CHAPTER 1

INTRODUCTION

Safety and Security became prime priority in living soundings for a civilized society. To identify unusual activities, maintain social discipline and provide security, the surveillance system establishes a vital role. The traditional surveillance system helps to record the activities in the designated area and by manually monitoring these data, the persons involved in the incident are identified. These video feeds provide relevant details for the authorities enabling them for more effective response. These systems thereby enhance the confidence of the public and thereby help to move freely in the monitored areas with a sense of security.

The video surveillance system plays an invaluable role in emergency services, disaster management and crime detection. The system provides critical information and prompt responses and effective coordination in emergency services. It also contributes assistance in traffic management, urban planning and crowd control by providing insights into public space usage. These data help the city planners to design better infrastructure and allocate resources effectively.

The surveillance systems in existence depends on employees to monitor various camera feeds, that may lead to stress, errors and fatigue. Lack of automated alerts, repeated talks and excessive volume of data results in delayed responses and hinder decision making. This expensive system can also lead to employee's health issues like back pain and eye strain, which limits in large-scale network usage.

The limitations of conventional monitoring system can be overwhelmed by automated surveillance system by exploring advance technologies like Artificial Intelligence (AI) to analyze large datasets, detect anomalies and generate immediate alerts. It can also ensure unswerving monitoring, speedy incident identification and proactive responses enhancing safety of public, optimization of operations and addressing complex urban needs. Automation of surveillance systems offers efficient, scalable and dependable remedies for the drawbacks of the manual system, making it vital for community security and urban management.

1.1 VIDEO SURVEILLANCE ARCHITECTURE

Video surveillance (VS) oversees the events and behaviors in chosen area by means of Closed-Circuit Television (CCTV) cameras to avert incidents and ensure security in homes, offices and public spaces. Figure 1.1 depicts the basic architecture of VS systems, integrating analogue and Internet Protocol (IP) cameras.

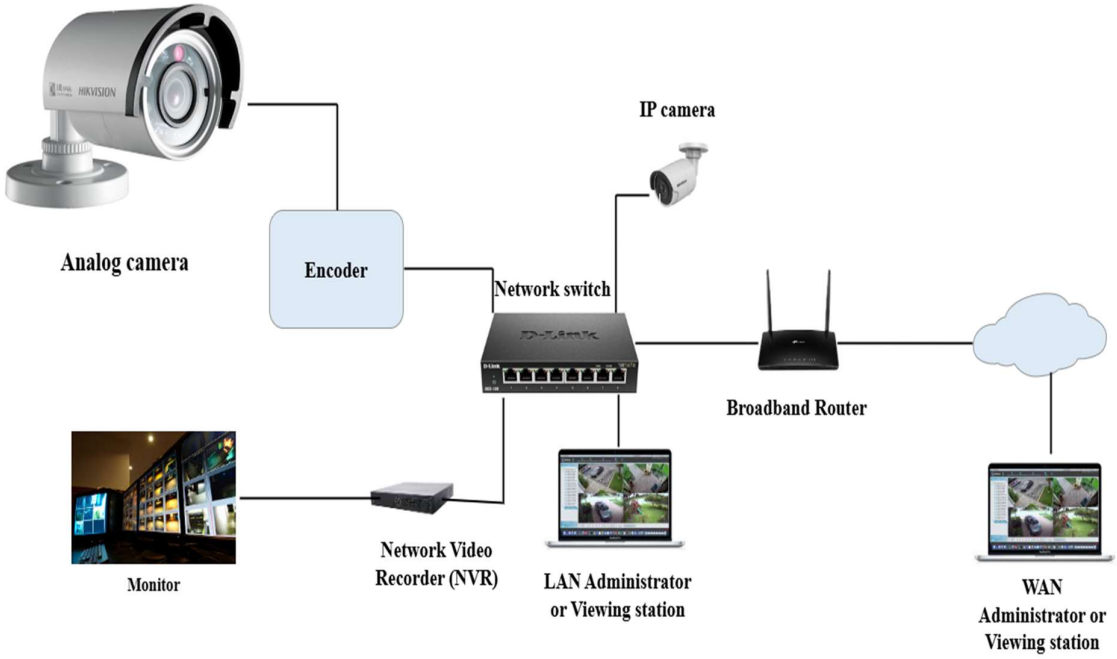

Figure 1.1 Architecture of a Simple VSS

In a video surveillance architecture, cameras act as the primary source of visual information, with their selection and deployment tailored to meet specific surveillance needs. The captured footage is encoded and transmitted through network infrastructure to a Network Video Recorder (NVR) via dedicated network switches. Various display interfaces, such as television and computer monitors, facilitate monitoring. The advent of IP cameras has transformed remote surveillance by enabling real-time video transmission across digital networks. Continuous monitoring is performed by network administrators, either stationed at a local base station within a private network or remotely managing a distributed network. These professionals ensure vigilant oversight, systematically analyzing video streams for potential security breaches or unusual activities (Tsakanikas & Dagiuklas, 2018). While

effective for basic security, the systems struggle with scalability and lack advanced features like automated detection and limiting compatibility with modern Artificial Intelligence (AI) technologies.

Digital video technologies or video analytics leverage advanced algorithms to analyze the visual data with pixel-level precision, enabling automated detection, tracking and anomaly identification within surveillance zones. These systems can autonomously recognize human movements, group activities and unusual events, eliminating the need for continuous human oversight. It significantly improves over traditional monitoring systems, addressing the inherent limitations (Cucchiara et al., 2005). Figure 1.2 explains the architecture of a video surveillance systems powered by video analytics.

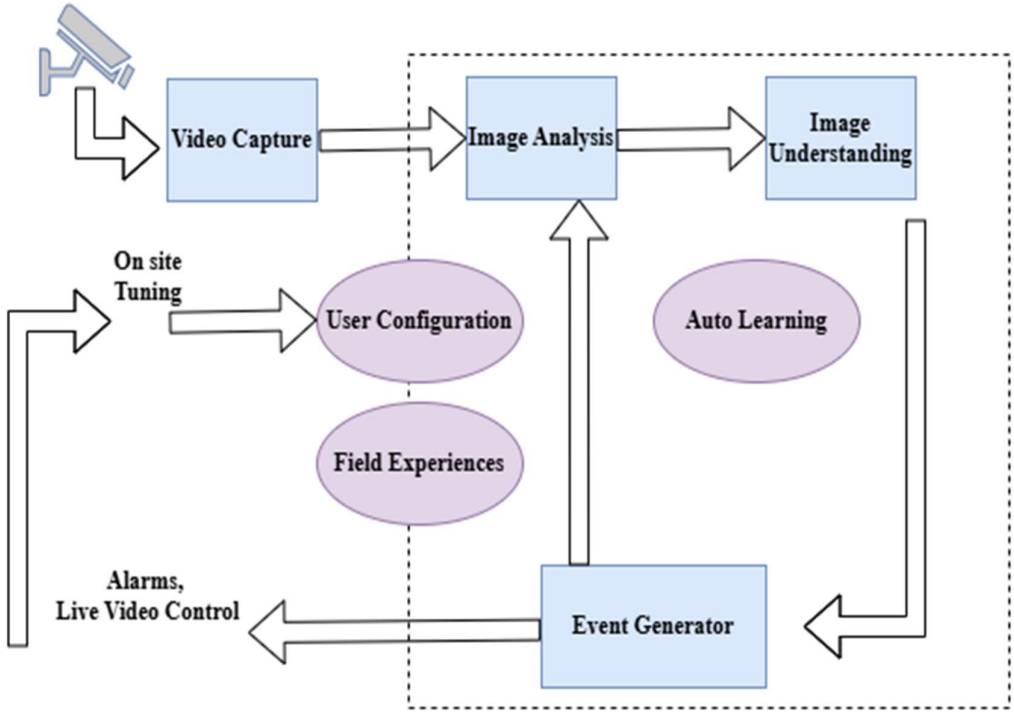

Figure 1.2 Video Analytics-Based Surveillance System

1.2 VIDEO ANOMALY DETECTION (VAD)

An anomaly is deviation from established behavioral patterns involving unexpected events or activities detected during surveillance, while Anomaly Detection (AD) identifies irregular behaviors that deviate from expected norms. Detecting anomalies requires

continuous monitoring of human and object behavior in surveillance area, as anomalous events can have significant consequences. Object-related anomalies include theft, misplacement or intentional abandonment, while human behaviors can be normal or anomalous (AL-Nawashi et al., 2027). As shown in Figure 1.3, anomalous behaviors include masked faces, repeated actions, touching protected objects, unwanted gestures and unexpected falls, which are critical for identifying potential threats. Normal activities like walking, eating and talking are less relevant for anomaly detection (Feng et al., 2021).

Figure 1.3 Classifications of Normal and Anomalous Events

1.2.1 Applications of Video Anomaly Detection (VAD)

VAD system is widely used across various fields to enhance safety, security and operational efficiency. In the military, VAD aids in threat detection, border surveillance and monitoring compact zones by analyzing real-time video feeds from drones or ground based cameras to identify potential dangers like hidden enemy movements. It also protects the military assets by detecting tampering or unauthorized access. In space exploration, VAD supports satellite image analysis, monitoring terrain changes, meteorological events or satellite damage and assists in identifying anomalies in celestial data for astronomical research. Transportation systems benefit from VAD by detecting equipment failures or unauthorized access in air, rail and maritime sectors. In contrast, infrastructure monitoring in the power sector helps to identify the hazards or failures in power plants, pipelines and grids.

VAD utilizes drone footage in agriculture sector, to monitor health of crops, condition of fields, irrigation facilities and unauthorized activities, also use it to prevent theft, fraud and crowd related anomalies. VAD can be contributed in wildlife conservation efforts by monitoring animal behavior and prevent poaching or habitual threats. Usage of VAD in smart cities comprises analysis of traffic patterns, road safety improvement and guarantee sustainable surveillance using unbiased algorithms.

In healthcare sector, applications include continuous observation of the patients, unsafe practices or equipment malfunctions, particularly in hospitals and elder care centers. VAD can also safeguard the cultural heritage by preventing theft or environmental damage to historical sites and museums. Monitoring of the anomalies like flooding or structural damages using VAD is vital in disaster preparation for effective resource prioritization and emergency rescue operations.

1.3 TRANSFER LEARNING TECHNIQUES IN VAD

Transfer learning (TL) is a significant Machine Learning (ML) approach (Pan & Yang, 2010), encompasses a pre-trained model for a new job. This method adopts acquired knowledge from large datasets like ImageNet, which contain millions of annotated images across various categories. By reusing pre-trained model layers, TL extracts feature from the

target dataset and fine-tunes the model for a specific job, significantly reducing training time, computational cost and the demand for widespread annotated data. This approach is mostly effective in fields where scarcity in availability of datasets, such as medical imaging, anomaly detection and Natural Language Processing (NLP). The conceptual framework of a TL model is pictured in Figure 1.4.

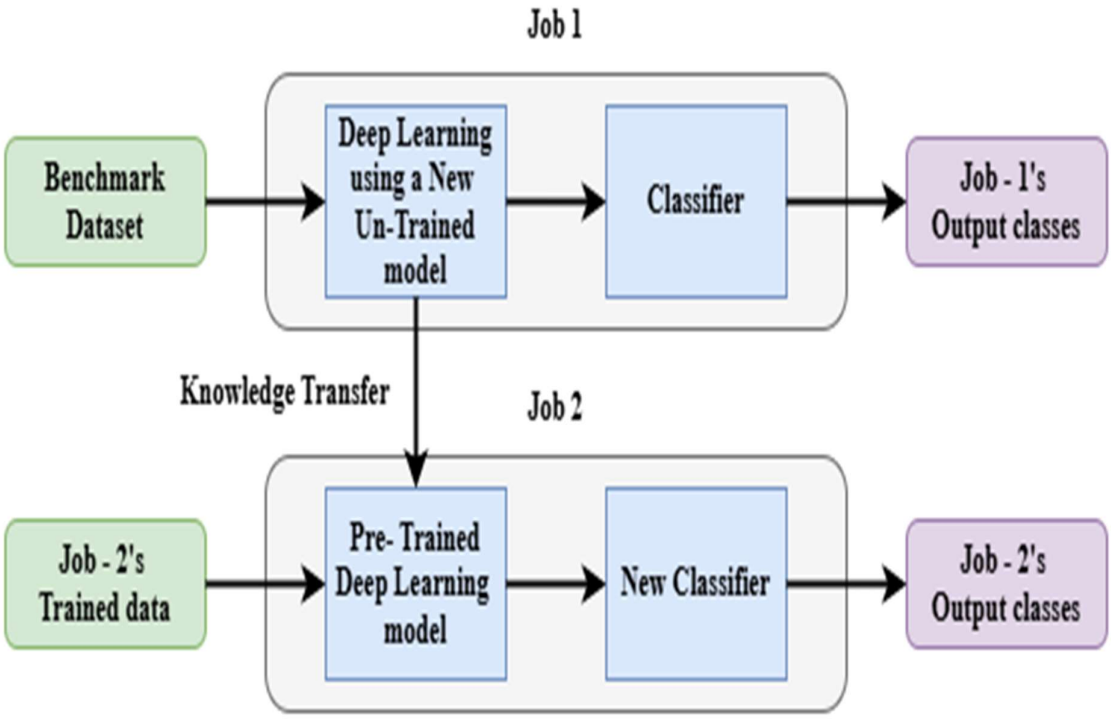

Figure 1.4 Transfer Learning Model

The TL model is built upon a pre-trained model developed using an extensive, general-purpose dataset, where the model learns generalizable features like edges, shapes or textures. These features are extracted through the pre-trained model's initial layers, which can remain frozen or undergo partial fine-tuning based on application. Next, the model is calibrated on a job-specific and compact dataset, reducing data requirements and training effort. To fulfil specific need of the new job, the model performs fine-tuning by replacing or improving the output layer, such as altering the output classes for classification. This diagram effectively demonstrates the flow of knowledge transfer from a board, generic domain to a specialized job, ensuring efficiency, adaptability and scalability, even in data-constrained environments (Sufian et al., 2020).

1.4 DEEP LEARNING ARCHITECTURES IN VAD

AI encompasses many technologies, with Deep Learning (DL) being one of its most transformative subsets. Figure 1.5 shows the AI, Machine Learning (ML) and DL Hierarchy.

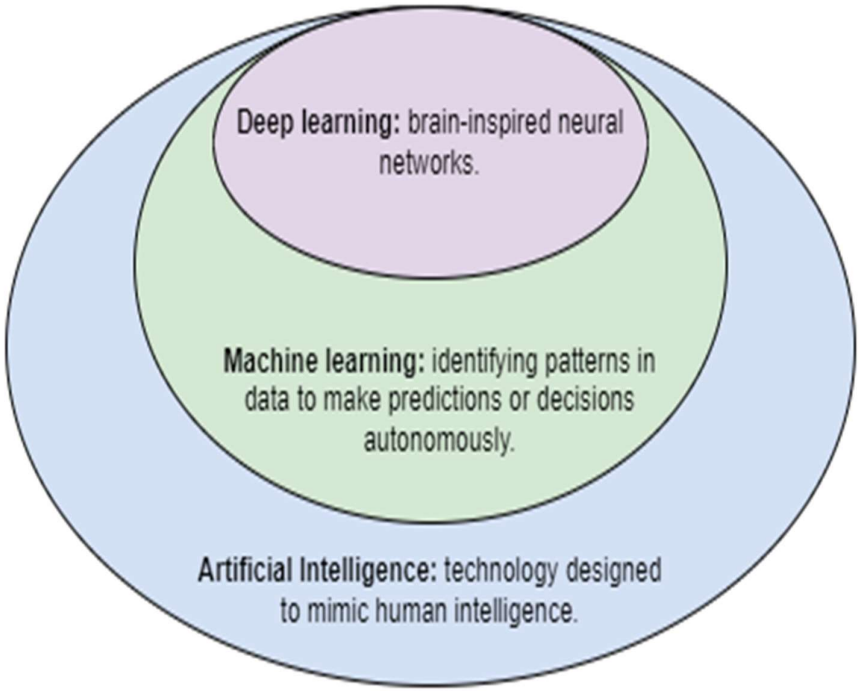

Figure 1.5 AI, ML and DL Hierarchy

AI encompasses Machine Learing (ML), which enables systems to extract data insights automatically. To model sophisticated patterns and representations DL which is a subsection of ML utilizes Neural Networks (NN). NN are computational models similar to human brain, envisioned to identify patterns and generate predictions. A neuron in NN processes inputs by computing a weighted sum of inputs, where w_i are weights, x_i are inputs and b is the bias. The calculated sum is fed into an activation function (eg. ReLU, Sigmoid), to produce non linearity which facilitates the neuron to model intricate associations (Goodfellow et al., 2026). The neuron's output is then fed to subsequent layers for further processing. The structure of NN comprises of interconnected layers of nodes (Neurons), as given in Figure 1.6.

To handle more complex patterns, NN scaled to a Multi-Layer Neural Network (MLNN) adding additional hidden layers. DL is an architecture with at least three hidden layers permitting automatic hierarchical feature extraction from data. These models can operate by means of supervised, unsupervised or semi-supervised learning techniques (Goodfellow et al., 2016).

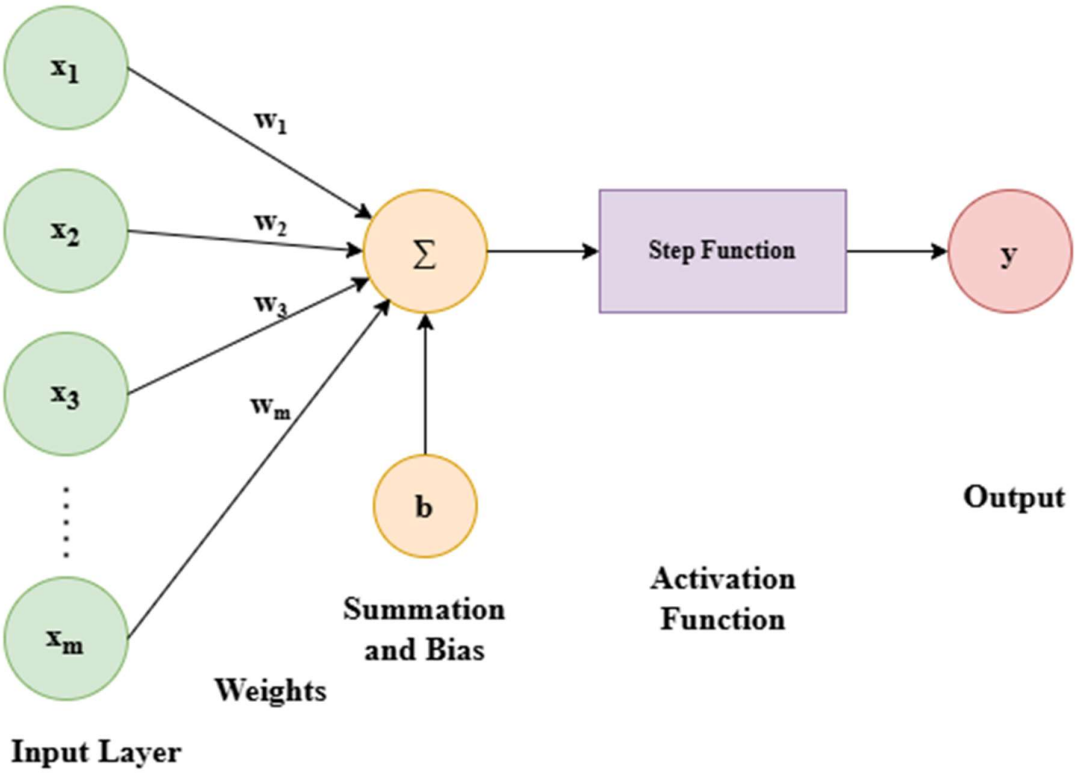

Figure 1.6 Construction of a Neural Network

1. **Supervised Learning:** It is the process of employing labeled data as input for training in order to identify the patterns as output. This technique is widely utilized in image classification and voice recognition applications.
2. **Unsupervised Learning:** It is the process of utilizing unlabeled data as input for training in order to identify the patterns as output. This method can be used for clustering and feature learning applications.
3. **Semi-Supervised Learning:** It is the process of applying a combination of labeled and unlabeled data as input for training to identify the patterns as output. This can be applied in biomedical image analysis and text classification.

A DL network figures upon the structure of NN, where neurons act as the fundamental processing units. Multiple layers of neurons arranged as input, hidden and output layers, which work together to execute feature extraction. The presence of more hidden layers ensures better feature extraction and learning representation of complex patters. Figure 1.7 provides the structure of DL, where the arrangement of neurons facilitates data flow and decision making across layers.

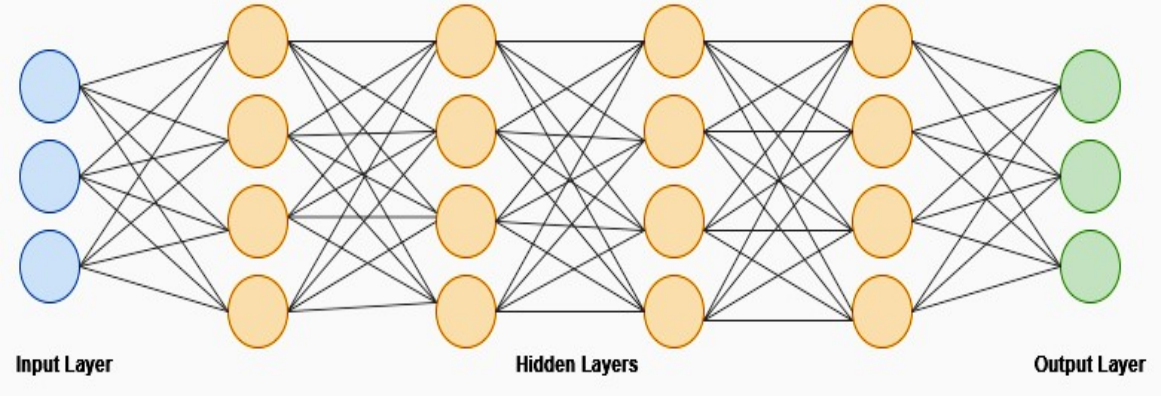

Figure 1.7 Structure of Deep Learning

The various DL models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), are greatly significant in VAD. By analyzing spatial and temporal information, these models detect complex patterns, detect anomalies and improve the accuracy of modern video surveillance process.

1.4.1 Convolutional Neural Network

CNN is characterized like the visual cortex in human beings, where neurons process designated regions of a visual region to extract spatial features. CNN has high ability to detect patterns such as shapes, edges and textures. CNNs are widely applied in image recognition, video analysis and object detection. CNNs leverage this concept to learn hierarchical patterns and features derived from unprocessed input, reducing the requirement for explicit feature design. Figure 1.8 depicts the architecture of CNN.

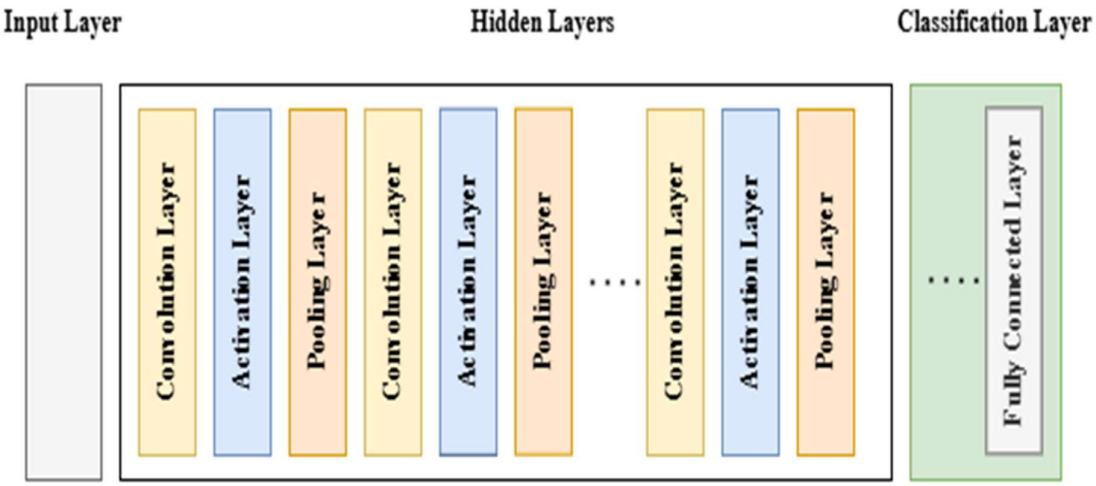

Figure 1.8 Architecture of CNN

The architecture of CNNs consists of multiple layers collaborating in tandem to derive and interpret features. The process begins with the input layer, which accepts the data, usually denoted as multi-dimensional arrays. For example, a 224×224 RGB image is described by dimensions 224×224×3, where the depth aligns to the color channels.

The convolutional layers form the backbone of CNNs. These layers leverage minimal, learnable filters to slide over the input's spatial dimensions, performing a dot product operation to create feature maps. Individual filter is crafted to capture distinct attributes like edges, textures or patterns. After performing convolution, a non-linear activation function, namely ReLU, is used to incorporate non-linearity, enabling the network to detect and imprint complex features and their correlations.

To further refine the extracted features, pooling layers are deployed. These layers diminish the spatial properties of extracted features by capturing localized patterns. Widely used pooling method like the max pooling extracts the highest value, whereas the average pooling computes the mean. Pooling layers help reduce computational costs, minimize overfitting and make the network invariant to small spatial translations.

Once feature extraction is complete, the results from the convolution and pooling layers are refined into a flat vector and traversed to fully connected layers. By combining the extracted features, these layers generate predictions. In classification problems, the output

layer utilizes an activation function such as Softmax to generate probability scores that indicate the most likely class.

The training process in CNNs involves backpropagation, where the network computes the error among forecasted and actual outputs using a loss function. By regulating the weights of the filters and neurons, the error is lessened, enabling the network to improve its predictions overtime.

CNNs have numerous benefits over conventional neural networks. CNN can automatically identify important features without manual intervention. The number of parameters can be reduced by distributing the weights to neurons and acquire hierarchical representations in order to perform complex roles in a CNN. The CNN can be applied in object detection and image classification (Alzubaidi et al., 2021).

1.4.2 Multiscale CNN

Figure 1.9 illustrates the Multiscale CNN (M-CNN) architecture, in which concurrently process input images at multiple spatial resolutions. The model can extract minute details at high resolution and broader understanding of information for low resolution inputs. This design is particularly effective for images where identifying patterns at various scale is essential. An input image of size 1024 x 1280 pixels, which is processed through multiple resolution pathways to capture diverse spatial features in a M-CNN architecture. Each pathway operates at a specific spatial resolution, with the highest resolution pathway preserving fine details. In contrast, subsequent pathways progressively down sample the image using pooling layers to extract global information efficiently. Pathways consist of convolutional layers with Kernel sizes like 5x5, succeeded by bias terms and nonlinear activation functions, allowing the network to learn hierarchical feature representations.

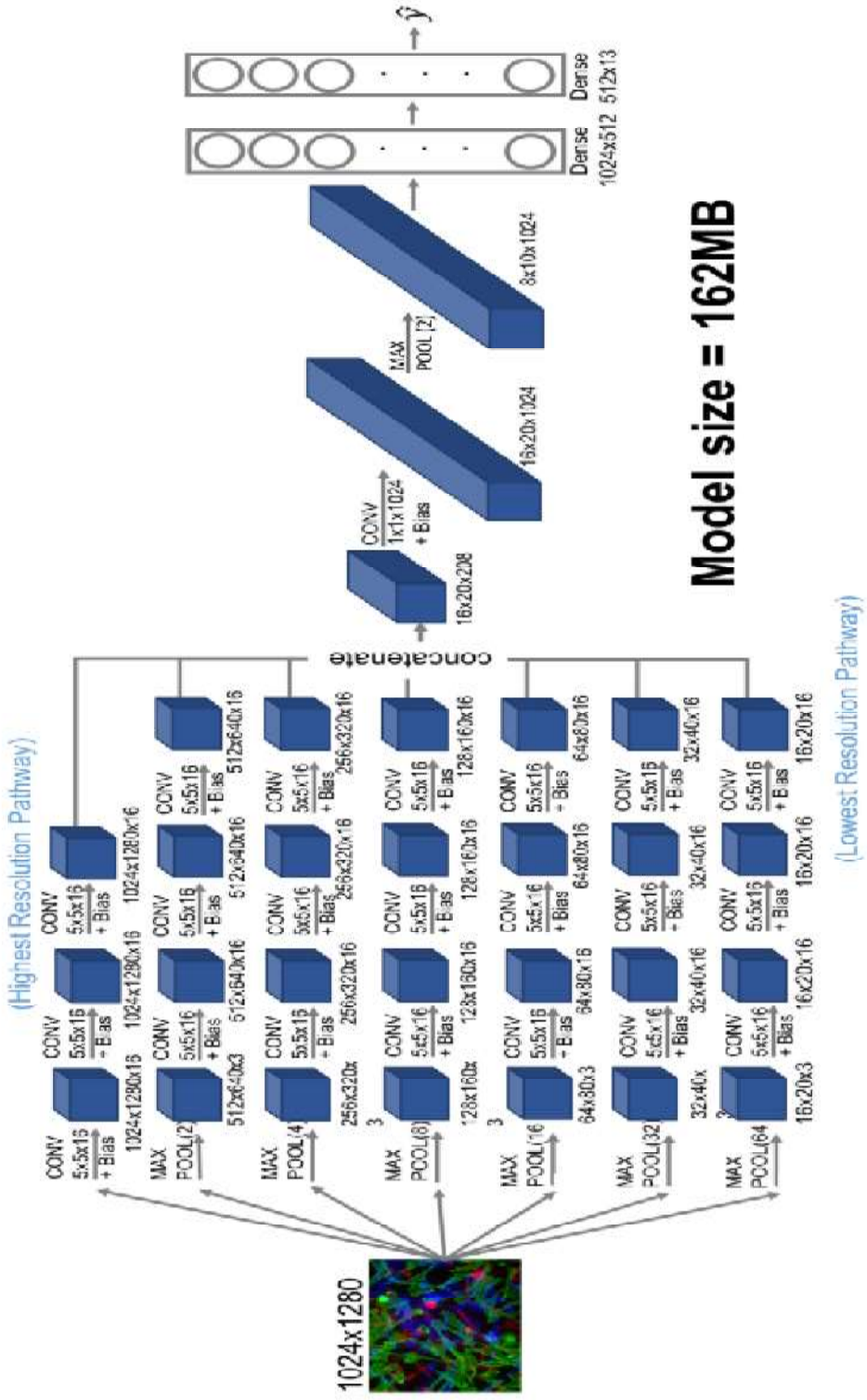

Figure 1.9 Kernels and Operations in the M-CNN model (Datta et al., 2019)

Outputs from all pathways are fused via concatenation, combing local and global information into a unified feature representation for robust learning. The fused representation is then fed in to a fully connected dense layers, progressively reducing dimensions to correlate the characteristics to the output classes. With a model size of 162MB, the architecture is efficient and scalable, ideal for managing massive datasets. The multiscale design optimizes computational performance with classification accuracy, making it exceptionally appropriate for analyzing images with carrying structures and patterns.

1.4.3 Long Short-Term Memory (LSTM)

The vanishing gradient concern can be solved by employing the LSTM (Hochreiter & Schmidhuber, 1997) which is the enhanced version of Recurrent Neural Network (RNN). LSTMs employ a dedicated memory cell that preserves significant information across long sequences. Three main gates, namely input, forget and output, regulate the LSTM cell. The input gate regulates which new data to be stored in the cell if necessary. Forget gate controls the unnecessary information to be discarded in the cell. The output gate controls the output of information stored in the cell. The gate architecture enables LSTM to enable long-term relationships and makes it suitable for sequential prediction. The capability of conserving necessary information helps the LSTM model to process the sequential data.

1.4.4 Convolutional Long Short-Term Memory

The Convolutional LSTM (ConvLSTM) (Huang H. et al., 2022) processes spatial and temporal dependencies of data, by integrating CNNs with LSTM. The model utilizes the new data stored in the input gate to retain spatial features. The redundant information are discarded using forget gates to focus on critical patterns. Convolutional layers enabling spatial feature extraction from video frames. The data flow of the cells is regulated by the output layer.

The model initially executes convolution operations and extracts spatial features. To identify the spatio-temporal relationship, temporal modeling is performed using LSTM gates. This model can be applied for spatiotemporal analysis task.

1.4.5 You Only Look Once (YOLO)

YOLO: "real-time object detection model". YOLO attain object detection by integrating the inputs into a single neural network during recognition phase. This feature results in faster response object detection (Thuan, 2021). For example, YOLOv1 uses a 7 x 7 grid structure, with each grid cells predicting parameters such as bounding box location, dimensions and confidence scores. Figure 1.10 displays the YOLOv1 Model with a 7 x 7 Grid structure. Each version of YOLO, from YOLOv1 to YOLOv8, has gradually advanced real-time object detection by increasing speed, accuracy and scalability. This technology is apt for applications like autonomous vehicles, surveillance and image analysis.

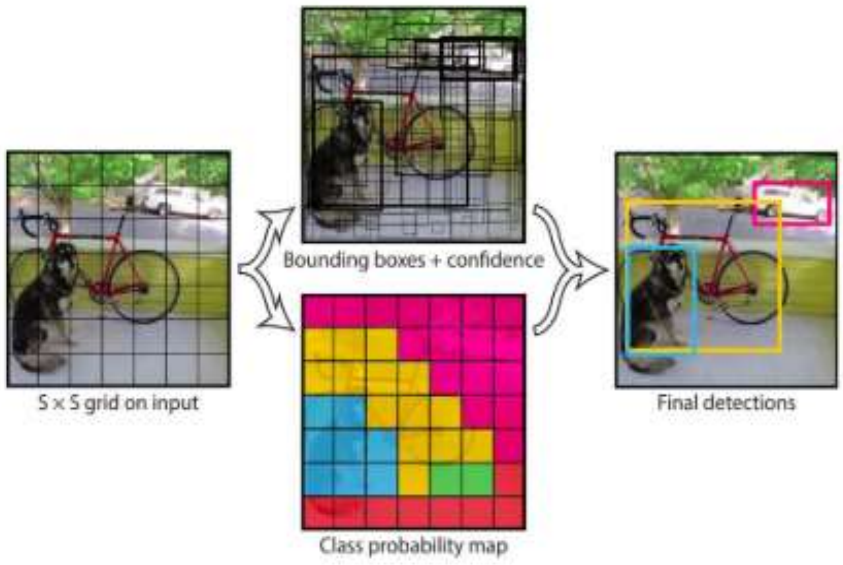

Figure 1.10 7x7 Grid Cell of YOLO model Applied to an Input Image (Thuan, 2021)

1.4.6 ResNet-50 Architecture

To address the vanishing gradient problem in deep architectures, ResNet-50 (He et al., 2015) is an alternate solution, which is a deep CNN that utilizes residual learning technique. The ResNet family includes models of varying depths, such as ResNet-18, ResNet-34, ResNet-50, ResNet-101 and ResNet-152, progressing from lighter to deeper architectures. Lighter models like ResNet-18 and ResNet-34 are idyllic for faster computation and resource constrained tasks, while deeper models such as ResNet-101 and ResNet-152 handle complex datasets and capture intricate spatial features. The ResNet-50

having moderate depth provides average complexity and performances making it appropriate for several image classification applications.

1.4.7 UNet Architecture

The U-shaped Network (UNet) (Rosenberger et al., 2015) is widely for image segmentation. The UNet structure consists of an encoding-decoding structure, that performs up and down sampling of inputs for feature extraction. Feature extracted from the earlier layer is transferred to the next level during upsampling and thereby improving feature extraction. The construction of UNet is imprinted in Figure 1.11.

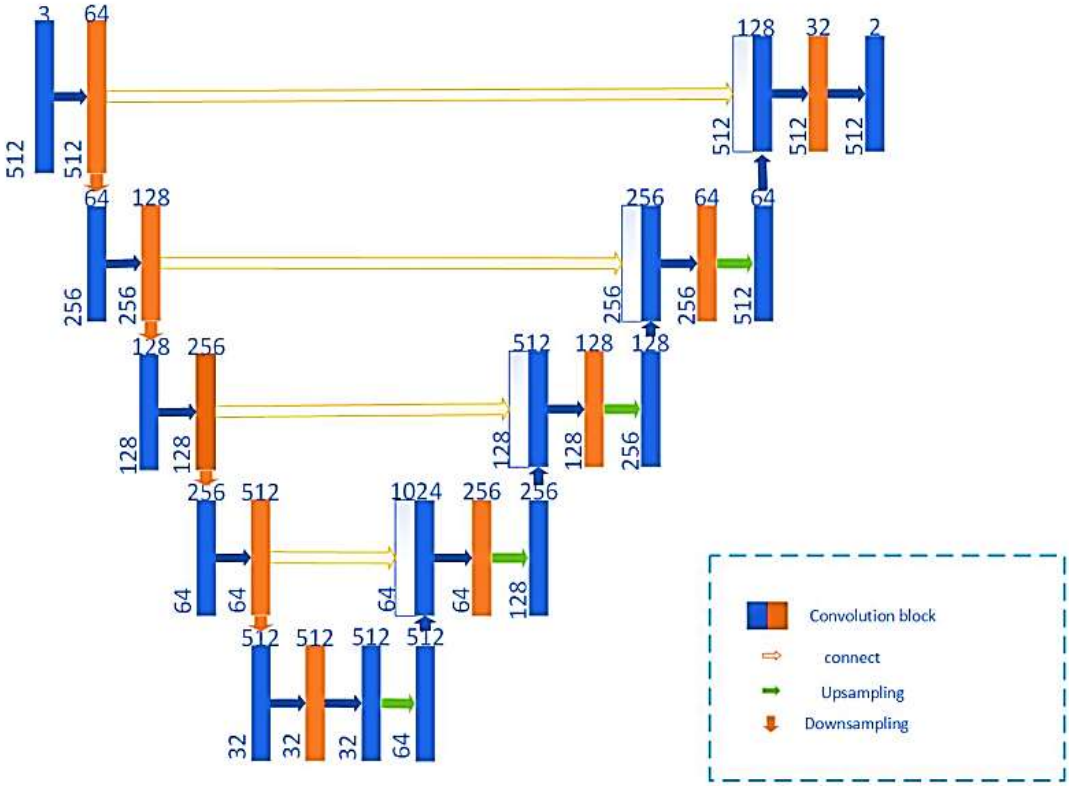

Figure 1.11 Structure of UNet (Chen et al., 2021)

UNet Structure can be applied not only in medical image segmentation, but also for various image segmentation and analysis tasks. UNet's capability to produce accurate pixel-level predictions has established it as a versatile solution for various image segmentation.

1.5 PROCESS OF VIDEO ANOMALY DETECTION USING DL

VAD encompasses a systematic channel to analyze the raw video input and differentiate the unusual patterns or activities. Figure 1.12 illustrates the process of the VAD.

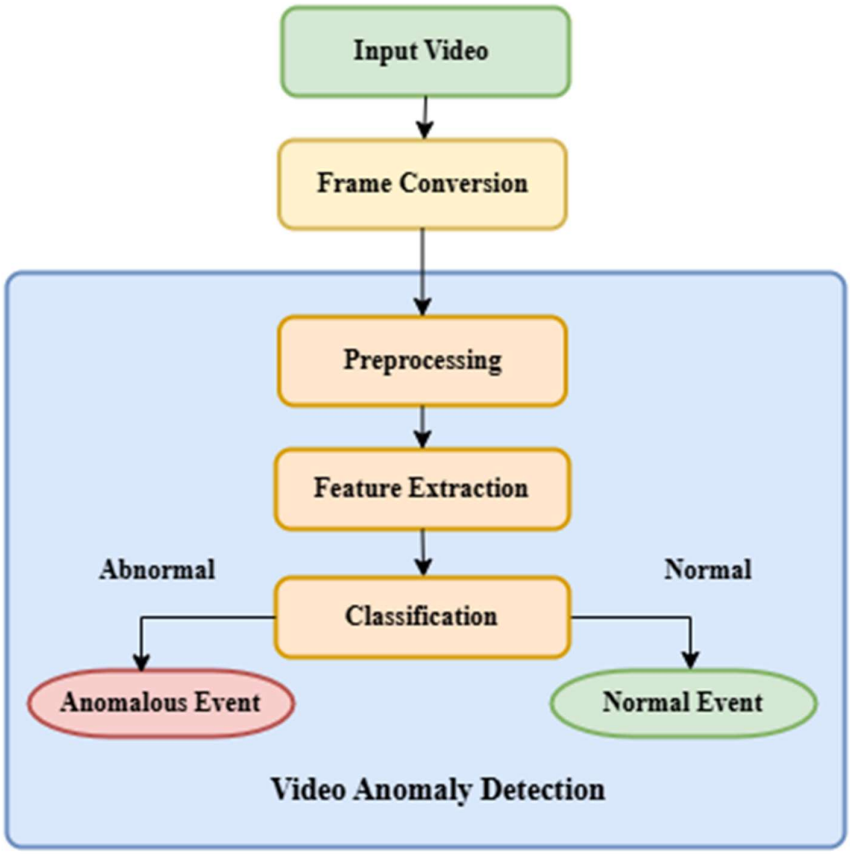

Figure 1.12 Process of Video Anomaly Detection

The process begins with acquiring video data captured from the surveillance systems or pre-recorded sources in video format. These videos are then decomposed into individual frames through frame conversion, allowing for frame-by-frame analysis. The frame extraction rate (e.g., 30 frames per second) determines the granularity of the analysis and provides the foundation for subsequent processing.

Once the frames are extracted, preprocessing is conducted to refine the quality and consistency of the information, which includes resizing the frames to standard dimensions, normalizing pixel values and utilizing the noise reduction techniques like Gaussian filtering

or histogram equalization. Data augmentation methods like flipping, rotation or cropping can be applied to create a more diverse dataset.

Following the processing step, feature extraction is conducted to recognize the spatial and temporal aspects of the video. Spatial features, which reflect the content of individual frames, are typically extracted using CNN models. Temporal features, essential for understanding motion and sequence patterns, are processed using RNN models.

The extracted features are the provided to the VAD component, where transfer learning techniques refine the models to adapt to the specific dataset and context to ensure that the system learns to identify domain-specific anomalies effectively. The derived features are finally channeled into the classifiers for prediction to categorize the frames or sequences as normal or anomalous. Post -processing techniques like smoothing can be applied across the sequences to ensure prediction stability.

1.6 MOTIVATION OF THE RESEARCH

The growing need for public safety underscores the importance of detecting abnormal events promptly to protect vulnerable groups such as women, children and the old. Rising incidents of theft, harassment and violence further emphasize the necessity of proactive systems that prevent harm rather than identifying culprits after an incident. The traditional surveillance systems vulnerable to significant restrictions, such as the impracticableness of manual video monitoring, increased false positive rates and impracticability in processing large-scale video data. In dynamic or crowded scenarios anomaly detecting accuracy is inadequate for existing VAD systems which delays emergency actions to mitigate the anomaly.

1.7 RESEARCH GAP

The inability of existing VAD to capture complex spatial and temporal relationships from video inputs results in inefficient anomaly detection. Present models have diminished anomaly distinguishing capacity and categorization ability. The absence of segmentation-based classification techniques restricts effective anomaly localization, reducing detection accuracy. All these limitations in properly detecting anomalies in various instance creates obstacles in taking mitigation and preventive measures against potential threats.

1.8 PROBLEM STATEMENT

Due to the inefficiency in processing extensive video data creates challenges to the existing surveillance systems and making hindrance for real time monitoring of surveillance areas. Increased false alarms due to high false positive rate divert attention from real threats. Poor integration of spatial and temporal features limits the accurate analysis of anomaly events. Difficulty in recognizing pattens in changing surroundings, inaccurate threat assessments reduce the surveillance effectiveness. Inadequate real-time processing delays emergency responses and dependance on manual monitoring increases the risk of fatigue, errors and inconsistencies. These restrictions reduce the reliability, scalability and overall efficiency of surveillance systems, enforces the need for advances automated solutions to enhance public safety.

1.9 CHALLENGES IN VAD

The primary challenges faced in the VAD research:

1. **High false alarm rates:** Many VAD models misclassify normal events as anomalous, leading to false positives that reduce the reliability of the system.
2. **Spatio-Temporal Complexity:** To identify anomalous events, a strong understanding of spatial and temporal patterns is required, and it is lacking in video frames consisting of dynamic and crowded scenes.
3. **Inefficient handling of Low-Resolution Videos:** Many surveillance systems operate with low-quality inputs, which may struggle to maintain accuracy.
4. **Overfitting and poor generalization:** Some models perform well during training but may fail during testing, especially in real-world scenarios having new or unseen anomaly patterns.
5. **Dependency on Manual Monitoring:** Manual monitoring increases the risk of biased or fatigue-related issues and inconsistencies, making automated solutions essential.

1.10 OBJECTIVES OF THE RESEARCH

The objectives of the research:

1. To build a reliable VAD model using feature extraction and improved object detection model and evaluate its performance with existing models.

2. To develop a hybrid model by integrating spatial feature extraction and temporal sequence learning.

3. To improve VAD by focusing on spatiotemporal segmentation technique to improve accuracy and diminish overfitting.

4. To enhance VAD in low-resolution video using hybrid deep learning approach by capturing spatiotemporal dependencies and minimizing noise.

1.11 CONTRIBUTIONS OF THE RESEARCH

The thesis pursues the possibilities of advanced deep learning methods to enhance VAD. It Implement hybrid VAD models and analyzes their performance, comparison and validation of results.

● **CNN-YOLO Model for AD**: In this work, CNN and YOLO architectures are hybridized for feature extraction and object detection. The spatial features extracted from CNN are optimized using Adam's Optimization during training. The testing of the model consists of a series of operations, like Histogram Equalization, Euclidean Distance Tracking, Post-Processing, Non-Max Suppression and Edge Based Segmentation. Giving priority to faster object detection, the CNN-YOLO model is developed by integrating a modified YOLOv4 network. This model processes a random frame out of every 100 frames to maintain a balance between speed and detection accuracy. The model offers increased anomaly detection capability and can be deployed effectively in low-resource hardware, enabling faster response in real-time applications.

● **ResNet-LSTM for Video Anomaly Recognition**: A hybrid model that captures the spatial and temporal features using ResNet-50 and LSTM architectures are incorporated in this research. The deep learning model, ResNet-50, explores deep feature extraction due to its residual connections and integrates LSTM for temporal dependency analysis of video input. Computational efficiency and performance are balanced by ResNet-50 and LSTM to precisely distinguish the normal and anomalous patterns. The model handles continuous video frames for anomaly detection in a dynamic environment.

● **Improved UNet-Cascade Sliding Window Technique (CWST) for AD:** The overfitting issue in VAD, which affects accuracy and efficiency, is addressed in this research using an Improved UNet (IUNet) architecture. Wiener filter preprocessing technique enhances the noise reduction, which leads to precise anomaly detection. The model uses an encoder-decoder UNet architecture for efficient feature extraction. This segmentation-based model's performance is improved by including a ConvLSTM layer in it, which enhances the spatial and temporal feature extraction. The CSWT is used to estimate the anomaly score and to categorize unusual events. The development of the IUNet-CSWT model enables improved performance in noisy video input frames by applying a Wiener filter and thereby improving anomaly detection capability.

● **Hierarchical Fusion Model for Intelligent Surveillance:** The anomaly detection in low-resolution and noisy videos, is performed by hybridization of Hierarchical Multiscale-CNN and LSTM. The preprocessing of input frames using Bilateral-Wave Denoise technique helps to reduce noise and preserve edges, thereby improving the image quality of low-resolution videos. The Hierarchical Multiscale-CNN enables the extraction of multiscale spatial features and the Spatial Pyramid Pooling (SPP) is integrated to ensure scale invariance during feature aggregation. Along with this, the temporal features are extracted using LSTM layers to achieve superior generalization and adaptability to real-time videos.

The hybrid Hierarchical Multiscale-CNN and LSTM model also employs advanced layers such as flatten, dense and SoftMax for classification of normal and anomalous events, which shows exceptional performance in VAD. The model provides improved detection accuracy and reconstruction quality on the benchmark dataset. The decreased EER indicates that the model has fewer classification errors. Additionally, the improved AUC reflects the model's ability to classify normal and anomalous events. The obtained performance metrics highlight the model's efficiency and reinforce its potential for real-time deployment in surveillance-based VAD systems. Automatic anomaly detection in real-time helps to enhance safety and security by providing timely updates and foresight, thereby preventing anomalous events.

1.12 ORGANIZATION OF THE THESIS

The structure of the thesis is as follows:

Chapter 1 explains surveillance systems and VAD methods and its applications. It also discusses the various DL architectures including CNNs, YOLOv4, ResNet-50, UNet and LSTM networks, highlighting their roles in enhancing system performance.

Chapter 2 describes a literature review on VAD. Transfer learning and DL techniques in VAD is narrated the present research trends and challenges.

Chapter 3 illustrates the CNN-YOLO object detection model for identifying unusual behavior in Video surveillance.

Chapter 4 proposes a hybrid ResNet-50 and LSTM model to improve VAD. Combination of spatial and temporal data analyzes to enhance anomaly detection accuracy is explained.

Chapter 5 explores the techniques for enhancing VAD, using an improved UNet architecture for segmentation and Cascade Sliding Window method for classification.

Chapter 6 describes the hybrid Multiscale CNN and LSTM anomaly detection approach that combines advanced feature extraction and temporal analysis techniques to improve detection accuracy in low-resolution video.

Chapter 7 presents the summarized performance analysis of the proposed algorithms for VAD.

Chapter 8 concludes with summary of findings, discussing the limitations and potential for future improvements.

1.13 SUMMARY

This chapter explores video surveillance systems and various VAD methods along with their applications. It also examines different deep learning architectures used in VAD.