algorithms

MDPI

*Article*

# SynthSecureNet: An Improved Deep Learning Architecture with Application to Intelligent Violence Detection

**Ntandoyenkosi Zungu, Peter Olukanmi * and Pitshou Bokoro**

Department of Electrical and Electronic Engineering Technology, Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg 2092, South Africa; 223048675@student.uj.ac.za (N.Z.); pitshoub@uj.ac.za (P.B.)

* Correspondence: polukanmi@uj.ac.za

**Abstract:** We present a new deep learning architecture, named SynthSecureNet, which hybridizes two popular architectures: MobileNetV2 and ResNetV2. The latter have been shown to be promising in violence detection. The aim of our architecture is to harness the combined strengths of the two known methods for improved accuracy. First, we leverage the pre-trained weights of MobileNetV2 and ResNet50V2 to initialize the network. Next, we fine-tune the network by training it on a dataset of labeled surveillance videos, with a focus on optimizing the fusion process between the two architectures. Experimental results demonstrate a significant improvement in accuracy compared with individual models. MobileNetV2 achieves an accuracy of 90%, while ResNet50V2 achieves a 94% accuracy in violence detection tasks. SynthSecureNet achieves an accuracy of 99.22%, surpassing the performance of individual models. The integration of MobileNetV2 and ResNet50V2 in SynthSecureNet offers a comprehensive solution that addresses the limitations of the existing architectures, paving the way for more effective surveillance and crime prevention strategies.

**Keywords:** ensemble model; hybrid model; SynthSecureNet; 2D CNN; deep transfer leaning; MobileNetV2; ResNetV2; surveillance alert; violence detection

check for updates

Academic Editor: Abdulsalam Yassine

Received: 30 October 2024
Revised: 24 December 2024
Accepted: 25 December 2024
Published: 10 January 2025

**Citation:** Zungu, N.; Olukanmi, P.; Bokoro, P. SynthSecureNet: An Improved Deep Learning Architecture with Application to Intelligent Violence Detection. *Algorithms* **2025**, *18*, 39. https://doi.org/10.3390/a18010039

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

The use of closed-circuit television (CCTV) systems is of critical importance in modern society for the purpose of ensuring public safety and security [1,2]. Surveillance systems function as a preventive measure [3] against unlawful activities and improper conduct by means of ongoing monitoring in residential areas, public spaces, transportation hubs, and transportation hubs [4–6]. In addition to preventing crime, surveillance technologies also help in quickly identifying and capturing people involved in illegal activities [7]. This enhances the capacity of law enforcement and allows for timely reactions to emerging threats. Furthermore, surveillance data are a crucial kind of forensic evidence in legal procedures, assisting in determining the guilt of criminals and guaranteeing justice for crime victims [8,9]. Surveillance is a fundamental aspect of proactive security measures, enhancing society's resilience against many possible threats and vulnerabilities [10,11].

Violence detection is one subfield of surveillance that is extremely important to society because it provides an early warning system for identifying and reducing physical violence, conflict, and injury [12,13]. Violence detection systems are able to automatically identify violent incidents on CCTV footage and report them for immediate intervention by utilizing advanced deep learning techniques, such as the proposed SynthSecureNet architecture. The proposed architecture can be integrated with violence systems that do not only allow law

enforcement and security personnel to respond quickly and improve public safety, but also give citizens a sense of security and assurance. Furthermore, assisting law enforcement in paying attention to tense situations before they go out of control will reduce chances of people/communities getting harmed, and also violence detection technologies are able to stop conflicts from getting worse. The proactive approach to maintaining peace and order in society that is showcased by the integration of violence detection systems into a surveillance infrastructure highlights the innovative influence of technology on public safety initiatives [14–16].

As a result, in this paper, a hybrid deep learning architecture is proposed with enhanced performance when compared with other deep learning architectures for violence detection.

"SynthSecureNet" is a unique name for the proposed ensemble architecture. The term "Synth" suggests synthesis or combining, and "SecureNet" implies a focus on security. This name effectively communicates that the proposed model is a synthesized ensemble designed for security-related tasks.

Zahid et al. [17] proposed a bagging framework known as IBaggedFCNet. This framework applies ensembles for classification and anomaly detection for motion pictures or videos. The authors used a complex Inception-v3 image classification network in their technique and did not use segmentation prior to extracting features from the video. This eliminated the chance for segmentation results from cross interference and reduced the memory usage. They demonstrated significant improvements on different benchmark datasets, with the best improvements recorded when using the UCF-Crime dataset. They conducted several experiments on several ensemble fusion techniques, such as static and dynamic fusions [17–19].

The SynthSecureNet architecture stands out from other ensemble architectures based on the combination of the ResNet50V2 and MobileNetV2 architectures. Previous studies have primarily dealt with integrating models that follow similar structural frameworks or using temporal–spatial data separately; on the other hand, SynthSecureNet combines the deep feature extraction strength of ResNet50V2 with the processing speed and light weight of MobileNetV2. This integration also helps to strengthen the model's ability to recognize potential signs of violence and increases the scope of implementing the system in real time with minimal computational power [20–22].

Additionally, the ensemble design for SynthSecureNet is carefully tuned and optimized so that it achieves the best performance in terms of both accuracy and speed. ResNet50V2 is a widely used solution that is ideal for capturing the details and complex patterns in the input data, while MobileNetV2 is designed for specific tasks for edge computing applications, and it has a very effective and simple architecture [23,24]. This combination allows SynthSecureNet to work quicker and achieve better results than other bagging methods that require speed at the expense of accuracy or vice versa. The architecture also implements the strategy for feature fusion and multi-level classification better than the original model so that it can recognize violent actions more intricately and effectively distinguish different actions of violence.

SynthSecureNet has several new qualities that are absent in other violence detection systems. First, it presents a dual-architectural framework that ensures improved flexibility to different surveillance setups, whether they refer to high-resolution urban footage or low-quality video streams from remote locations [21,22]. The fact that the architecture can work effectively on edge devices like smart cameras or mobile units is a step forward because it eliminates the time delay in transmitting data to centralized servers for processing [13].

This hybrid alert system ensures that critical alerts are delivered in real time even in environments where connectivity is limited. The proposed model was trained using a

large dataset. Overall, these advancements facilitate the application of SynthSecureNet as a complete predictive system that can comprehensively provide real-time violence detection on CCTV systems for addressing the missing pieces in prior ensemble architectures.

### *1.1. State-of-the-Art Overview*

Recent advancements in deep learning have significantly improved the performance of violence detection systems. Techniques such as ensemble learning, temporal–spatial data modeling, and hybrid architectures have demonstrated their effectiveness. For example, Zahid et al. [17] introduced IBaggedFCNet, a bagging framework leveraging Inception-v3 for anomaly detection. However, this approach lacks segmentation capabilities, which limits its application to complex real-world scenarios. Other approaches, such as Conv3D-based networks, employ temporal feature extraction but are computationally expensive [7]. Despite these advancements, challenges remain in achieving high accuracy while maintaining real-time performance.

### *1.2. Proposed Contribution*

This study presents SynthSecureNet, a novel hybrid architecture combining the strengths of popular MobileNetV2 and ResNet50V2. SynthSecureNet improves violence detection accuracy through feature fusion while maintaining computational efficiency, making it suitable for real-time deployment in edge devices. The architecture achieves a remarkable 99.22% accuracy, surpassing the original methods, and offers a robust solution for public safety initiatives.

## 2. Methodology

### *2.1. The Architecture*

The SynthSecureNet architecture was built by integrating two well-known deep learning architectures, namely, MobileNetV2 [25] and ResNet50V2 [26]. Our aim was to work on both architectures and fuse them to obtain improved performance for violence detection in videos. MobileNetV2 is a convolutional neural network architecture that is based on the previous multiple versions of MobileNets, which aims to design more efficient deep networks used for mobile and edge devices [27]. It uses the concept of depthwise separable convolutions and inverted residual blocks, which helps in minimizing the number of parameters and computational cost with high performance.

ResNet50V2 (short for residual networks) introduces residual connections to deep Convolutional Networks, which helps in solving a vanishing gradient issue [28]. It is a series of residual blocks. That is where the 'residual' in 'ResNet' comes from because each block makes it very easy to learn residual functions with respect to the layer inputs and help you create much deeper networks [29].

Although state-of-the-art methods such as SlowFast [30], ResNet 3D [31,32], and Video Swin Transformer [33] offer advanced temporal–spatial modeling for video-based tasks, their inclusion was not feasible in this study due to computational limitations. These methods require extensive hardware resources, which were beyond the scope of the current experimental setup. Future research will benchmark SynthSecureNet against these methods to further validate its performance and adaptability.

### *2.2. Combining ResNet and MobileNet*

As stated in the above paragraph, the SynthSecureNet architecture is basically a combination of two architectures, namely, ResNetV2 and MobileNetV2 architectures. When initializing the model, MobileNetV2 and ResNet50V2 are initialized through pre-trained weights on ImageNet data. The pre-trained weights are a great starting point for feature

extraction as they have learned useful representations from a large-scale image classification problem [34,35].

Both MobileNetV2 and ResNet50V2 models are used as feature extractors, which involves removing their classification layers and utilizing the outputs from their last convolutional layers as feature maps. For a better explanation, let $f_m(x)$ and $f_r(x)$ depict the feature maps extracted by MobileNetV2 and ResNet50V2, respectively, for an input x.

Along the channel dimension, features extracted from MobileNetV2 and ResNet50V2 were concatenated. This concatenation ensures that the model learns from the non-duplicated information in both architectures. The aforementioned concatenated feature map $f_c(x)$ can be mathematically shown as (1). The number of channels on the resulting feature map $f_c(x)$ is greater than single feature maps [34,36].

$$f_c(x) = contat(f_m(x),\ f_r(x)) \tag{1}$$

Then, the concatenated feature map $f_c(x)$ is processed by extra layers to extract higher-level representations and perform classification. Additional layers include a flattened layer where the spatial feature map is converted into a one-dimensional vector. Additional layers also include dense layers, which are fully connected and learn non-linear combinations of features. Further, they also include dropout layers, which are a technique that performs regulation to prevent overfitting, where units are dropped out during the training duration [37].

This can also be denoted as the output of those extra layers (2), as follows:

$$y = g(f_c(x)) \tag{2}$$

The function of $g()$ corresponds to the mapping learned by additional layers, which takes the input as a concatenated feature map $f_c(x)$ and the output as y.

On the output layer, a sigmoid activation function is used in the final dense layer that produces binary output to classify whether the video contains violence or not [38,39]. Here, the output is squashed between 0 and 1; hence, it can be considered as a probability value. The output of the SynthSecureNet model can be defined as follows (3):

$$p(y \mid x) = \sigma(w \cdot g(f_c(x)) + b) \tag{3}$$

where

$w$ is a weight vector;

$b$ is a bias term;

$w \cdot g(fc(x)) + b$ is a linear combination of the transformed features.

We can consider the following sigma equation:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Thus,

$$p(y \mid x) = \frac{1}{1 + e^{-(w \cdot g(f_c(x)) + b)}} \tag{5}$$

This equation models the probability $p(y|x)$ as a sigmoid function of the linear transformation of the features obtained from the concatenated outputs of MobileNetV2 and ResNetV2 [38]. The function corresponds to the mapping learned by additional layers, which takes the input of concatenated features.

The SynthSecureNet model was trained using the binary cross-entropy loss function, which measured the dissimilarity between the predicted probabilities and the true labels.

The Adam optimizer was used to update the model's weights based on the gradients computed during backpropagation [40,41].

### *2.3. SynthSecureNet Architecture Network Development*

The input to the SynthSecureNet model is a video, which undergoes further preprocessing and is converted into a sequence of frames. The input video is denoted by V, whereas N represents the number of frames in it; refer to the following Equation (5):

$$V = \{f_1, f_2, \ldots, f_N\} \tag{6}$$

where $f_i$ is an image (each frame).

In a preprocessing phase, before exposure to the model, the frame (or stack of frames) goes through data augmentation processing that increases training sample diversity and improves generalization. Some of these techniques that were used in augmentation include the following [18]:

1. Random flipping;
2. Random zooming;
3. Random brightness adjustment;
4. Random rotation.

For better explanation purposes, we denote an augmented frame as $f_i^{aug}$.

In the feature extraction phase of the proposed system, the augmented frames were then passed through MobileNetV2 and ResNet50V2 to extract the features. These models were initialized with pre-trained weights of what is commonly known as ImageNet dataset [29].

MobileNetV2 is made up of a series of depthwise separable convolutions and inverted residual blocks.

The output of the MobileNetV2 model for the frame $f_i^{aug}$ is denoted as follows:

$$f_m\left(f_i^{aug}\right) = MobileNetV2\left(f_i^{aug}\right) \tag{7}$$

where the feature map $f_m(f_i^{aug})$ has the shape of $(h_m, w_m, c_m)$, where $h_m$, $w_m$, and $c_m$ are the height, width, and number of channels, respectively.

Given that the ResNet50V2 model is made up of a sequence of residual blocks, each block learns a residual function.

The output of this ResNet50V2 model for the frame $f_i^{aug}$ can be written as follows:

$$f_r\left(f_i^{aug}\right) = ResNet50V2\left(f_i^{aug}\right) \tag{8}$$

where the feature map $f_r(f_i^{aug})$ has the shape of $(h_r, w_r, c_r)$, where $h_r$, $w_r$, and $c_r$ are the height, width, and number of channels, respectively.

The final feature map is acquired by concatenating feature maps from MobileNetV2 and ResNet50V2 along the channel dimension. The concatenation operation can be represented as follows:

$$f_c\left(f_i^{aug}\right) = concat\left(f_m\left(f_i^{aug}\right), f_r\left(f_i^{aug}\right)\right) \tag{9}$$

The resulting feature map $f_c(f_i^{aug})$ has the shape of $(h_c, w_c, c_c)$, where $h_c = min(h_m, h_r)$, $w_c = min(w_m, w_r)$, and $c_c = c_m + c_r$.

The concatenated feature map $f_c(f_i^{aug})$ is passed through additional layers to learn higher-level representations and perform classification [34]. Then the flattened layer con-

verts the spatial feature map into a 1D vector. The output of the flattened layer can be represented as follows:

$$v_i = flatten\left(f_c\left(f_i^{aug}\right)\right) \tag{10}$$

where the flattened vector $v_i$ has the shape of $(c_c \times h_c \times w_c)$.

In dense layers for the proposed model, the flattened vector $v_i$ is passed through a series of dense (fully connected) layers.

Each dense layer applies a linear transformation followed by a non-linear activation function like ReLU.

The output of the $j$-th dense layer can be represented as follows:

$$d_j(v_i) = ReLU\left(W_j \bullet v_i + b_j\right) \tag{11}$$

Here, $W_j$ and $b_j$ are the learnable weights and biases of the $j$-th dense layer, respectively.

Dropout is implemented for regularization where some of the neurons/synapses are omitted randomly during the training process in a given iteration to overcome overfitting. The output of the dropout layer can be represented as follows:

$$d_j^{drop}(v_i) = dropout\left(d_j(v_i),\ rate\right) \tag{12}$$

where rate is the dropout rate, specifying the fraction of units to be dropped.

In an output layer, after the final dense layer, the output goes through a sigmoid activation function in order to make a binary prediction, as follows:

$$p(y_i|f_i^{aug}) = sigmoid\left(W_{out} \bullet d_{final}(v_i) + b_{out}\right) = \sigma\left(W_{out} \bullet d_{final}(v_i) + b_{out}\right) \tag{13}$$

where $W_{out}$ and $b_{out}$ are the learnable weights and bias of the output layer, respectively. The sigmoid function is defined as as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{14}$$

During the training process, we trained the proposed SynthSecureNet model with the binary cross-entropy loss function, which is computed according to the predicted probabilities and the true labels. The loss for a single frame $f_i^{aug}$ can be computed as follows:

$$L(y_i,\ p(y_i|f_i^{aug})) = -[y_i \bullet log(p(y_i|f_i^{aug})) + (1 - y_i) \bullet log(1 - p(y_i|f_i^{aug}))] \tag{15}$$

where the $y_i \bullet log(p(y_i|f_i^{aug}))$ term is large (more negative) when the true label $y_i$ is 1 and the predicted probability $p(y_i|f_i^{aug})$ is close to 0.

Additionally, the $(1 - y_i) \bullet log(1 - p(y_i|f_i^{aug}))$ term is large (more negative) when the true label $y_i$ is 0 and the predicted probability $1 - p(y_i|f_i^{aug})$ is close to 0. The overall loss $L(y_i,\ p(y_i|f_i^{aug}))$ is the sum of these two terms, ensuring that the model is penalized for making incorrect predictions, and the penalty increases as the prediction confidence in the wrong direction increases for the proposed model.

The overall loss for the entire video is the average of the frame-level losses as shown in the above equation.

$$L = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\ p(y_i|f_i^{aug})) \tag{16}$$

The weights for the proposed model were updated using the Adam optimizer, which computed adaptive learning rates for each parameter based on the first and second moments of the gradients [40,42].

In the inference stage, the preprocessed frames of a video are passed through the trained SynthSecureNet model for obtaining frame-level predictions. The final prediction for each video is initialized on top of the frame-level predictions by majority averaging.

### *2.4. Ensemble Model (Benefits)*

Ensemble models are preferred in many machine learning tasks compared with individual models. Due to their performance that surpasses the performance of an individual model, the benefits of using ensemble models are briefly discussed in the paragraphs below.

#### 2.4.1. Better Predictive Performance

Ensemble models have the capacity to outperform individual models in terms of prediction. Ensembling generally decreases inaccuracy compared with utilizing a single model for making predictions. Ensemble models achieve this by gaining and leveraging the optimal characteristics of each individual model, reducing the flaws of each model [43–45].

#### 2.4.2. Less Overfitting

Overfitting arises when a model acquires the ability to perfectly match the random fluctuations included in the training data, leading to inadequate performance when applied to new, unseen data. Ensemble methods can mitigate overfitting by combining the predictions of numerous models in a way that achieves a more accurate final prediction. Each individual model may exhibit its own method of overfitting the training data. However, when the predictions of the ensemble are combined, the ensemble demonstrates superior generalization performance. Ensembling, which combines multiple models together, effectively mitigates any overfitting tendencies exhibited by individual models [46].

To address overfitting, dropout layers were incorporated into the SynthSecureNet architecture. This technique randomly drops units during training, helping to generalize the model to unseen data. However, other advanced methods for preventing overfitting, such as early stopping and advanced regularization techniques (e.g., L2 regularization, weight decay), were not utilized in this study. Additionally, ensemble diversity metrics, which measure the uniqueness and complementarity of the models combined in an ensemble, could further enhance the robustness of SynthSecureNet. These methods will be explored in future work to improve the model's performance and generalization.

#### 2.4.3. Less Increased Robustness

Ensemble models have demonstrated greater resilience to noise, outliers, and fluctuations in the data as compared with individual models. While individual models may be susceptible to errors or highly influenced by certain data points, the ensemble can nevertheless yield meaningful predictions when the majority (or all) of the models demonstrate strong performance. The predictions generated by the ensemble are less likely to be as vulnerable as those generated through an individual model [47].

#### 2.4.4. Dealing with Complex Relationships

Ensemble models have the ability to acquire intricate linkages and patterns within the data. The aspects or collections of the data that one individual model accurately represents may differ from those of the others. By combining these models, the ensemble can impart a superior and more intricate depiction of the authentic patterns [48]. Ensemble approaches address non-linear linkages, interactions, and dependencies that individual models may find challenging [49–51].

#### 2.4.5. Bias-Variance Reduction

Ensemble models help mitigate the errors caused by bias and variance in deep learning, resulting in bias-variance reduction. Bias refers to the discrepancy caused by approximating a complicated real-world situation with a simplified model, whereas variance pertains to the model's susceptibility to minor fluctuations in the training data. Ensemble approaches, like bagging and boosting, mitigate bias by employing numerous weak learners and generate a superior and more easily interpretable model [22,52]. Such averaging can aid in lowering prediction variance, resulting in smoother predictions (that is, some models may have more fluctuating predictions, while other models may have more stabilizing predictions with the two impacts canceling each other out).

#### 2.4.6. Dealing with Data Diversity

Ensemble models offer advantages when working with varied or dissimilar data. Various unique models can be trained on distinct subsets or representations of the data, capturing diverse elements or properties. By integrating these models, the ensemble can take advantage of the variety of information and generate more enlightened predictions. Ensemble approaches are capable of managing data that have varying scales, distributions, or modalities. This is because each individual model within the ensemble can focus on a single property of the data [53].

#### 2.4.7. Flexibility and Modularity

Ensemble models possess a high degree of flexibility and modularity in their design and application. The models may be the result of the creation of a hybrid of algorithm models, such as decision trees, neural networks, or SVMs. This provides the ability to combine expertise in a certain field and utilize the advantages of unique algorithms. They may be readily expanded with more ensemble techniques as they develop and diverge, allowing for the modular inclusion and removal of individual models to accommodate new requirements or data [47].

#### 2.4.8. Parallel Processing

Ensemble models may benefit from the parallel processing capabilities that are accessible. Training and prediction for individual models can be performed independently and simultaneously utilizing several processors or a distributed computational resource. This can lead to reduced training time and improved scalability, particularly when dealing with huge datasets or models [18].

#### 2.4.9. Interpretability and Insights

Ensemble models lack interpretability on their own but can offer insights by assessing the influence and importance of individual models [23]. By analyzing the models or features that have the most impact on the predictions of the ensemble, analysts can gain a deeper understanding of the fundamental patterns and relationships within the data. Ensemble methods can be combined with interpretable models to provide the advantages of both high performance and interpretability [54].

#### 2.4.10. Robustness to Concept Drift

Ensemble models have greater resilience to idea drift, which refers to the alteration of the statistical characteristics of the target variable over time. Ensemble approaches can effectively handle changes in data distributions by incorporating models trained on different time periods or utilizing alternative algorithms. Unlike fixed models that require periodic tuning, ensemble models can be updated dynamically. They can also be weighted

to prioritize the most essential models based on incoming data, making them robust in non-stationary situations [51,55].

### *2.5. Block Diagram*

In the process of synthesizing the required architecture, transfer learning was performed, which incorporated MobileNetV2 and ResNet50V2 models as feature extractors. Transfer learning is the idea of using features learned in one task to another different but related task, which helps the model to obtain benefits of pre-learned representations, and thus, the training from scratch is not always necessary. Figure 1 illustrates the development process for the proposed architecture from the preprocessing phase up to inferencing, as shown in Figure 1 below.

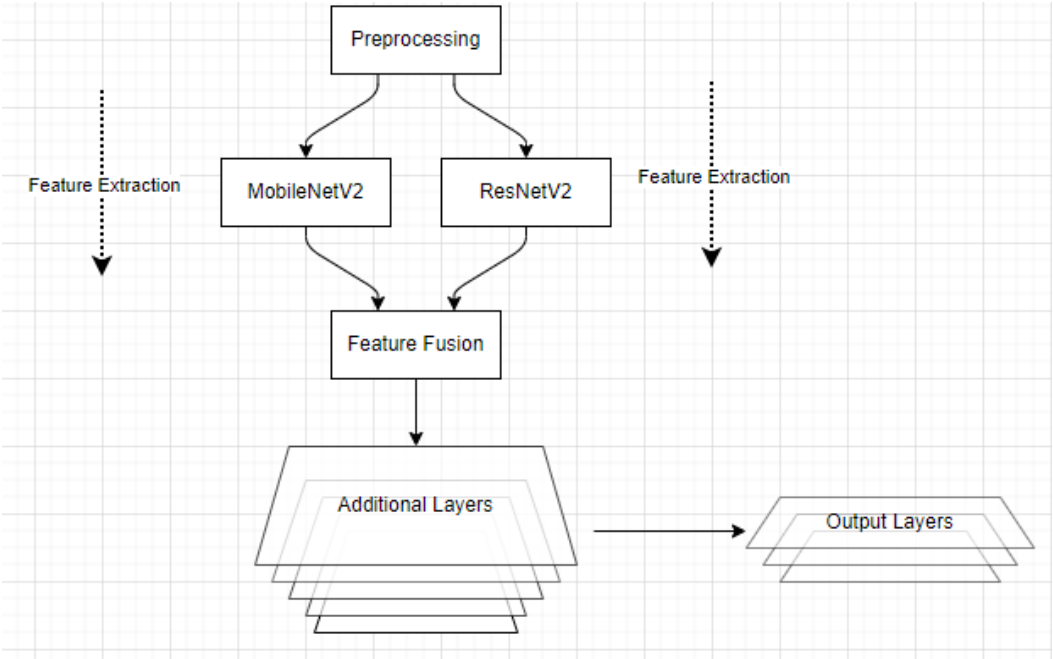



**Figure 1.** Block diagram of the SynthSecureNet architecture.

The algorithm used in SynthSecureNet is summarized in the above block diagram. From the input videos in the preprocessing phase, frames are extracted and augmented further with techniques that include flipping randomly, zooming randomly, adjusting the brightness, and rotating randomly to make the training dataset more diverse. During the feature extraction, the preprocessed frames are then analyzed using MobileNetV2 and ResNet50V2 models, respectively, which are used for feature extraction. MobileNetV2 and ResNet50V2 are convolutional neural networks (CNNs) that have been pre-trained on large-scale image databases including ImageNet and are capable of recognizing quality features within Images. To achieve this, the last convolutional layers of both MobileNetV2 and ResNet50V2 architectures are fed with the input frames to generate their feature maps, which preserves more high-level representations. In this last stage of the network (feature fusion phase), MobileNetV2 and ResNet50V2 feature maps are concatenated along with the channel dimension. Temporal concatenation integrates the two models by merging the feature maps along the channel axis, outputting a fused feature map that includes information from the two models. This fusion step enables the subsequent layers to make use of the complementary features being learned by both models.

In the Additional Layers phase, as shown in Figure 1 above, the fused feature map is for further processing of getting features and making predictions. These layers typically include a flattened layer that flattens the fused feature map obtained from the 3D convolu-

tion into a one-dimensional vector. Dense layers are fully connected layers that specifically calculate complex combinations of features that are extracted out of the flattened structure. Dropout layers come into play during the training process, where a certain part of the units is randomly removed in order to reduce overfitting by the use of this technique.

In the output layer phase, the final dense layer employs the sigmoid activation function in classifying the inputs as a binary output to signify the presence or the absence of violence in the input frame.

During the training phase, the binary cross-entropy function is used as the loss function for the SynthSecureNet model as it calculates the difference between the probability estimates produced by the model and the ground truth. The model is trained using the Adam optimizer because it allows for the learning rate to be updated for each parameter on the basis of its gradient history. As a post-processing of the individual frames of a video during inference, the preprocessed frames are input to the SynthSecureNet model for frame-level classification. To this end, the frame-level predictions are then combined based on methods such as majority voting or averaging to arrive at the final video-level prediction.

Furthermore, together with transfer learning, feature fusion, and additional layers, the proposed SynthSecureNet was designed to learn and classify violence in videos. The pre-trained models, when taken as feature extractors, are already powerful, as they employ and build from prior experience with large datasets of images. Incorporating features of multiple models promotes the models' ability to focus on different and complementary aspects. These added layers are used in order to gain feature combinations and use them in training the model to learn accurate patterns and make precise predictions based on the fused features.

### *2.6. Model Training Dataset*

The model was trained using the Real-Life Violence Situations Dataset during the training procedure. The data comprise clips that were captured in real-world scenarios, such as sports, as well as CCTV footage. The dataset is accessible online and is available for use [56]. The dataset comprises 2000 videos, of which 1000 videos depict violent activities, and the remaining 1000 videos depict non-violent activities. Only 70% of the videos from the Real-Life Violence Situations Dataset were allocated to the training set. Only 15% of the videos from the Real-Life Violence Situations Dataset were allocated to the validation set. The testing set was allocated 15% of the videos from the Real-Life Violence Situations Dataset.

The Real-Life Violence Situations Dataset, used for training and testing in this study, provides a balanced representation of violent and non-violent activities in real-world scenarios. However, it does not encompass all possible contexts of violence, such as variations in cultural environments, lighting conditions (e.g., low-light settings). These factors may significantly influence the performance of violence detection models in diverse applications.

In addition to the Real-Life Violence Situations Dataset, other datasets such as the Surveillance Camera Fight dataset [57], RWF-2000 [58], and Bus Violence dataset [59] offer diverse scenarios and environmental conditions. While these datasets could provide additional insights into violence detection across varied contexts, their inclusion was outside the scope of this study due to resource constraints and the focus on establishing the efficacy of SynthSecureNet on the Real-Life Violence Situations Dataset. Future work will incorporate these datasets to further evaluate and generalize the proposed approach.

## 3. Experimental Results

In this section, the performances of three architectures, namely, SynthSecureNet, ResNetV2, and MobileNetV2 Architectures, are compared.

### *3.1. Performance Metrics*

The performances of these architectures are evaluated using five parameters, which include accuracy, recall, F1 score, precision, and training duration. The confusion matrix was also used and produced for each performance evaluation. The evaluation was in terms of accuracy, precision, recall, and F1 score [60]. Accuracy was used to evaluate overall correctness for a model. These factors that were used to evaluate performance are calculated using TP (true positive), FP (false positive), FN (false negative), and TN (true negative). Violence is a binary classification task where an event or human activity is determined as violent (positive class) or nonviolent (negative class) [61].

Accuracy is a measure of how many predictions are correct out of the total predictions. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{17}$$

Accuracy provides a general idea of the correctness of the trained model. Precision, commonly referred to as Positive Predictive Value, indicates the percentage of the model's positive predictions that turn out to be true. The formula to generate the precision is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

Recall, which is a measure of how many actual positive cases the model accurately predicted, is sometimes referred to as Sensitivity or True Positive Rate. It is computed as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

The harmonic mean of recall and precision is known as F1 score. It offers a fair assessment that considers both false positives and false negatives. The formula for calculating the F1 score is as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

Recall and precision are weighed equally in the F1 score [60].

In summary, accuracy gives an overall idea of how correct the model is during classification. Precision focuses on minimizing false positives. Recall focuses on minimizing false negatives. The F1 score balances precision and recall.

### *3.2. Experimental Procedure*

The performance evaluation was performed by changing the size of the dataset that was used to train all the models. The datasets that were used to evaluate the all performances of the three models consist of a 400-video dataset, 800-video dataset, 1200-video dataset, 1600-video dataset, and 2000-video dataset, respectively. Across all sub-dataset sizes, the sets were allocated as follows: 70% for training set, 15% for validation set, and 15% for testing set.

All these datasets were extracted from one dataset that consists of 2000 videos in total. The dataset consists of two classes with 1000 videos each; the first class consists of violence videos, and the second class constitutes nonviolence videos. Sub-datasets were extracted evenly from both classes (e.g., 1200-video dataset = 600 violence videos + 600 nonviolence videos) of which 70%, 15%, and 15% were allocated to the training set, validation set, testing set, respectively. An online open and free programming platform called Kaggle Notebook (Python language) was used when developing the proposed new architecture and generating results for performance evaluation.

### *3.3. Performance Evaluation*

#### 3.3.1. Dataset of 400 Videos

In Table 1, a 400-video dataset was used to train, validate, and test the models. It can be observed in the table that SynthSecureNet outperformed both MobileNetV2 and ResNet50V2. Results in Table 1 were obtained during the testing phase, and the training duration that was observed during the training phase is also included in Table 1.

**Table 1.** Performance comparison using the 400-video dataset.

| Model | F1 Score | Recall | Precision | Accuracy | Training Duration |
|---|---|---|---|---|---|
| SynthSecureNet | 96% | 96% | 96% | 96% | 15 min |
| ResNetV2 | 91% | 88% | 89% | 89% | 9 min |
| MobileNetV2 | 89% | 89% | 88% | 88% | 5 min |

The SynthSecureNet architecture achieved an F1 score of 96%, a recall of 96%, a precision of 96%, and an accuracy of 96%, demonstrating a high level of performance for this model in violence detection. On the other hand, ResNetV2 achieved an F1 score of 91%, while MobileNetV2 achieved an F1 score of 89%, which shows that ResNetV2 outperformed MobileNetV2.

The training duration of the SynthSecureNet model (15 min) took longer than that of ResNet50V2 (9 min) and MobileNetV2 (5 min). When compared with both MobileNetV2 and ResNet50V2, the SynthSecureNet model's high performance came at a trade-off of a small increase in training period.

Figure 2 shows the confusion matrices obtained after testing. SynthSecureNet made 2253 correct predictions and only 53 wrong predictions, demonstrating its effectiveness in accurately classifying violent and non-violent videos.

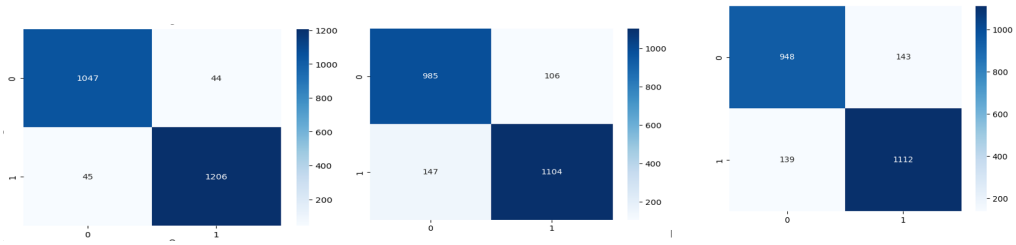



**Figure 2.** Confusion matrices for SynthSecureNet, ResNetV2, and MobileNetV2 (400-video dataset).

#### 3.3.2. Dataset of 800 Videos

The models were trained, validated, and tested using a dataset of 800 videos, as shown in Table 2. SynthSecureNet performed better than both MobileNetV2 and ResNet50V2. The results in Table 2 were obtained during the testing phase, and the training duration that was observed during the training phase is also included in Table 2.

**Table 2.** Performance comparison using the 800-video dataset.

| Model | F1 Score | Recall | Precision | Accuracy | Training Duration |
|---|---|---|---|---|---|
| SynthSecureNet | 98% | 98% | 98% | 98% | 30 min |
| ResNetV2 | 90% | 88% | 92% | 89% | 14 min |
| MobileNetV2 | 89% | 89% | 89% | 89% | 9 min |

The SynthSecureNet architecture demonstrated a high degree of performance in violence detection with an F1 score of 98%, a recall of 98%, a precision of 98%, and an accuracy of 98%. However, ResNetV2 outperformed MobileNetV2 with an F1 score of 90% compared with 89% for MobileNetV2.

The training duration of the SynthSecureNet model, which lasted for 30 min, exceeded that of the ResNet50V2 model, which took 14 min, and the MobileNetV2 model, which took 9 min.

Figure 3 displays the confusion matrices acquired throughout the testing process. SynthSecureNet achieved a high accuracy in classifying between violent and non-violent activities, with 4567 right predictions and just 110 incorrect predictions.

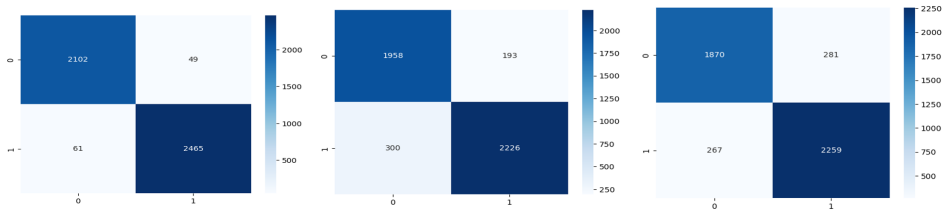



**Figure 3.** Confusion matrices for SynthSecureNet, ResNetV2, and MobileNetV2 (800-video dataset).

### 3.3.3. Dataset of 1200 Videos

In Table 3, the 1200-video dataset was used to train, validate, and test the models. It can be observed that SynthSecureNet outperformed both MobileNetV2 and ResNet50V2. The results in Table 3 were obtained during the testing phase, and the training duration that was observed during the training phase is also included in Table 3.

**Table 3.** Performance comparison using the 1200-video dataset.

| Model | F1 Score | Recall | Precision | Accuracy | Training Duration |
|---|---|---|---|---|---|
| SynthSecureNet | 98.31% | 98.25% | 98.37% | 98.15% | 46 min |
| ResNetV2 | 91% | 89% | 90% | 90% | 26 min |
| MobileNetV2 | 91% | 92% | 90% | 90% | 19 min |

The SynthSecureNet architecture achieved an F1 score of 98.31%, a recall of 98.25%, a precision of 98.37%, and an accuracy of 98.15%, demonstrating a high level of performance for this model in violence detection. On the other hand, both ResNetV2 and MobileNetV2 achieved an F1 score of 91%, indicating that the level of reliability of both models is the same.

The training duration of the SynthSecureNet model (46 min) took longer than that of ResNet50V2 (26 min) and MobileNetV2 (19 min).

Figure 4 shows the confusion matrices obtained in the testing process. SynthSecureNet made 7374 correct predictions and only 139 wrong predictions, demonstrating its effectiveness in accurately classifying violent and non-violent activities.

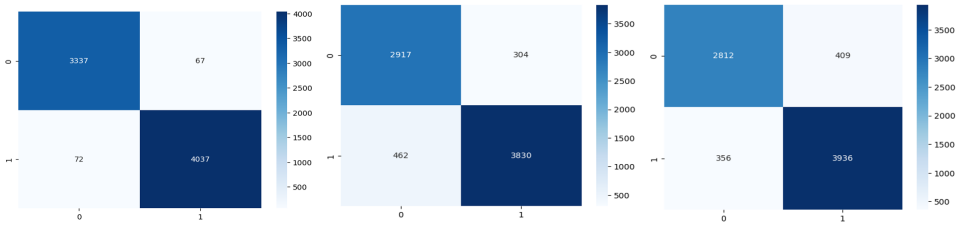



**Figure 4.** Confusion matrices for SynthSecureNet, ResNetV2, and MobileNetV2 (1200-video dataset).

### 3.3.4. Dataset of 1600 Videos

The models were trained, validated, and tested using a 1600-video dataset, as indicated in Table 4. The table illustrates that SynthSecureNet outperformed both ResNet50V2 and MobileNetV2. The results in Table 4 were obtained during the testing phase, and the training duration that was observed during the training phase is also included in Table 4.

**Table 4.** Performance comparison using the 1600-video dataset.

| Model | F1 Score | Recall | Precision | Accuracy | Training Duration |
|---|---|---|---|---|---|
| SynthSecureNet | 98.88% | 98.89% | 98.87% | 98.77% | 1 h 40 min |
| ResNetV2 | 91% | 90% | 91% | 91% | 30 min |
| MobileNetV2 | 91% | 91% | 90% | 90% | 1 h 11 min |

The SynthSecureNet architecture demonstrated a high level of performance in the detection of violence, as demonstrated by its F1 score of 98.88%, recall of 98.89%, precision of 98.87%, and accuracy of 98.77%. However, ResNet50V2 and MobileNetV2 both achieved an F1 score of 91%, indicating that their reliability levels are equivalent.

The SynthSecureNet model required an extended training duration (1 h and 40 min) compared with ResNet50V2 (30 min) and MobileNetV2 (1 h and 11 min).

Figure 5 illustrates the confusion matrices acquired during the testing process. The effectiveness of SynthSecureNet in accurately classifying violent and non-violent activities was demonstrated by its 9896 correct predictions and 123 incorrect predictions.

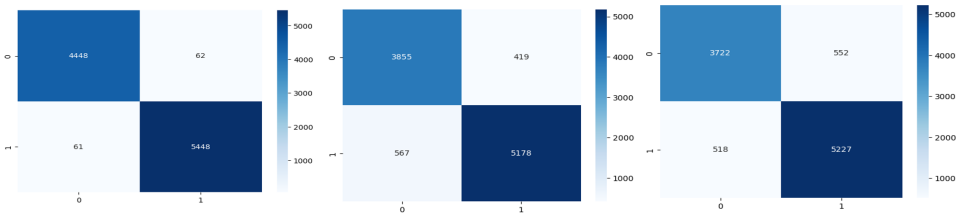



**Figure 5.** Confusion matrices for SynthSecureNet, ResNetV2, and MobileNetV2 (1600-video dataset).

### 3.3.5. Dataset of 2000 Videos

In Table 5, the 2000-video dataset was used to train, validate, and test the models. It can again be observed that SynthSecureNet outperformed both MobileNetV2 and ResNet50V2. The results in Table 5 were obtained during the testing phase, and the training duration that was observed during the training phase is also included in Table 5.

**Table 5.** Performance comparison using the 2000-video dataset.

| Model | F1 Score | Recall | Precision | Accuracy | Training Duration |
|---|---|---|---|---|---|
| SynthSecureNet | 99.24% | 99.23% | 99.26% | 99.22% | 6 h 5 min |
| ResNetV2 | 94% | 95% | 94% | 94% | 3 h 9 min |
| MobileNetV2 | 91% | 91% | 90% | 90% | 2 h 30 min |

The SynthSecureNet architecture achieved an F1 score of 99.24%, a recall of 99.23%, a precision of 99.26%, and an accuracy of 99.22%. On the other hand, ResNet50V2 achieved an F1 score of 94%, while MobileNetV2 achieved an F1 score of 91%, which means that ResNet50V2 is more reliable when compared with MobileNetV2.

The training duration of the SynthSecureNet model (6 h and 5 min) took longer than that of ResNet50V2 (3 h and 9 min) and MobileNetV2 (2 h and 30 min).

Figure 6 shows the confusion matrices obtained in the testing process of all three models. SynthSecureNet made 12,454 correct predictions and only 98 wrong predictions, demonstrating its effectiveness in accurately classifying violent and non-violent activities.

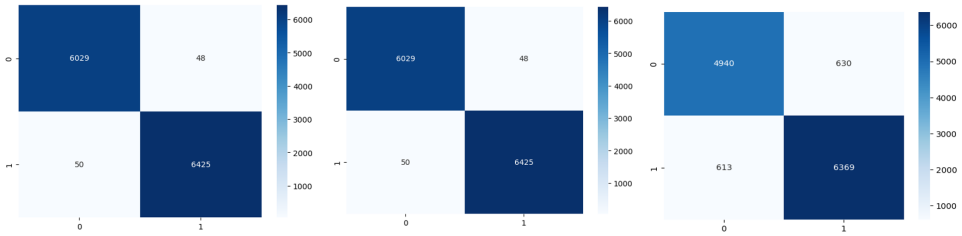



**Figure 6.** Confusion matrices for SynthSecureNet, ResNetV2, and MobileNetV2 (2000-video dataset).

### *3.4. Summary*

In conclusion, it can be observed that the proposed SynthSecureNet architecture yields better results as compared with ResNet50V2 and MobileNetV2 in the task of violence detection from videos. SynthSecureNet achieved higher F1 scores, recall rate, precision level, and accuracy in all the sizes of the datasets considered, indicating its efficiency in classifying violence accurately. Despite the fact that the training times were longer than those of the other models, SynthSecureNet presented excellent results that seem to be justified by extra computational time.

The analysis of the evaluation results shows that SynthSecureNet is a promising violence detection system and can be considered rather effective and accurate. Therefore, the capacity for scaling big data and maintaining high performance makes it a suitable solution for actual use. The viability of the architecture reflects its durability and adaptability, which may make it appropriate for use in different areas with violence identification as a priority, for example, video surveillance, content filtering, safety, and security.

The consistent superior performance of SynthSecureNet across various dataset sizes demonstrates its robustness and scalability. This is particularly important in real-world applications where the volume of data can vary significantly. The architecture's ability to maintain high accuracy levels even with larger datasets suggests that it could be effectively deployed in large-scale surveillance systems without compromising on detection quality.

Moreover, the trade-off between increased training time and improved performance is a crucial consideration. While SynthSecureNet requires more computational resources during the training phase, the enhanced accuracy in violence detection could lead to more efficient and effective real-time monitoring systems. This could potentially result in faster response times to violent incidents and improved public safety outcomes.

The flexibility of SynthSecureNet, as evidenced by its performance across different dataset sizes, also indicates its potential for adaptation to various contexts and scenarios. This versatility could be valuable in different sectors such as law enforcement, private security, and content moderation platforms, where the nature and prevalence of violent content may vary.

In summary, the experimental results strongly support the efficacy of the SynthSecureNet architecture for violence detection tasks. Its consistent outperformance of established models like ResNet50V2 and MobileNetV2 positions it as a promising advancement in the field of automated violence detection. Future work could focus on further optimizing the training process to reduce computational time while maintaining or even improving the current high levels of accuracy.

## 4. Conclusions

In this study, a new architecture named SynthSecureNet, was proposed as a feature-level fused model of the MobileNetV2 and ResNet50V2 architectures, which outperforms the individual models across various datasets independently. The results of the comprehensive evaluation on datasets containing 400 to 2000 videos show that our proposed model improves the performance of the existing ones.

The applications of the proposed architecture are relevant to multiple domains, such as video surveillance, content moderation, and public safety or security settings. The high accuracy and reliability of SynthSecureNet can help in the automation of violence detection, which can minimize the load of moderators and help in taking necessary precautions.

In this paper, feature-level fusion is also highlighted as the most suitable method to improve violence detection performance in the study. Hence, SynthSecureNet integrates parts of the models from both MobileNetV2 and ResNet50V2, which in turn enhances its predictive accuracy and generalization capability.

Future work can include exploring possibilities of minimizing the training duration and computational resource consumption.

**Author Contributions:** Conceptualization, N.Z. and P.O.; methodology, N.Z. and P.B; software, N.Z.; validation, P.O. and P.B.; writing—original draft preparation, N.Z.; writing—review and editing, N.Z., P.O. and P.B.; visualization, N.Z.; supervision, P.O. and P.B.; project administration, P.O. and P.B.; funding acquisition, P.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** No new datasets were created. All relevant results data are available in the article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Inbavalli, A.; Jarshini, T.; Muralikrishnaa, M. Efficient Aggressive Behaviour Detection And Alert System Employing Deep Learning Techniques. In Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM), Dehradun, India, 14–16 March 2024; IEEE: Piscataway, HJ, USA, 2024; pp. 404–410.
2. Tyagi, S.; Tyagi, S.; Bagga, V.; Bansal, S.; Goswami, S.; Verma, Y.; Tyagi, M.; Singh, S.P. Deep learning solution for real-time violence detection in video streams. In *Advances in AI for Biomedical Instrumentation, Electronics and Computing*; CRC Press: Boca Raton, FL, USA, 2024; pp. 142–146.
3. Thomas, M.; Balamurugan, P. Real-Time Violence Detection and Alert System using MobileNetV2 and Cloud Firestore. In Proceedings of the 2024 2nd International Conference on Networking and Communications (ICNWC), Chennai, India, 2–4 April 2024; IEEE: Piscataway, HJ, USA, 2024; pp. 1–9.
4. Maheswari, G.; Balaji M.U.A.; Asik, A.; Adams, S.R.; Thanikavel, B. Public Space Real-Time Violence Detection and Notifier. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 28–29 March 2024; IEEE: Piscataway, HJ, USA, 2024; pp. 1–6.
5. Huillcen Baca, H.A.; Palomino Valdivia, F.d.L.; Gutierrez Caceres, J.C. Efficient human violence recognition for surveillance in real time. *Sensors* **2024**, *24*, 668. [CrossRef] [PubMed]
6. Arroyo, R.; Yebes, J.J.; Bergasa, L.M.; Daza, I.G.; Almazán, J. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Syst. Appl.* **2015**, *42*, 7991–8005. [CrossRef]
7. Park, J.H.; Mahmoud, M.; Kang, H.S. Conv3D-based video violence detection network using optical flow and RGB data. *Sensors* **2024**, *24*, 317. [CrossRef] [PubMed]
8. Gautam, V.; Maheshwari, H.; Tiwari, R.G.; Agarwal, A.K.; Trivedi, N.K. Federated Learning Empowered Violence Recognition in CCTV Footage: A YOLO and ResNet-50 Fusion Approach. In Proceedings of the 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 4–6 January 2024; IEEE: Piscataway, HJ, USA, 2024; pp. 01–06.
9. Mumtaz, A.; Sargano, A.B.; Habib, Z. Violence detection in surveillance videos with deep network using transfer learning. In Proceedings of the 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 20–22 December 2018; IEEE: Piscataway, HJ, USA, 2018; pp. 558–563.
10. Akinci, G.M. The purposes and meanings of surveillance: A case study in a shopping mall in Ankara, Turkey. *Secur. J.* **2015**, *28*, 39–53. [CrossRef]
11. Lee, H.E.; Ermakova, T.; Ververis, V.; Fabian, B. Detecting child sexual abuse material: A comprehensive survey. *Forensic Sci. Int. Digit. Investig.* **2020**, *34*, 301022. [CrossRef]
12. Ballard, C.; Brady, L. Violence Prevention in Georgia's Rural Public School Systems: A Comparison of Perceptions of School Superintendents 1995–2005. *J. Sch. Violence* **2007**, *6*, 105–129. [CrossRef]

13. Honarjoo, N.; Abdari, A.; Mansouri, A. Violence detection using pre-trained models. In Proceedings of the 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), Kashan, Iran, 3–4 March 2021; IEEE: Piscataway, HJ, USA, 2021; pp. 1–4.
14. Negre, P.; Alonso, R.S.; González-Briones, A.; Prieto, J.; Rodríguez-González, S. Literature Review of Deep-Learning-Based Detection of Violence in Video. *Sensors* **2024**, *24*, 4016. [CrossRef]
15. Haque, M.; Nyeem, H.; Afsha, S. BrutNet: A novel approach for violence detection and classification using DCNN with GRU. *J. Eng.* **2024**, *2024*, e12375. [CrossRef]
16. Contardo, P.; Tomassini, S.; Falcionelli, N.; Dragoni, A.F.; Sernani, P. Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In Proceedings of the 5th International Conference on Recent Trends and Applications in Computer Science and Information Technology, RTA-CSIT 2023, Tirana, Albania, 26–27 April 2023.
17. Zahid, Y.; Tahir, M.A.; Durrani, N.M.; Bouridane, A. IBaggedFCNet: An Ensemble Framework for Anomaly Detection in Surveillance Videos. *IEEE Access* **2020**, *8*, 220620–220630. [CrossRef]
18. Young, S.; Abdou, T.; Bener, A. Deep Super Learner: A Deep Ensemble for Classification Problems. In *Advances in Artificial Intelligence*; Bagheri, E., Cheung, J.C., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 10832, pp. 84–95. [CrossRef]
19. Jayaswal, R.; Dixit, M. A Framework for Anomaly Classification Using Deep Transfer Learning Approach. *Rev. d'Intell. Artif.* **2021**, *35*, 255–263. [CrossRef]
20. Kumar, V.; Recupero, D.R.; Riboni, D.; Helaoui, R. Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes. *IEEE Access* **2021**, *9*, 7107–7126. [CrossRef]
21. Tariq, U.; Lin, K.H.; Li, Z.; Zhou, X.; Wang, Z.; Le, V.; Huang, T.S.; Lv, X.; Han, T.X. Recognizing Emotions from an Ensemble of Features. *IEEE Trans. Syst. Man Cybern. Part (Cybern.)* **2012**, *42*, 1017–1026. [CrossRef]
22. Zheng, J.; Cao, X.; Zhang, B.; Zhen, X.; Su, X. Deep Ensemble Machine for Video Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 553–565. [CrossRef]
23. Shripriya, C.; Akshaya, J.; Sowmya, R.; Poonkodi, M. Violence Detection System Using Resnet. In Proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, Tamil Nadu, 2–4 December 2021; IEEE: Piscataway, HJ, USA, 2021; pp. 1069–1072.
24. Jain, M.; Kumar, M. A Review of Violence Detection Techniques. In Proceedings of the 2024 2nd International Conference on Computer, Communication and Control (IC4), Indore, India, 8–10 February 2024; IEEE: Piscataway, HJ, USA, 2024; pp. 1–6.
25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
26. Koonce, B. ResNet 50. In *Convolutional Neural Networks with Swift for Tensorflow*; Apress: Berkeley, CA, USA, 2021; pp. 63–72. [CrossRef]
27. Jain, A.; Vishwakarma, D.K. State-of-the-arts Violence Detection using ConvNets. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 28–30 July 2020; IEEE: Piscataway, HJ, USA, 2020; pp. 0813–0817.
28. Bermeo, M.; Morocho-Cayamcela, M.E.; Cuenca, E. Unraveling the Power of 4D Residual Blocks and Transfer Learning in Violence Detection. In *Information and Communication Technologies*; Maldonado-Mahauad, J., Herrera-Tapia, J., Zambrano-Martínez, J.L., Berrezueta, S., Eds.; Communications in Computer and Information Science; Springer Nature: Cham, Switzerland, 2023; Volume 1885, pp. 207–219. [CrossRef]
29. Jianjie, S.; Weijun, Z. Violence detection based on three-dimensional convolutional neural network with inception-ResNet. In Proceedings of the 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 11–13 December 2020; IEEE: Piscataway, HJ, USA, 2020; pp. 145–150.
30. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
31. Uemura, T.; Näppi, J.J.; Hironaka, T.; Kim, H.; Yoshida, H. Comparative performance of 3D-DenseNet, 3D-ResNet, and 3D-VGG models in polyp detection for CT colonography. In Proceedings of the Medical Imaging 2020: Computer-Aided Diagnosis, Houston, TX, USA, 16–19 February 2020; SPIE: Bellingham, WA, USA, 2020; Volume 11314, pp. 736–741.
32. Dubey, S.; Boragule, A.; Jeon, M. 3d resnet with ranking loss function for abnormal activity detection in videos. In Proceedings of the 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), Chengdu, China, 23–26 October 2019; IEEE: Piscataway, HJ, USA, 2019; pp. 1–6.
33. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
34. Nguyen, L.D.; Gao, R.; Lin, D.; Lin, Z. Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 15455–15467. [CrossRef]

35. Traoré, A.; Akhloufi, M.A. Violence detection in videos using deep recurrent and convolutional neural networks. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; IEEE: Piscataway, HJ, USA, 2020; pp. 154–159.
36. Nguyen, L.D.; Lin, D.; Lin, Z.; Cao, J. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 27–30 May 2018; IEEE: Piscataway, HJ, USA, 2018; pp. 1–5.
37. Che, W.; Liu, Y.; Wang, Y.; Zheng, B.; Liu, T. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. *arXiv* **2018**, arXiv:1807.03121.
38. Lu, Y.; Huang, X.; Huang, Y.; Liu, D. Sigmoid function model for a PFM power electronic converter. *IEEE Trans. Power Electron.* **2019**, *35*, 4233–4241. [CrossRef]
39. Qin, Y.; Wang, X.; Zou, J. The optimized deep belief networks with improved logistic sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines. *IEEE Trans. Ind. Electron.* **2018**, *66*, 3814–3824. [CrossRef]
40. Zhang, Z. Improved adam optimizer for deep neural networks. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada, 4–6 June 2018; IEEE: Piscataway, HJ, USA, 2018; pp. 1–2.
41. Jayasimhan, A.; Pabitha, P. A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset. In Proceedings of the 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 15–16 December 2022; IEEE: Piscataway, HJ, USA, 2022; pp. 1–5.
42. Feng, J.; Wang, Z.; Zha, M.; Cao, X. Flower Recognition Based on Transfer Learning and Adam Deep Learning Optimization Algorithm. In Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China, 20–22 September 2019; pp. 598–604. [CrossRef]
43. Das, A.; Hoque, M.M.; Sharif, O.; Dewan, M.A.A.; Siddique, N. TEmoX: Classification of Textual Emotion Using Ensemble of Transformers. *IEEE Access* **2023**, *11*, 109803–109818. [CrossRef]
44. Sarman, S.; Sert, M. Audio based violent scene classification using ensemble learning. In Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey, 22–25 March 2018; IEEE: Piscataway, HJ, USA, 2018; pp. 1–5.
45. Ju, C.; Bibaut, A.; Van Der Laan, M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* **2018**, *45*, 2800–2818. [CrossRef]
46. Kumari, A.; Kaur, T.; Ranjan, P.; Chopra, S.; Sarkar, S.; Baitha, U. Workplace violence against doctors: Characteristics, risk factors, and mitigation strategies. *J. Postgrad. Med.* **2020**, *66*, 149. [CrossRef]
47. Liu, H.; Zhu, X.; Fujii, T. Ensemble deep learning based cooperative spectrum sensing with semi-soft stacking fusion center. In Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakech, Morocco, 15–19 April 2019; IEEE: Piscataway, HJ, USA, 2019; pp. 1–6.
48. Vo-Le, C.; Vo, H.S.; Vu, T.D.; Son, N.H. Violence Detection using Feature Fusion of Optical Flow and 3D CNN on AICS-Violence Dataset. In Proceedings of the 2022 IEEE Ninth International Conference on Communications and Electronics (ICCE), Nha Trang, Vietnam, 27–29 July 2022; IEEE: Piscataway, HJ, USA, 2022; pp. 395–399.
49. Lv, Q. Classification of Grapevine Leaf Images with Deep Learning Ensemble Models. In Proceedings of the 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 12–14 May 2023; pp. 191–194. [CrossRef]
50. Wang, J.; Hazarika, S.; Li, C.; Shen, H.W. Visualization and visual analysis of ensemble data: A survey. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 2853–2872. [CrossRef] [PubMed]
51. Catolino, G.; Ferrucci, F. Ensemble techniques for software change prediction: A preliminary investigation. In Proceedings of the 2018 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE), Campobasso, Italy, 20 March 2018; IEEE: Piscataway, HJ, USA, 2018; pp. 25–30.
52. Ramzan, M.; Abid, A.; Khan, H.U.; Awan, S.M.; Ismail, A.; Ahmed, M.; Ilyas, M.; Mahmood, A. A review on state-of-the-art violence detection techniques. *IEEE Access* **2019**, *7*, 107560–107575. [CrossRef]
53. Oh, H. A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model. *IEEE Access* **2021**, *9*, 144121–144128. [CrossRef]
54. Naik, A.J.; Gopalakrishna, M.T. Deep-violence: individual person violent activity detection in video. *Multimed. Tools Appl.* **2021**, *80*, 18365–18380. [CrossRef]
55. Mohandes, M.; Deriche, M.; Aliyu, S.O. Classifiers combination techniques: A comprehensive review. *IEEE Access* **2018**, *6*, 19626–19639. [CrossRef]
56. Soliman.; Nashed, K.; Mostafa, C.K. Real Life Violence Situations Dataset. Available online: https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset (accessed on 29 October 2024).
57. Aktı, Ş.; Tataroğlu, G.A.; Ekenel, H.K. Vision-based fight detection from surveillance cameras. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; IEEE: Piscataway, HJ, USA, 2019; pp. 1–6.

58. Cheng, M.; Cai, K.; Li, M. RWF-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, HJ, USA, 2021; pp. 4183–4190.
59. Ciampi, L.; Foszner, P.; Messina, N.; Staniszewski, M.; Gennaro, C.; Falchi, F.; Serao, G.; Cogiel, M.; Golba, D.; Szczęsna, A. Bus violence: An open benchmark for video violence detection on public transport. *Sensors* **2022**, *22*, 8345. [CrossRef]
60. Fourure, D.; Javaid, M.U.; Posocco, N.; Tihon, S. Anomaly Detection: How to Artificially Increase Your F1-Score with a Biased Evaluation Protocol. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*; Dong, Y., Kourtellis, N., Hammer, B., Lozano, J.A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2021; Volume 12978, pp. 3–18. [CrossRef]
61. Yacouby, R.; Axman, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online, 16 March 2020; pp. 79–91.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.