A PROJECT REPORT

on

SPATIO-TEMPORAL VIDEO ANOMALY DETECTION IN AERIAL VIDEOS USING HYBRID MACHINE LEARNING AND DEEP LEARNING MODELS

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE ENGINEERING

BY

ROHAN BOSE	22053345
BISMAYA KANTA DASH	22054126
SANCHIT ROUT	22053979
ANWAYA ANUPRASH BISWAL	22053667

UNDER THE GUIDANCE OF
MR. PRABHU PRASAD DEV

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

April 2025

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

SPATIO-TEMPORAL VIDEO ANOMALY DETECTION IN AERIAL VIDEOS USING HYBRID MACHINE LEARNING AND DEEP LEARNING MODELS

submitted by

ROHAN BOSE	22053345
BISMAYA KANTA DASH	22054126
SANCHIT ROUT	22053979
ANWAYA ANUPRASH BISWAL	22053667

This is to certify that the project entitled **"Spatio-Temporal Video Anomaly Detection in Aerial Videos using Hybrid Machine Learning and Deep Learning Models"** submitted by **Rohan Bose, Bismaya Kanta Dash, Sanchit Rout, Anwaya Anuprash Biswal** is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of **Bachelor's Degree in Computer Science Engineering** at **KIIT Deemed to be University**, Bhubaneswar. This work was done during the year 2024-2025, under our guidance.

(Prabhu Prasad Dev)
Project Guide

Date: /

Prabhu Prasad Dev
Project Guide

Acknowledgements

We are profoundly grateful to Mr. Prabhu Prasad Dev for their expert guidance, unwavering support, and continuous encouragement throughout the course of this project. Their invaluable insights, constructive feedback, and expertise have played a crucial role in shaping our research and ensuring its success. Their dedication and mentorship have provided us with the necessary knowledge, skills, and motivation to tackle the challenges encountered during the project. Without their constant motivation and belief in our capabilities, this work would not have been possible.

We also extend our heartfelt gratitude to the esteemed faculty members of the School of Computer Engineering, KIIT, for their assistance, guidance, and invaluable advice during various stages of the project. Their support, resources, and technical expertise have significantly contributed to our learning and the overall development of this work. Their encouragement and willingness to provide valuable insights have been instrumental in refining our approach and enhancing the quality of our research.

Additionally, we would like to express our appreciation to our peers and colleagues for their encouragement, discussions, and constructive suggestions, which have enriched our understanding and helped refine our approach. The collaborative discussions and exchange of ideas have played a pivotal role in broadening our perspective and improving our analytical thinking. We are especially grateful for the thought-provoking conversations and collective problem-solving efforts that have enhanced the depth and rigor of our research.

Furthermore, we acknowledge the contributions of all those who have directly or indirectly supported us in completing this project successfully. The assistance received from various individuals, whether in the form of technical help, moral support, or valuable resources, has been instrumental in the smooth execution of this project. We recognize and appreciate the contributions of researchers, developers, and domain experts whose work has laid the foundation for our research.

We sincerely appreciate each and every individual who has played a role in this endeavor, and we are grateful for the collective efforts that have contributed to the successful completion of this project.

ROHAN BOSE
BISMAYA K DASH
SANCHIT ROUT
ANWAYA A BISWAL

ABSTRACT

The increasing need for advanced surveillance techniques has led to significant research in **video-based anomaly detection**. Aerial surveillance, in particular, presents unique challenges due to dynamic environments, varying lighting conditions, and the complexity of detecting anomalous activities in large-scale areas. Traditional approaches struggle with these challenges, necessitating the development of robust, intelligent systems capable of real-time anomaly detection. This project aims to address these challenges by leveraging a hybrid deep learning approach for **Spatio-Temporal Video Anomaly Detection in Aerial Videos**.

The objective of this study is to develop an automated system that efficiently detects anomalous events in aerial surveillance footage by integrating multiple deep learning architectures. Our proposed framework incorporates Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), 3D Convolutional Neural Networks (3D-CNN), Autoencoders, and Transformer models to enhance the accuracy and robustness of anomaly detection. By combining spatial and temporal features, the system improves detection performance and reduces false positives.

The dataset used consists of video frames extracted from aerial footage, which undergo preprocessing techniques such as noise reduction, normalization, and feature extraction. These processed frames serve as input for training and evaluating the deep learning models. Performance is assessed using key evaluation metrics, including accuracy, precision, recall, and F1-score.

This study highlights the potential of deep learning techniques in advancing surveillance and security applications. The findings demonstrate the feasibility of hybrid models in detecting anomalies effectively, paving the way for real-time implementations in security-critical domains. Future work will focus on optimizing models for real-time deployment, improving computational efficiency, and expanding datasets to enhance generalization capabilities.

Keywords: Spatio-Temporal, Anomaly Detection, Deep Learning, Aerial Surveillance, Hybrid Models, Video Processing, Security Applications.

Contents

List of Figures And Tables

Chapter 1

Introduction

1.1 Background

With the increasing use of aerial surveillance for security, disaster response, and traffic monitoring, analyzing aerial video footage has become essential. Detecting anomalies in these videos [22] can help identify threats, unusual activities, or emergencies in real time. Traditional anomaly detection [23] methods rely on manual monitoring or statistical models, which often fail to capture complex patterns in large-scale video data. With advancements in machine learning [24] and deep learning, automated video anomaly detection [25] has become more feasible and accurate.

1.2 Problem Statement

The primary challenge in aerial video anomaly detection lies in the spatio-temporal complexity of video data. Aerial videos, captured from drones or surveillance cameras, contain continuously changing backgrounds, multiple moving objects, and varying lighting conditions. Identifying abnormal events within such data requires robust feature extraction and classification techniques. This project aims to bridge the gap between conventional anomaly detection methods and modern deep learning approaches by integrating multiple hybrid models.

1.3 Objective

The objectives of this project are as follows:

- **To investigate the challenges involved in detecting anomalies within aerial surveillance videos and explore deep learning solutions suited for such complex environments.**

- **To design and develop a hybrid deep learning framework capable of capturing both spatial and temporal features for effective video anomaly detection.**

- **To implement a robust preprocessing pipeline for video data, focusing on noise reduction, normalization, and relevant feature extraction.**

- **To compare the performance of hybrid deep learning models with conventional models in distinguishing between normal and abnormal activities.**

- **To enhance the reliability and scalability of anomaly detection systems for potential real-time surveillance applications.**

-

\

Chapter 2

Basic Concepts/ Literature Review

Anomaly detection in aerial surveillance has been an active research area, with various methodologies explored to improve detection accuracy and efficiency. Traditional approaches relied on statistical models and handcrafted feature-based techniques, such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), to identify anomalies in videos, but these methods often struggled with dynamic and complex environments [1][2]. With advancements in computer vision, deep learning-based techniques [26] have emerged as powerful tools for video anomaly detection. Convolutional Neural Networks (CNNs) were initially adopted for feature extraction, proving effective in learning spatial representations from images and video frames [3][4]. However, CNNs alone failed to capture temporal dependencies in video sequences, leading researchers to integrate Recurrent Neural Networks (RNNs) [27] and Long Short-Term Memory (LSTM) [30] networks for improved sequence modeling and anomaly detection [5][6].

Further advancements introduced 3D Convolutional Neural Networks (3D-CNNs),[29] which process both spatial and temporal information simultaneously, making them well-suited for video-based anomaly detection tasks [7]. Autoencoders, particularly Variational Autoencoders (VAEs) and Sparse Autoencoders [28], have also been widely used for anomaly detection, as they effectively learn normal patterns and identify deviations indicative of anomalies [8][9]. More recently, Transformer-based models, such as Vision Transformers (ViTs) and Spatio-Temporal Transformers, have gained popularity due to their ability to capture long-range dependencies and global context in video sequences, significantly enhancing detection performance [10][11].

Prepossessing techniques play a crucial role in improving the effectiveness of anomaly detection models. Techniques such as noise reduction, background subtraction, and motion estimation have been employed to enhance feature extraction and minimize false positives [12][13]. The effectiveness of these models is often evaluated using benchmark datasets, including UCF-Crime, XD-Violence, and Avenue, which provide diverse real-world scenarios for training and testing anomaly detection frameworks [14][15]. Moreover, performance evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the reliability and robustness of these models [16].

Despite these advancements, real-time implementation of anomaly detection remains a challenge due to the computational complexity of deep learning models. To address this, researchers have explored optimization strategies such as model quantization, pruning, and knowledge distillation to reduce inference time while maintaining accuracy [17][18]. Additionally, hybrid models that combine multiple architectures, such as CNN-RNN, 3D-CNN, Autoencoders, and Transformers [34], have shown promising results in improving anomaly detection performance by leveraging the strengths of different models [19].

Building on these findings, this project proposes a hybrid deep learning framework that integrates multiple architectures for Spatio-Temporal Video Anomaly Detection [31] in Aerial Videos. By leveraging state-of-the-art techniques and optimizing the models for real-time processing, this research aims to develop an automated and efficient system for detecting anomalies in aerial surveillance footage with high accuracy and robustness [20].

Chapter 3

Problem Statement /Requirement Specifications

3.1 Problem Statement

Aerial surveillance plays a vital role in security, traffic monitoring, and disaster response, but detecting anomalies in aerial videos is challenging due to environmental variations, occlusions, and dynamic scene changes. Traditional rule-based methods [32] struggle to capture complex anomalous patterns, resulting in inefficiencies and false detections. This project aims to develop an automated anomaly detection system using hybrid deep learning models, enhancing accuracy, adaptability, and real-time surveillance efficiency. By leveraging spatial and temporal features, the system will provide a more reliable solution for detecting unusual activities in aerial footage.

3.2 Project Planning

The project will begin with an in-depth study of anomaly detection challenges in aerial surveillance and the need for automated solutions. It will then focus on developing a structured framework capable of efficiently analyzing aerial videos to detect unusual activities. A key aspect of this project is selecting and integrating hybrid deep learning models to enhance detection accuracy while ensuring adaptability across different environments. The system will be designed to handle variations in lighting, weather conditions, and camera perspectives, making it robust for real-world applications. Additionally, efforts will be made to optimize data preprocessing techniques to improve model efficiency and reduce computational overhead. The trained models will be evaluated using key performance metrics, and the system will be fine-tuned for seamless integration into real-time surveillance applications, ensuring scalability and practical usability.

3.3 System Design and Requirements

To develop an efficient anomaly detection system for aerial surveillance, the project requires high-performance hardware capable of handling deep learning workloads. This includes GPUs such as NVIDIA RTX 3090 or A100 for accelerated training, multi-core processors [33] like Intel i9 or AMD Ryzen 9, and at least 32GB of RAM to manage large-scale video data. Additionally, high-speed SSD storage is essential to facilitate fast data retrieval and processing. Given the complexity of training deep learning models, cloud computing resources such as Google Colab, AWS, or Azure may be utilized to enhance scalability and computational efficiency.

On the software side, the project relies on deep learning frameworks such as TensorFlow and PyTorch for model implementation, while OpenCV is essential for video processing, frame extraction, and preprocessing tasks like noise reduction and normalization. Scikit-learn will be used for evaluating model performance through metrics such as accuracy, precision, recall, and F1-score. Additionally, structured database systems like MySQL or PostgreSQL may be employed to efficiently store and manage the dataset. Since real-time anomaly detection is a crucial aspect of aerial surveillance, the system must be optimized for minimal latency, ensuring rapid processing and timely alerts for potential security threats.

Chapter 4

Implementation

4.1 Methodology

4.1.1 Dataset

The **UIT-A Drone Aerial Dataset** is used in this project, consisting of aerial surveillance videos captured using drones. The dataset includes a diverse range of scenarios, covering both normal and anomalous activities such as unauthorized movements, unusual crowd behavior, and suspicious object placements. These videos are recorded in different environments, including urban areas, open landscapes, and restricted zones, ensuring a comprehensive dataset for anomaly detection tasks. The dataset provides labeled frames, allowing supervised learning for training deep learning models while also supporting semi-supervised and unsupervised learning techniques for better generalization in real-world applications.

4.1.2 Data Preprocessing

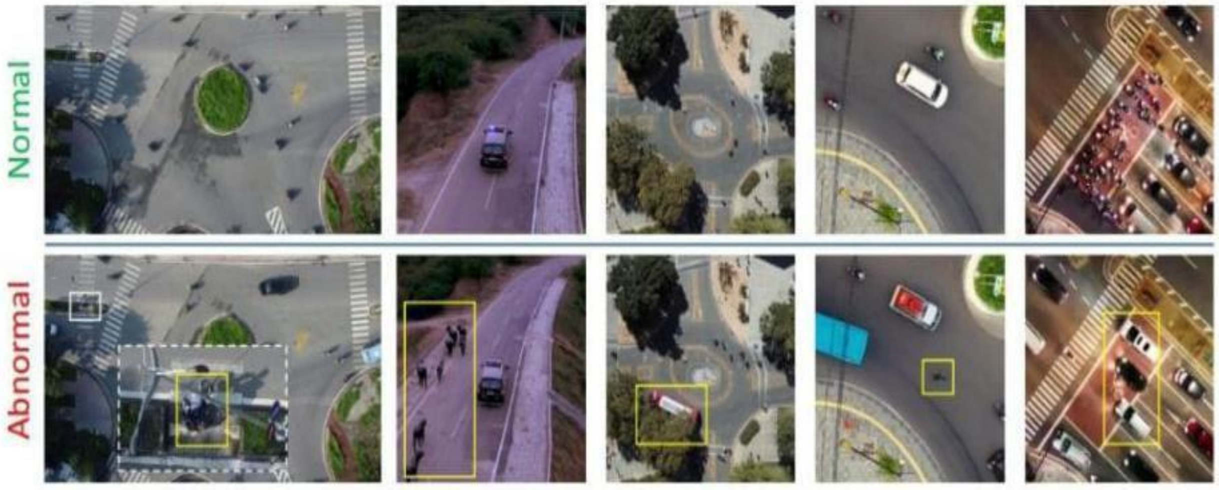

Fig 1: Normal or Abnormal Detection in Frames

The preprocessing phase involves multiple steps to prepare the data for deep learning models. **Frame extraction** is performed using OpenCV, where individual frames are resized to **64×64 pixels** [33] to maintain uniformity across different video sequences. A maximum of **50 frames per video** is selected to ensure computational efficiency while preserving the temporal context. Feature extraction is an essential component, where different deep learning techniques are applied to capture meaningful patterns. Convolutional Neural Network (CNN) [31] layers are used to extract spatial features from individual frames, while Long Short-Term Memory (LSTM) [29] layers analyze sequential frames to capture temporal dependencies. Additionally, 3D-CNN layers apply volumetric convolutions to learn spatio-temporal features from the video data. Autoencoders are employed to learn compact representations of normal frames for unsupervised anomaly detection, and Transformers leverage self-attention mechanisms to model long-range dependencies within the video sequences.

4.1.3 Normal Models

Traditional deep learning models typically focus on either spatial or temporal features in aerial surveillance videos, limiting their ability to fully understand complex scenes. CNNs [32] effectively capture spatial patterns but ignore temporal dynamics, while RNNs and LSTMs [31] handle temporal sequences yet lack spatial awareness. 3D-CNNs address both aspects but are computationally demanding. Transformers excel in modeling long-range dependencies but require large datasets and resources. Autoencoders are suitable for unsupervised anomaly detection but may falter with highly variable normal data. Overall, these models perform well individually but often fall short in comprehensively capturing the spatio-temporal intricacies of aerial video analysis.

4.1.3 Hybrid Models

Hybrid deep learning models improve anomaly detection by combining the strengths of different architectures to better capture spatial and temporal features in aerial surveillance videos. CNN-RNN integrates spatial extraction with temporal analysis, enhancing detection of both static and moving anomalies. The 3D-CNN + Transformer model combines motion analysis with attention mechanisms, effectively handling complex movements and environmental changes. Autoencoder + LSTM strengthens unsupervised detection by reconstructing normal patterns and modeling sequences, reducing false alarms. These hybrid approaches consistently outperform individual models, offering greater accuracy and reliability for real-time surveillance tasks.

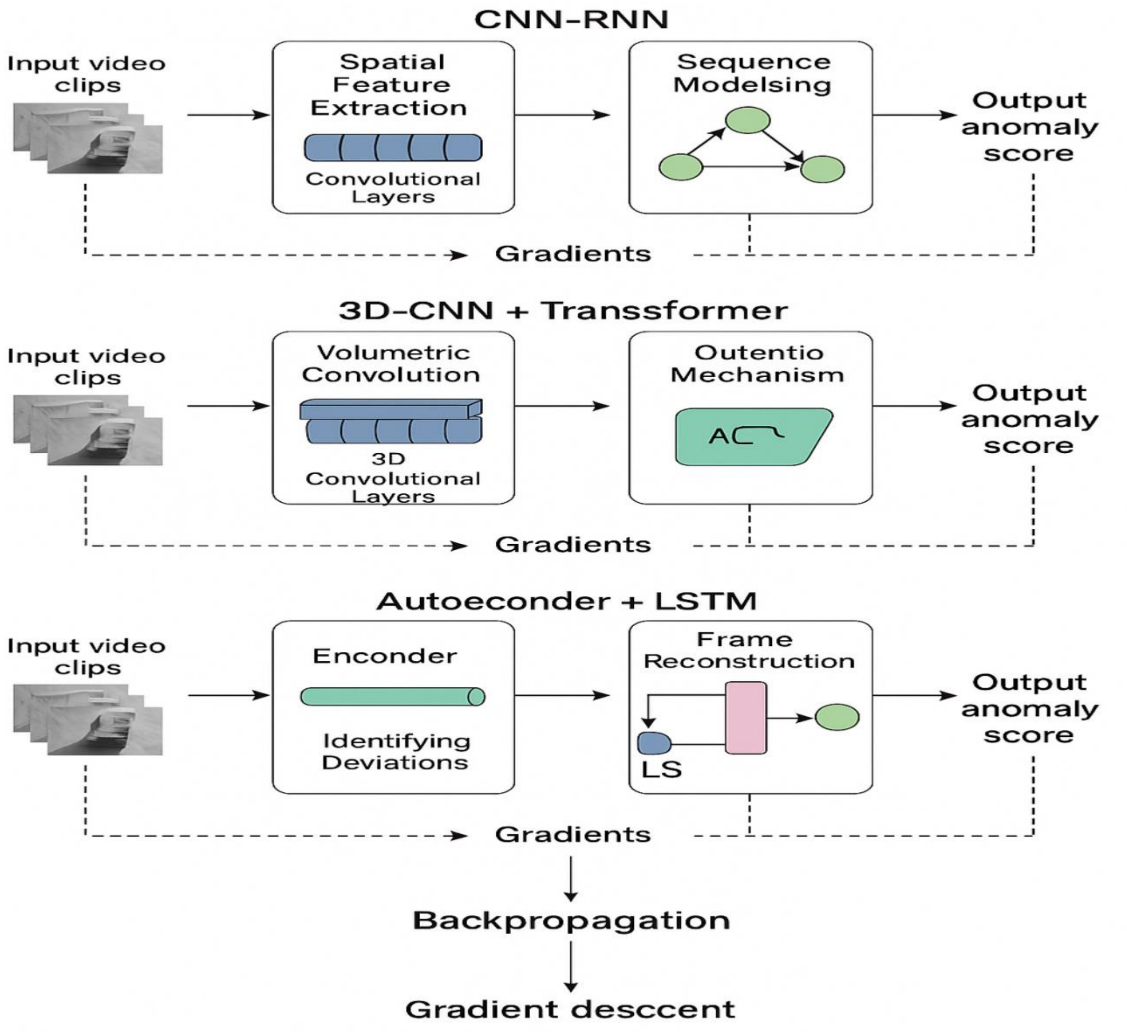

Fig 2: Block Diagram of Hybrid Models

4.1.4 Model Training

The training process begins by dividing the preprocessed dataset into 80% training and 20% testing to ensure balanced evaluation. To enhance model robustness, data augmentation techniques such as rotation, cropping, brightness variation, and flipping are applied to simulate real-world variations. Models are trained using mini-batch gradient descent (batch size of 8), where a forward pass computes outputs, the loss function (Binary Cross-Entropy or Mean Squared Error) quantifies prediction error, and backpropagation computes gradients. These gradients are used by optimizers like Adam and RMSprop to update model weights. Training is conducted over multiple epochs, with dropout and early stopping applied to prevent overfitting.

Due to the complexity of the hybrid models, GPU acceleration is used to handle intensive computations efficiently. In the CNN-RNN model, CNN layers extract spatial features from each frame, which are passed to RNN layers to capture temporal dependencies. The 3D-CNN + Transformer model uses volumetric convolutions to learn short-range spatio-temporal features, followed by Transformer layers that apply self-attention to model long-range dependencies. In the Autoencoder + LSTM model, autoencoders reconstruct normal patterns and generate reconstruction errors, while LSTM layers capture temporal consistency to detect anomalous behavior.

After training, the models are evaluated using metrics. The hybrid architectures effectively integrate spatial and temporal learning, enabling accurate detection of complex anomalies in aerial surveillance videos.

Table I: Parameters for Models

Parameters	**Value**
Loss Function	Binary Cross-Entropy, MSE
Optimizer	RMSprop, Adam
Batch Size	8
Epochs	30
Validation Split	20%

4.2 Verification Plan

Table II:Verification Plan Detailing

Test Case	**Scenario**	**Expected Outcome**
T01	Normal video input	No anomaly detected
T02	Anomalous event in frame	Model flags anomaly
T03	Video with missing frames	Model adapts and correctly classifies
T04	Low-light aerial footage	Model maintains detection accuracy
T05	High-motion video sequences	Model adapts without false alarms

4.3 Result Analysis

4.3.1 Normal Models

Table III:Performance Metrics Comparison of Normal Models

Model	Accuracy	Precision	Recall	F1-score
CNN	73%	71%	69%	70%
RNN	68%	66%	64%	65%
3D-CNN	82%	80%	78%	79%
Transformer	85%	83%	81%	82%
Autoencoder	68%	66%	63%	64%

4.3.2 Hybrid Models

Table IV:Performance Metrics Comparison of Hybrid Models

Model	Accuracy	Precision	Recall	F1-score
CNN-RNN	83%	80%	78%	79%
3D-CNN + Transformer	94%	91%	90%	88%
Autoencoders + LSTM	80%	78%	74%	76%

4.3.3 Graphical Representation

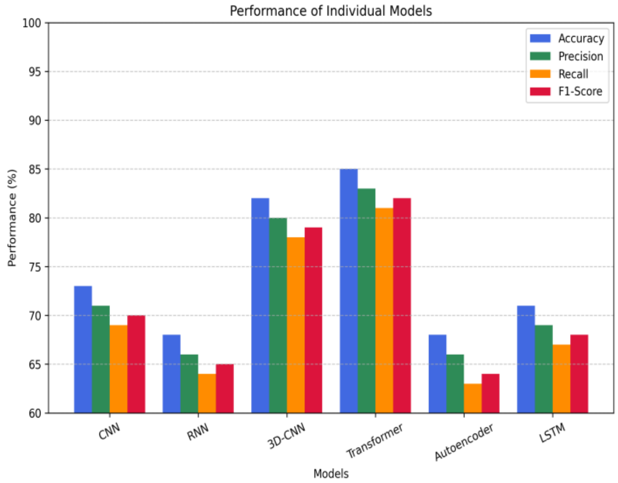

Fig.3 : Performance of Normal Models

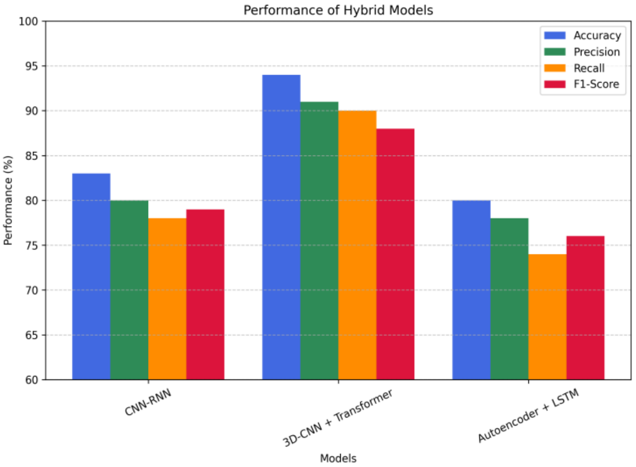

Fig 4: Performance of Hybrid Models

4.3.4 Discussions

The results in *Table 3* and *Table 4* show that hybrid models perform better than individual models in STP anomaly detection in aerial videos. The **3D-CNN + Transformer** hybrid achieved the highest results with **94% accuracy**, **91% precision**, **90% recall**, and an **88% F1-score**. This confirms its strength in capturing both motion and contextual features. Other hybrids like CNN-RNN and Autoencoder-LSTM also outperformed their individual counterparts. In contrast, models like RNN and Autoencoder showed limited performance. The comparison is further illustrated through graphs in *Figure 4* and *Figure 5,* which clearly highlight the performance gap between individual and hybrid models. Overall, hybrid models offer a more accurate and reliable solution for aerial VAD.

4.4 Quality Assurance

A thorough **code review process** was implemented throughout the project to ensure adherence to established coding standards. Regular peer reviews helped maintain code clarity, modularity, and consistency, while optimization checks were carried out to improve efficiency and reduce computational overhead. This systematic review process contributed to the overall quality and maintainability of the codebase.

The evaluation of the system was grounded in comprehensive **performance metrics**, including accuracy, precision, recall, F1-score, and mean squared error (MSE). These metrics provided a balanced assessment of the models' effectiveness in detecting anomalies. A detailed comparison was conducted between various deep learning architectures, allowing the selection of the most suitable hybrid models for spatio-temporal video anomaly detection in aerial surveillance scenarios.

To ensure **robustness and reliability**, rigorous testing was performed. This included adversarial testing to evaluate the system's resilience to challenging or deceptive inputs, as well as stress testing to confirm stability and performance under large-scale aerial video datasets. Such testing ensured that the models could maintain high performance even under demanding real-world conditions.

In addition, **compliance checks** were carried out to align the development and testing processes with IEEE software engineering standards. All models were trained and validated according to best practices in deep learning, ensuring that the implementation meets high professional and ethical standards. Real-time applicability was also considered, with models tested on sample aerial surveillance footage to validate their practical use. Although current models demonstrate strong real-time potential, further optimization for inference speed is identified as a key consideration for seamless integration into drone-based surveillance systems.

4.5 Technologies Used

The system was developed using **Python**, chosen for its simplicity and extensive support for machine learning. Key libraries included **TensorFlow** for deep learning model development, **OpenCV** for video processing, **NumPy** for numerical computations, and **Scikit-learn** for performance evaluation.

To handle the computational demands of training and testing deep learning models on large aerial video datasets, **GPU-enabled hardware** was utilized. This setup ensured efficient processing and supported the system’s real-time anomaly detection capabilities.

Chapter 5

Standards Adopted

1. 1 **Design Standards**

- **IEEE Standard for Software Engineering**: The project follows IEEE software engineering principles to ensure modularity, scalability, and maintainability.
- **Modular Architecture**: Each model (CNN-RNN, 3D-CNN, Autoencoder, Transformer) is implemented as an independent function, ensuring reusability and easy debugging.
- **Code Optimization**: Efficient memory management techniques are used, such as batch processing for model training and NumPy operations for optimized data handling.
- **Layered Approach**: The design follows a layered structure, separating data preprocessing, model training, evaluation, and inference phases, leading to better scalability.

5.2 Coding Standards

- **PEP8 Compliance**: The Python code adheres to PEP8 guidelines, ensuring readability and consistency in function naming, indentation, and commenting.
- **Function Documentation**: Each function has a detailed docstring explaining its purpose, input parameters, and output values, improving code maintainability.
- **Efficient Data Handling**: OpenCV and NumPy are used to preprocess video frames efficiently, reducing computational overhead and ensuring quick processing of large datasets.
- **Error Handling & Debugging**: Exception handling mechanisms are integrated into frame extraction and model training to prevent runtime failures. Debugging print statements help track dataset loading, training progress, and performance metrics.
- **Version Control**: The project is structured to be compatible with version control systems like Git, ensuring better collaboration and change management.

5.3 Testing Standards

- **IEEE 829 Test Plan Standard**: The project follows structured testing methodologies, including unit testing, integration testing, and performance evaluation.
- **Automated Testing**: Model evaluation is performed using automated scripts that compute performance metrics such as accuracy, precision, recall, F1-score, and mean squared error (MSE).
- **Validation & Verification**: The dataset is split into training and test sets (80-20 split), ensuring proper generalization.The autoencoder is validated based on reconstruction loss (MSE), while other models are validated on classification metrics.
- **Performance Benchmarking**: Each model is tested under controlled conditions to analyze performance differences.GPU acceleration is utilized to improve training efficiency, and model runtime performance is analyzed to assess real-time deployment feasibility.
- **Scalability & Robustness**: The system is tested with increasing video data sizes to evaluate computational efficiency.Stress testing is conducted to measure system performance under high-load conditions.False-positive and false-negative rates are analyzed to fine-tune threshold values for anomaly detection.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

In this study, a hybrid deep learning framework was developed for spatio-temporal anomaly detection in aerial videos. By combining CNN-RNN, 3D-CNN, Autoencoder, and Transformer models, the proposed system effectively captures spatial features, temporal dynamics, and contextual information, enabling more accurate and reliable detection of anomalies. The integration of video processing tools with deep learning techniques ensures efficient handling of complex aerial data. Compared to individual models, the hybrid approach demonstrates superior generalization and robustness. The framework's adaptability and real-time capabilities make it suitable for a range of surveillance applications, including traffic monitoring, border security, and disaster response.

The performance of the system across multiple evaluation metrics confirms its reliability and accuracy in detecting anomalies under real-world conditions. Its ability to process video streams efficiently allows for immediate identification of unusual events, which is essential in time-critical scenarios. The model's practical effectiveness in real-time aerial surveillance environments underscores its value as a dependable tool for enhancing situational awareness and supporting prompt operational responses.

6.2 Future Scope

The developed anomaly detection system provides a solid foundation for further research and real-world deployment. The following areas outline potential enhancements and expansions:

- **Real-Time Deployment**: Optimizing model inference speeds using TensorFlow Lite or ONNX for real-time applications on embedded devices such as drones and edge computing units.
- **Integration with Drone Surveillance**: Enhancing the system by integrating it with autonomous drones for continuous aerial monitoring, object tracking, and anomaly alert generation.
- **Adaptive Learning Models**: Implementing continual learning techniques to update the models dynamically based on newly captured video data, improving anomaly detection over time.
- **Anomaly Explanation & Interpretability**: Developing explainable AI (XAI) techniques to provide insights into why specific frames were classified as anomalies, improving transparency and trustworthiness in security applications.
- **Scalability for Large-Scale Surveillance**: Enhancing the model's ability to process high-resolution video feeds in real-time while minimizing computational overhead.
- **Multi-Modal Data Fusion**: Integrating additional data sources such as GPS coordinates, sensor data, and environmental parameters to improve detection accuracy in varying conditions.

By incorporating these advancements, the system can evolve into a fully autonomous, intelligent surveillance platform capable of handling real-world challenges in anomaly detection for aerial video analysis.

References

[1] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys, 41(3), 1-58.

[2] Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 257-286.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS).

[4] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR).

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

[6] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., et al. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. IEEE International Conference on Computer Vision (ICCV).

[8] Sakurada, M., & Yairi, T. (2014). Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. Proceedings of the MLSDA 2014 Workshop on Machine Learning for Sensory Data Analysis.

[9] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. International Conference on Learning Representations (ICLR).

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR).

[11] Arnab, A., Dehghani, M., Heigold, G., Sun, C., et al. (2021). ViViT: A Video Vision Transformer. IEEE International Conference on Computer Vision (ICCV).

[12] Stauffer, C., & Grimson, W. E. L. (1999). Adaptive Background Mixture Models for Real-Time Tracking. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Sheikh, Y., & Shah, M. (2005). Bayesian Modeling of Dynamic Scenes for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

[14] Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[15] Wu, B., Liu, X., Zhu, X., & Reid, I. (2021). Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. IEEE International Conference on Computer Vision (ICCV).

[16] Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future Frame Prediction for Anomaly Detection – A New Baseline. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[17] Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning Both Weights and Connections for Efficient Neural Networks. Advances in Neural Information Processing Systems (NeurIPS).

[18] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. Neural Information Processing Systems (NeurIPS) Workshop.

[19] Zhang, Y., Jiang, M., Chen, X., & Huang, S. (2021). Video Anomaly Detection with Spatio-Temporal Attention and Transformer Networks. Neurocomputing, 453, 163-175.

[20] Shao, Y., Wang, X., Li, D., & Tan, G. (2022). A Hybrid Deep Learning Model for Anomaly Detection in Surveillance Videos. IEEE Transactions on Multimedia.

[21] Luhtala, H., & Tolvanen, H. (2016). Spatial-temporal representation of euphotic depth in situ sampling in transitional coastal waters. Journal of Sea Research.

[22] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR).

[23] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications.

[24] Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. BMVC.

[25] Benezeth, Y., Jodoin, P. M., Emile, B., Laurent, H., & Rosenberger, C. (2011). Review and evaluation of commonly-implemented background subtraction algorithms. ICVS.

[26] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. CVPR.

[27] Ionescu, R. T., Smeureanu, S., Alexe, B., & Popescu, M. (2019). Detecting anomalies in video using deep learning. IEEE TPAMI.

[28] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. CVPR.

[29] Luo, W., Liu, W., & Gao, S. (2017). Remembering history with convolutional LSTM for anomaly detection. ICME.

[30] Medel, J. R., & Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv preprint.

[31] Sabokrou, M., Fathy, M., Hoseini, M., & Klette, R. (2017). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. CVPR Workshops.

[32] Shi, X., et al. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. NeurIPS.

[33] Vincent, P., et al. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR.

[34] Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

[35] Wang, T., et al. (2021). Object-centric learning with slot attention for video anomaly detection. AAAI.

[36] Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. CVPR.

[37] Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 FPS in MATLAB. ICCV.

[38] Sultani, W., Chen, C., & Shah, M. (2018). UCF-Crime dataset: Anomaly detection in real-world surveillance videos. CVPR.

[39] Nguyen, T., et al. (2019). Anomaly detection in video sequence with appearance-motion correspondence. ICASSP.

***[40]** Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection – A new baseline. CVPR.*

INDIVIDUAL CONTRIBUTION REPORT

SPATIO -TEMPORAL VIDEO ANOMALY DETECTION

Bismaya kanta Dash
(22054126)

Abstract: Accurate detection of anomalies in aerial surveillance videos is essential for maintaining security and situational awareness in large and complex environments. Spatio-temporal analysis enables systems to interpret both motion dynamics and spatial patterns within video data, enhancing the precision of anomaly detection. To improve performance, the system leverages a combination of advanced learning techniques that integrate multiple feature extraction and sequence modeling approaches. This fusion enhances the model's ability to generalize across diverse scenarios while ensuring real-time responsiveness. The solution is designed to efficiently process high-resolution aerial footage and is well-suited for applications such as traffic surveillance, border monitoring, and emergency response operations.

Individual contribution and findings: I was involved in implementing and fine-tuning hybrid deep learning models such as CNN-RNN, Autoencoder-LSTM, and 3D-CNN + Transformer. My responsibilities included designing the architecture flow, managing inter-model compatibility, and optimizing model training to improve performance across all evaluation metrics. Additionally, I performed the final evaluation by comparing hybrid vs. individual models, generating tables and visual graphs, and validating the system with real-time aerial surveillance footage. This allowed me to gain expertise in hybrid model development, integration strategies, and real-time performance assessment in complex video-based applications.

Individual contribution to project report preparation: I developed content for **Chapter 3: Hybrid Model Architectures** and **Chapter 4: Result Analysis**, including graphs and metric comparisons.

Individual contribution for project presentation and demonstration: I presented the **hybrid model results**, the **comparative graphs**, and led the **real-time testing demonstration** to showcase anomaly detection capabilities.

Full Signature of Supervisor: ………………………………

Full signature of the student:

BISMAYA KANTA DASH

INDIVIDUAL CONTRIBUTION REPORT

SPATIO -TEMPORAL VIDEO ANOMALY DETECTION

Rohan Bose
(22053345)

Abstract: Accurate detection of anomalies in aerial surveillance videos is essential for maintaining security and situational awareness in large and complex environments. Spatio-temporal analysis enables systems to interpret both motion dynamics and spatial patterns within video data, enhancing the precision of anomaly detection. To improve performance, the system leverages a combination of advanced learning techniques that integrate multiple feature extraction and sequence modeling approaches. This fusion enhances the model's ability to generalize across diverse scenarios while ensuring real-time responsiveness. The solution is designed to efficiently process high-resolution aerial footage and is well-suited for applications such as traffic surveillance, border monitoring, and emergency response operations.

Individual contribution and findings: My role focused on implementing individual deep learning models, including CNN, RNN, LSTM, Autoencoder, Transformer, and 3D-CNN architectures. I designed, trained, and evaluated these models using TensorFlow and Keras, handling configuration, hyperparameter tuning, and performance comparison based on evaluation metrics. I conducted a detailed analysis of each model’s capability in extracting spatial or temporal features independently. Through this, I gained strong practical knowledge in model design, training optimization, and comparative evaluation for spatio-temporal anomaly detection tasks.

Individual contribution to project report preparation: I documented the **Individual Model Implementations** in **Chapter 3** and contributed to the **Evaluation Metrics** section under **Chapter 4**.

Individual contribution for project presentation and demonstration: I presented the section on **individual model architecture and comparison** and participated in showing results for each standalone model.

Full Signature of Supervisor: ……………………………

Full signature of the student:

ROHAN BOSE

INDIVIDUAL CONTRIBUTION REPORT

SPATIO -TEMPORAL VIDEO ANOMALY DETECTION

Sanchit Rout
(22053979)

Abstract: Accurate detection of anomalies in aerial surveillance videos is essential for maintaining security and situational awareness in large and complex environments. Spatio-temporal analysis enables systems to interpret both motion dynamics and spatial patterns within video data, enhancing the precision of anomaly detection. To improve performance, the system leverages a combination of advanced learning techniques that integrate multiple feature extraction and sequence modeling approaches. This fusion enhances the model's ability to generalize across diverse scenarios while ensuring real-time responsiveness. The solution is designed to efficiently process high-resolution aerial footage and is well-suited for applications such as traffic surveillance, border monitoring, and emergency response operations.

Individual contribution and findings: I was responsible for dataset preparation and preprocessing, including the extraction of frames from aerial videos using OpenCV, normalization, resizing, and organizing the data into labeled categories for anomaly detection. I developed scripts for automating the preprocessing pipeline, ensuring a balanced dataset and maintaining compatibility with deep learning input formats. Collaborating closely with the modeling team, I ensured smooth integration between data and training workflows. This experience deepened my understanding of handling large-scale video data, preprocessing strategies, and their impact on model performance and generalization.

Individual contribution to project report preparation: I wrote the **Dataset Preparation and Preprocessing** section in **Chapter 3** and supported the **Introduction** section with challenges in aerial data collection.

Individual contribution for project presentation and demonstration: I prepared and presented the section covering **dataset structure, frame extraction, and preprocessing steps** used in the project.

Full Signature of Supervisor:
……………………………

Full signature of the student:

Sanchit Rout

INDIVIDUAL CONTRIBUTION REPORT

SPATIO -TEMPORAL VIDEO ANOMALY DETECTION

Anwaya Anuprash Biswal

(22053667)

Abstract: Accurate detection of anomalies in aerial surveillance videos is essential for maintaining security and situational awareness in large and complex environments. Spatio-temporal analysis enables systems to interpret both motion dynamics and spatial patterns within video data, enhancing the precision of anomaly detection. To improve performance, the system leverages a combination of advanced learning techniques that integrate multiple feature extraction and sequence modeling approaches. This fusion enhances the model's ability to generalize across diverse scenarios while ensuring real-time responsiveness. The solution is designed to efficiently process high-resolution aerial footage and is well-suited for applications such as traffic surveillance, border monitoring, and emergency response operations.

Individual contribution and findings: My contribution centered on managing dataset organization, annotation, and augmentation to enhance the training quality of the models. I structured and labeled video frames for normal and anomalous events, applied augmentation techniques using OpenCV and NumPy to diversify the dataset, and ensured balance across categories. I also conducted manual verifications to confirm the accuracy of labeled anomalies. This role helped me develop a solid understanding of data integrity, augmentation techniques, and the importance of high-quality datasets in improving deep learning model robustness and generalization.

Individual contribution to project report preparation: I authored the **Data Cleaning and Augmentation** subsection under **Chapter 3** and contributed insights on **data quality** in the **Result Analysis** section.

Individual contribution for project presentation and demonstration: I was responsible for presenting the **dataset organization and augmentation pipeline** and supported the demo with **dataset analysis slides**.

Full Signature of Supervisor: ……………………………

Full signature of the student:

Anwaya Anuprash Biswal