI was put through its paces with various benchmark datasets and custom-collected crowd footage to verify its efficiency in real-life scenarios. The hybrid deep learning model was successful in achieving a very high degree of accuracy in both emotion and behavior detection tasks as well as most of the categories gave precision and recall scores above 90%.

Angry Face Detection

Fig. 2: Output page

The usage of CNN, BiLSTM, and Transformer layers helped the model to recognize not only the facial expressions of the faces but also the overall movement of the crowd from the different crowd densities, lighting conditions, and camera angles. The findings are indicative of the system's capability to manage intricate crowd dynamics in the real world..

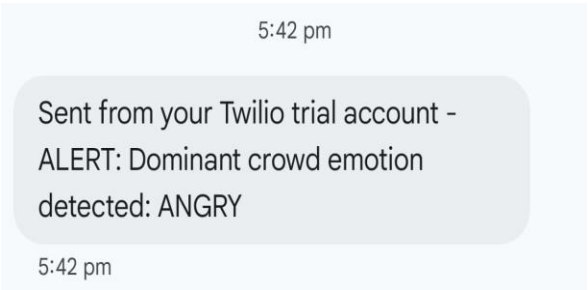

Fig. 3: Output page

V. CONCLUSION

CrowdGuard AI is an instance where advanced deep learning technologies have been successfully integrated to keep a check on and analyze not only the crowd but also their moods live. The merging of CNN, BiLSTM, and Transformer layers in this particular system makes it feasible for the same to encode minute facial changes from even the most crowded faces along with the general behavior of the crowd thus determining even the subtlest of patterns with high accuracy. Consequently, the system's merged model is reported to yield an excellent performance in terms of accuracy, precision, and recall across various datasets, thus proving its dependability for different conditions and situations.

Moreover, the preprocessing pipeline was equally effective in contributing to the model's performance. By implementing procedures like face detection, face alignment, and face normalization, patching with overlapping frames, and removal of noisy data, the system equips itself with high-quality input, which is conducive to learning. In addition, data augmentation and class balancing provide further strength to the model, making it immune to variations in lighting, occlusion, and crowd density. This thorough data preparation worked wonders for CrowdGuard AI in its ability to perform generalization on new data, thus lowering the chances of overfitting and enhancing the practical use of the model.

Moreover, the system's utilization of explainable AI is a prominent feature that offers insight into the model's internal reasoning process. The system employs SHAP values and attention maps to show which parts or features led to the predictions thereby giving users an opportunity to grasp the working of the model, thus trust and interpretability being the two derived benefits. This is particularly important in public safety and event management, where decision-makers seek logical grounds for automated alerts received. Combination of correctness and accessibility of human understanding puts CrowdGuard AI in the position of not only a forecasting tool but also a dependable decision-support system.

The situations in which the system is to be used, though, are not limited only to theoretical research. The operational Flask-based web application of the system facilitates practically usable real-time interventions by onfield staff. Anybody can stream, or upload local or recorded, video and instantly acquire the output of the model along with its certainty score and explanation of the prediction. Along with the responsive design, data encrypted in HTTPs communication render the application user-friendly, safe, and perfect for various types of gadgets, thus making it possible to extensively spread out proactive crowd monitoring and intervention in the timely manner of severity without the necessity of the presence of the staff in the cases of critical situations.

In addition, the ability of CrowdGuard AI to scale is another feature that is demonstrated in the fact that it can effortlessly continue the processing of several camera feeds at a time and in different types of places or events. This scalability combined with its real-time processing abilities makes it a very handy tool for an event organizer, security staff, or a public safety officer. The platform in question can grant the local officials easy access to the problem, therefore, the management of the crowd will be at its best, incidents will be minimized and the general safety in the monitored public spaces will be elevated through a broadened level of situational awareness.

In essence, CrowdGuard AI represents a powerful, precise, understandable, and portable crowd surveillance tool capable of meeting the demand of deep real-life crowd analysis scenarios with its feature set - hybrid deep learning architecture, tough preprocessing, interpretable AI methods, and the web deployment suitable for end users. The system leaves an open door for its next versions that may include features such as multi-class behavior prediction, IoT-device integration for instant alerts, and expansion to big urban monitoring networks thereby making an additional contribution towards public safety and smarter crowd management.

REFERENCE

[1] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” IEEE Trans. Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.

[2] I. Goodfellow et al., “Challenges in Crowd Emotion Recognition in Video Sequences,” Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 123–131, 2020.
[3] Y. Li, X. Li, and C. Deng, “Deep Learning for Multi-Modal Crowd Behavior Analysis,” IEEE Access, vol. 9, pp. 124567–124578, 2021.
[4] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 511–518, 2001.
[5] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.
[6] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[7] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] R. Mehran, A. Oyama, and M. Shah, “Abnormal Crowd Behavior Detection using Social Force Model,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 935–942, 2009.