Expert Systems With Applications 189 (2022) 116114

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Attention based CNN model for fire detection and localization in real-world images

Saima Majid [a], Fayadh Alenezi [b], Sarfaraz Masood [a,*], Musheer Ahmad [a], Emine Selda Gündüz [c], Kemal Polat [d,*]

[a] Department of Computer Engineering, Jamia Millia Islamia, New Delhi 110025, India
[b] Department of Electrical Engineering, College of Engineering, Jouf University, Saudi Arabia
[c] First and Emergency Aid Programme, Akdeniz University Vocational School of Health Services, Antalya, Turkey
[d] Department of Electrical and Electronics Engineering, Bolu Abant Izzet Baysal University, Bolu, Turkey

ARTICLE INFO

Keywords:
Fire detection
CNN
Attention mechanism
Transfer learning
Grad-CAM

ABSTRACT

Fire is a severe natural calamity that causes significant harm to human lives and the environment. Recent works have proposed the use of computer vision for developing a cost-effective automated fire detection system. This paper presents a custom framework for detecting fire using transfer learning with state-of-the-art CNNs trained over real-world fire breakout images. The framework also uses the Grad-CAM method for the visualization and localization of fire in the images. The model also uses an attention mechanism that has significantly assisted the network in achieving better performances. It was observed through Grad-CAM results that the proposed use of attention led the model towards better localization of fire in the images. Among the plethora of models explored, the EfficientNetB0 emerged as the best-suited network choice for the problem. For the selected real-world fire image dataset, a test accuracy of 95.40% strongly supports the model's efficiency in detecting fire from the presented image samples. Also, a very high recall of 97.61 highlights that the model has negligible false negatives, suggesting the network to be reliable for fire detection.

1. Introduction

Fire is a destructive natural disaster causing massive vandalism to both human life and the ecological environment. Fire detection in outdoor environments has become a prime concern as well as a challenging task for the safety of human lives. Several large-scale wildfires and forest fires erupted around the world in 2020, including the Australian bushfires that began in 2019 and lasted until March 2020. In this incident, almost half a billion animals were burnt alive. A similar blaze in the US state of California was also devastating as it killed numerous people ("Wildfires, forest fires around world in, 2020). In recent years, fire detection systems have received a lot of attention and have aided in the protection of people and property from fire threats. Some aspects of fire, such as light, heat, and smoke, can be detected using sensor detection systems (F. Saeed, A. Paul, P. Karthigaikumar, & A. Nayyar, 2019).

To minimize the destruction caused by various accidents of fire, various types of fire detection algorithms have been recently introduced with different technologies (Arpit Jadon, Omama, Ansari, & Sharma, 2019). Conventional fire detection methods use sensors that detect smoke, size of the fire, location of initial flame, atmosphere temperature, etc. (Yin, Wan, Yuan, Xia, and Shi, 2017). These sensors are very popular and have been extensively used as they are low-cost and simple to operate (Yin et al., 2017). However, these system detectors can have few shortcomings such as the late triggering of alarm defeating the purpose of early warning, space coverage, and signal transmission. Excluding the early fire detection problem, existing systems also turn out to be ineffective to the false triggering of the alarm (Jadon et al., 2019). Many alarms are operative in a confined space, thus becoming inefficient for a wide-open space, for instance, for outdoors or public spaces, huge infrastructures such as stadiums, aircraft hangers. Furthermore, almost all the sensors require proximity to the fire/smoke (Jadon et al., 2019). Since the sensors are usually set up in the ceiling, the time taken by smoke to reach up to the ceiling results in delay and thus defeating the purpose of an early warning.

Monitoring fire detection systems based on camera feeds have received significant recognition in computer vision research communities, specifically in convolutional neural networks (CNNs), during the past few years (Li, Chen, Wu, & Liu, 2020). However these fire detection

* Corresponding authorsat: Department of Computer Engineeering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi 110025, India.
E-mail addresses: fshenezi@ju.edu.sa (F. Alenezi), smasood@jmi.ac.in (S. Masood), seldagunduz@akdeniz.edu.tr (E.S. Gündüz), kpolat@ibu.edu.tr (K. Polat).

https://doi.org/10.1016/j.eswa.2021.116114
Received 18 September 2021; Received in revised form 14 October 2021; Accepted 16 October 2021
Available online 26 October 2021

methods still have some challenges to overcome. Many researchers have presented solutions creating proposal regions by choosing features manually. This type of process of creating the proposal regions by determining one after the other ignores the use of CNNs to the overall procedure of detection. Hence this leads to a huge amount of computation overhead and slow detection speed (Li and Zhao, 2020). Generally, studying and exploring the static characteristics of smoke and diverse flame in a vision system is not an easy task as it needs a vast amount of knowledge. Thus, the problem comes with a demand for an effective and stable algorithm for the detection of fire with high accuracy and automated feature selection that can prevent large-scale damage to both human lives and the natural environment (Cao, Yang, Tang, & Lu, 2019).

For the safety of human life, fire detection in outdoor environments has become a major worry as well as a difficult task. Fires pose a severe hazard to industries, crowded events, and highly populated locations around the world, according to statistics. These kinds of incidents can destroy property, harm the environment, and endanger human and animal life. These incidents lead to loss of human and animal life, severe damage to financial infrastructures as well as environments. However, if immediate action is taken then the loss incurred by these incidents can be hugely minimized. Definitely, vision based automated systems can prove to be a boon in detecting such incidents. This work is also motivated with this idea, and hence proposes a deep learning-based fire detection framework with high accuracy and significant recall.

In this paper, the designed framework's main objective was to detect and localize fire by using real-world images. Transfer learning was applied to make the implementation process more efficient. Reusing pre-trained fire detection models comes with advantages like high accuracy thereby yielding the potential to detect fire flames effectively. Also an attention mechanism was implemented in the proposed model to draw model's focus to the relevant sections of the image. The fundamental procedures implemented in this paper are outlined below:

- Considering the limitations of traditional sensor-based fire detectors, an effective CNN framework for fire detection from real-world images is proposed. The proposed framework avoids the long-drawn procedure of feature engineering and spontaneously learns sample features from data.
- Datasets of fire images were not abundantly present and difficult to obtain. Hence, a composite dataset was created by collecting images collected from the data of well-known public datasets used in recent works on this problem (Kim and Lee, 2019).
- Motivated by the transfer learning strategies, numerous state-of-the-art CNN architectures such as Resnet50, VGG16, GoogLeNetV3, and EfficientNetB0 were explored for this work.
- The model was extensively fine-tuned as such fine-tuned models yield better performances with a different number of epochs (Luo, Zhao, Liu, and Huang, 2017). Also, the computational complexity of the proposed framework was stabilized along with the accuracy and size of the model, making it a good system for detection (Khan, Muhammad, Mumtaz, Baik, & de Albuquerque, 2019).
- The framework design uses a Global Average Pooling 2D (GAP) scheme to extract features along with an attention mechanism that directs the model's focus to different regions of an image for improved efficacy.
- For better visual identification and localization, the proposed work also uses the Grad-CAM method so that the part or region of the image gets highlighted. This helps to recognize the impact on the prediction of the class of that image and also gives insights about the failure modes of the model.

The remaining content of the paper is structured as follows. Section 2 presents the recent related works in the area of fire detection. This is followed by Section 3 which gives an overview of the proposed framework along with its components. The results obtained from the various experiments are discussed in the Section 4. Lastly, Section 5 presents conclusion to this work along with the directions for future work.

2. Related works

For various research communities, automatic detection of fire by making use of computer vision techniques and deep learning models has now become an open challenge because of many reasons like the resemblance with other natural objects such as sunlight, and lightings (Khan et al., 2019). The conventional feature engineering/extraction based methods like (Mohdiwale, Sahu, Sinha, & Bhateja, 2021) and (Khare, Bajaj, & Sinha, 2020) do seem to be promising, but do not appear to be the ideal choice to work on image based problems. Hence, these days deep learning methods have attained a state-of-the-art performance on the tasks of computer vision (Xu, Zhang, Zhang, Lin, & Wang, 2017). Also, deep learning has various applications like object detection/classification in images, videos and is now also used for real-time detection of any activity, speech recognition, and natural language processing, etc. (Kim and Lee, 2019). This section thus provides an insight into the researches made recently on vision-based early fire detection systems.

Currently, research on fire detection based on computational vision have proposed solutions based deep neural networks like CNNs have shown promising results. Therefore, few researchers have initiated studies on the field of fire detection using CNNs to further improve performances. The present literature highlights that some proposed solutions for fire detection systems include shape, color, texture, and motion features. For instance, (Luo et al., 2017) designed an algorithm for smoke detection by studying the motion elements of smoke. Making use of an automatic learning-based approach, suspected features were generated by CNN and the methodology used was background dynamic update and dark channel prior algorithm, resulting in a general and easily implementable approach. Also, a strategy of implicit enlargement was used where the suspected smoke regions were amplified that improved the timeliness of the detection.

In the work of (Saeed, Paul, Karthigaikumar, & Nayyar, 2019), a hybrid model was developed which comprised of an Adaboost-MLP neural model to predict the fire. Subsequently, an Adaboost-LBP model was designed to create the Regions of Interests (ROIs) and lastly a CNN for fire detection with videos and images, captured from the cameras installed for the surveillance. In (Khan et al., 2019) an economical system using deep CNNs for early detection of smoke in normal as well as foggy IoT environments was proposed. The designed process was applied to a self-made dataset and used VGG-16 architecture, for training.

(Kim and Lee, 2019) designed a Faster Region-based Convolutional Neural Network (R-CNN) with a video sequence for detection of the suspected areas of fire and non-fire based on the spatial features. Additionally, the features summed up were collected by LSTM for classification. In (Muhammad, Khan, Elhoseny, Ahmed, & Baik, 2019) a deployable model was designed on mobile devices and a lightweight deep neural networks system was created for detection of fire using videos taken in unknown surveillance scenarios. The proposed network was computationally inexpensive as it has no dense fully connected layers. (Kang, Wang, Chou, Chen, & Chang, 2019) developed a lightweight framework for the detection of fire using images based on deep learning using tiny-YOLOv3 architecture. The developed model achieved better detection accuracy with lower complexity in the training stage by some parameter adjusting.

(Muhammad, Ahmad, Mehmood, Rho, & Baik, 2018) proposed a low-cost CNN architecture for the detection of fire using surveillance videos. Knowing the suitable computational complexity for the problem, the designed model was trained on Google Net architecture. In (Maksymiv, Rak, & Peleshko, 2017) a method based on deep learning for detection of emergencies linking with fire and smoke using the processed data taken from the camera was developed. (Yin et al., 2017)

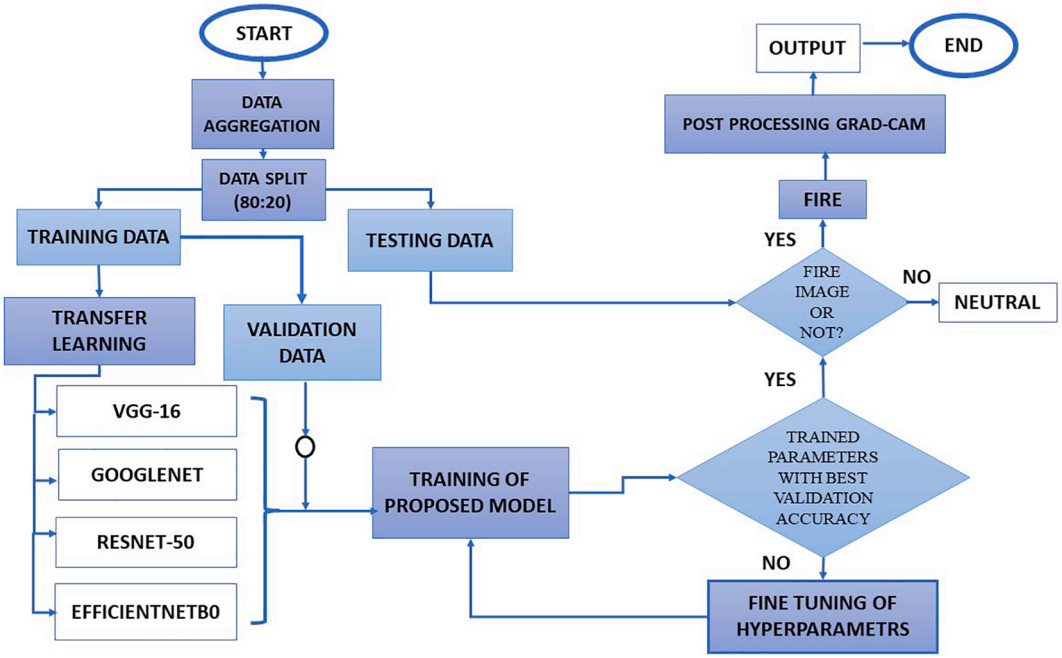

Fig. 1. Flowchart describing the adopted framework for this work.

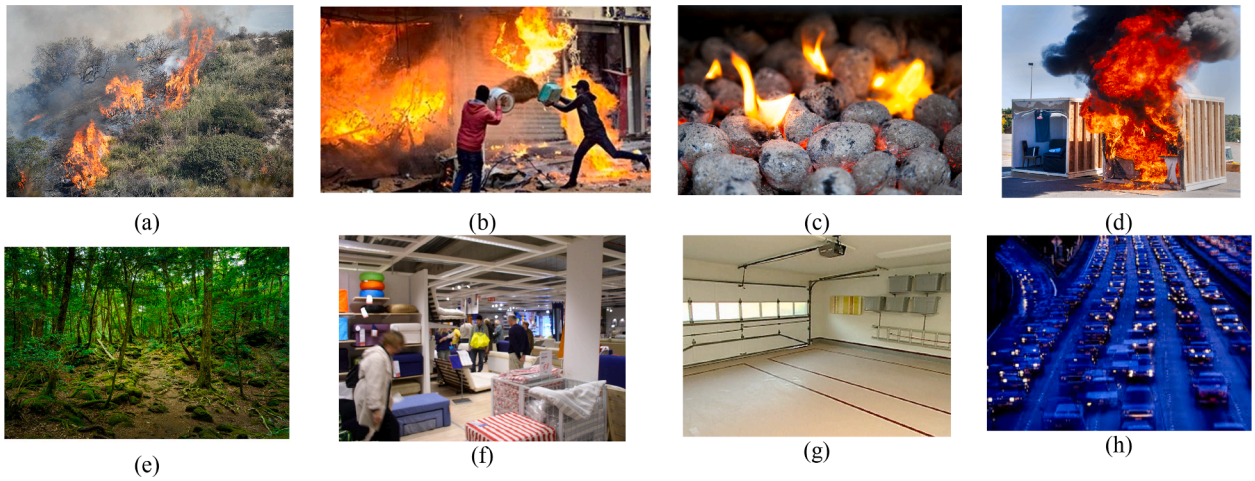

Fig. 2. Sample images from dataset (a)-(d) Fire images, (e)-(f) Neutral Images.

proposed a deep normalization and CNN (DNCNN) model which has 14 layers and was used for automated feature extraction and classification.

It can be observed from these works that a better performing model for vision based fire detection system is yet to be explored. Also the present solutions are challenged by the scarcity of datasets for the problem which limits the researcher while in analyzing the robustness of the proposed model. Hence this work explores for a computationally light, yet an effective solution for the vision based fire detection system trained and test over a significantly larger dataset.

3. Proposed methodology

This section describes the methodology and datasets used in this work along with the details about the evaluation and comparison of the proposed method. The primary purpose of the proposed contribution is to detect fire using an attention-based CNN model and also to visualize the localization with the Grad-CAM method. The proposed model was implemented using TensorFlow and Keras framework, while the training and testing of the model were performed on the Google Colab portal (Saeed et al., 2019). The GPU implementation was beneficial to accelerate the training process, especially during the fine-tuning process. For the known classification problem, different state-of-the-art models were used to train the network. The motivation for using the state-of-the-art models was good classification accuracy and speed efficiency (Muhammad et al., 2018). The flowchart describing the adopted framework for this work is depicted in Fig. 1.

3.1. Dataset

The aforementioned literature shows that the fire image samples for training and testing are not abundantly available. Most of the publicly available datasets available for this problem were too small to yield reliably efficient model. Hence a range of multiple datasets used in recent research works which contain fire as well as non-fire image samples were fused to form a single composite large dataset for this study. The composite dataset prepared for this work comprised of images from (Dataset, 2021; DeepQuestAI, 2021; Saied, 2020; Carlo, 2021; Bansal, 2021). The training data adds up to a total of 3988 fire images

Table 1
Details Of The Composite Dataset.

Dataset	Fire	Neutral	Total
Train	3417	3417	6834
Test	571	572	1143
Total	3988	3989	7977

and 3989 non-fire images. These images were from diverse real world scenery environments such as streets, buildings, people, indoors, halls, and forest, which assisted in building a robust model. The dataset posed significant challenges for the model as it included confusing colored objects such as sunlight scenes and lightings that made the task of fire detection even more difficult.

Fig. 2 shows some of the sample images from the composite dataset used for training and testing the model. Class wise composition of the prepared dataset are described in Table 1.

For effective training and testing, the dataset was split into train and test sets with a ratio of 80:20. And the validated data was 10% of the training data. The images in the dataset were uniformly resized to a size of 800×600. The input images were converted in the form of NumPy arrays and were resized to 224x224 pixels to reduce the storage size (Prabhu Ram, Gokul Kannan, Gowdham, & Arul Vignesh, 2020).

3.2. Proposed model architecture

Motivated by the process of visual perception of living creatures CNNs have been developed. The first widely-known architecture was the LeNet which was proposed in 1998, had shown good performance for hand-written digits classification (Lecun, Bottou, Bengio, & Haffner, 1998). In later years, many variants of CNNs have been proposed. Applications of CNNs include object detection/classification, action recognition, pose estimation, image segmentation, and scene labeling and also for understanding Natural Language Processing (NLP) and speech recognition. CNNs are broadly used to work with image based problems, attaining promising results over large-scale datasets. These networks have become the go-to models in deep learning because of their architecture which eliminates the need for hand crafted feature extraction phase. The detailed model architecture is presented in Fig. 3.

3.2.1. Convolutional layer

The convolutional layer is performs the crucial task of feature extraction from the input images. A Kernel/Filter is used for the convolution operation in the initial part of a convolution layer. The filters have the same depth as that of the input image for colored images, having multiple channels (RGB). These filters are of varied size and are employed along with the input data to create feature maps. After the extraction of high-level features, these maps are then subjected to a pooling layer.

3.2.2. Pooling layer

The pooling layer, also known as the subsampling layer, plays the role to minimize the spatial size of the features convolved. A special pooling layer known as the Global Average Pooling Layer (GAP) was applied in the proposed model after the backbone section of the network. This layer computes the average output of each feature map in the previous layer that provides extracted spatial features for each image. After the pooling process was completed Batch normalization was also applied to standardize the inputs to a layer for each mini-batch with a momentum of 0.99. The feature vector becomes the input to the fully connected layer.

3.2.3. Fully connected layer

The fully connected layer of the model consists of two or more hidden layers where the inputs are of higher representations. The FC layer learns to recognize full objects in different shapes and positions. In the designed model, the dense layer was used along with L2 norm regularization with a regularization factor of 0.01 which makes the model generalize better. Subsequently, the output was fed into a trainable layer i.e., to the spatial attention layers, where it learns to attention weights to get an attentionally-pooled feature vector for the classification section. The classification layer of the model attempts to learn a non-linear

Table 2
Hyperparameter space for fine-tuning the model.

Hyperparameter	Hyperparameter Space
Input Image Size Dimension	224×224
Pooling layer type in CNN	Max, Global Average Pooling
Dropout Rate	[**0.2**, 0.4, 0.6]
Learning Rate	[0.1, **0.01**, 0.001, 0.0001]
Optimizer	Adam, **AdaGrad**, SGD, RMSProp, NAdam
Batch Size	**32**, 64, 128
Epochs	10, **20**, 30, 50
Generalization Loss used	Cross-entropy
Regularization used	L1, **L2**
Activation function	Sigmoid, Rectified Linear unit (ReLu)

Table 3
Results of the Proposed Model.

Model	Test Accuracy	Precision	Recall	F-score
GoogLeNet	88.01	94.46	80.73	87.06
	91.25	86.17	98.24	91.81
VGG16	64.48	58.46	99.82	73.73
	92.21	93.97	90.19	92.04
ResNet50	92.54	88.64	98.42	93.27
	94.96	92.88	95.97	94.40
EfficientNetB0	**92.68**	**89.74**	**98.73**	**94.00**
	95.40	**91.77**	**97.61**	**94.76**

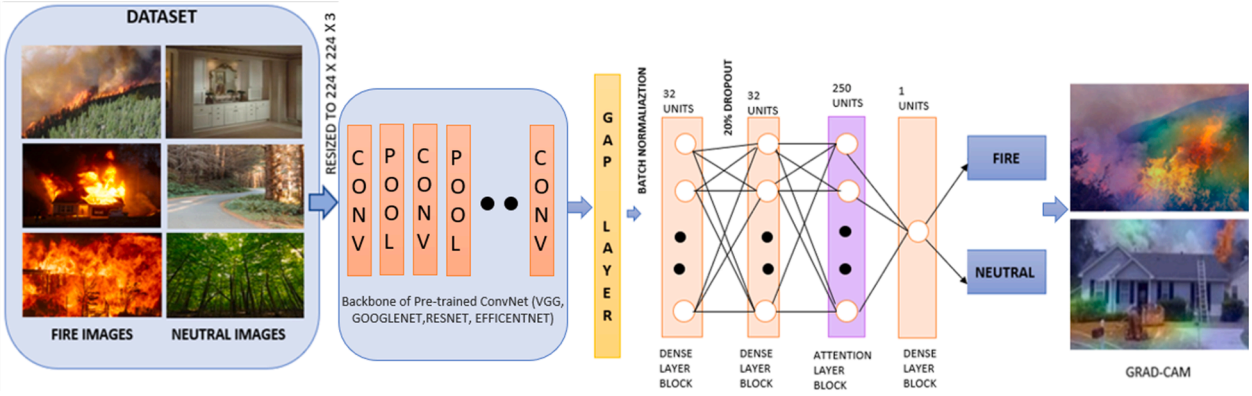

Fig. 3. Layered architecture of the Proposed Model.

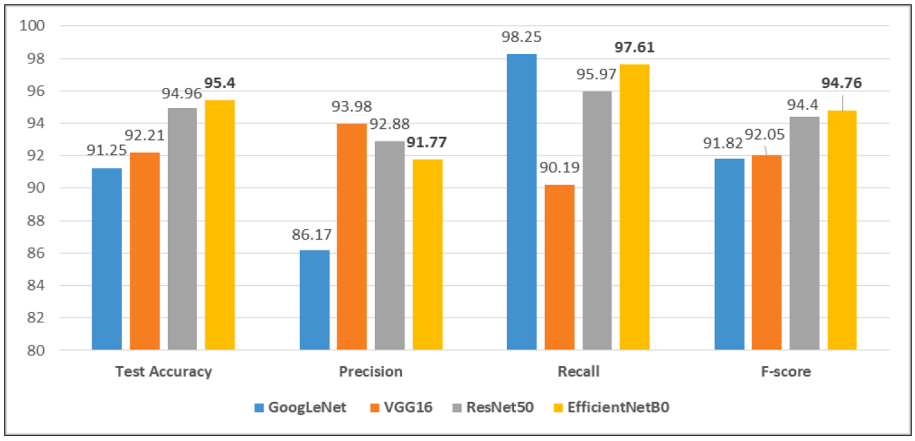

Fig. 4. Performance metric values for all the explored network variants.

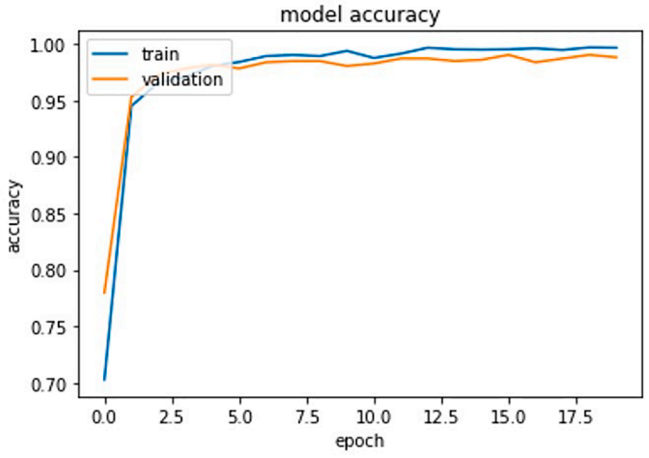

Fig. 5. Training and Validation Accuracy Curves for the EfficientNetB0 based model.

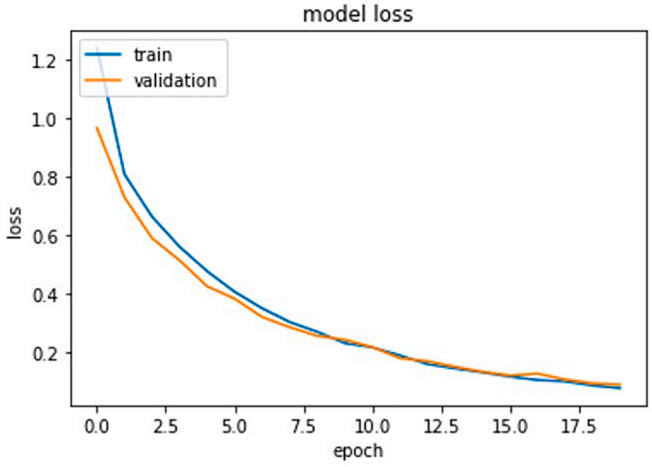

Fig. 6. Training and Validation Loss Curves for the EfficientNetB0 based model.

mapping (Moniruzzaman, Yin, & Qin, 2019).

An activation function of Rectified Linear Unit (Relu) was placed between all the layers excluding the last layer, which had sigmoid as its activation function. This function provides a probability distribution

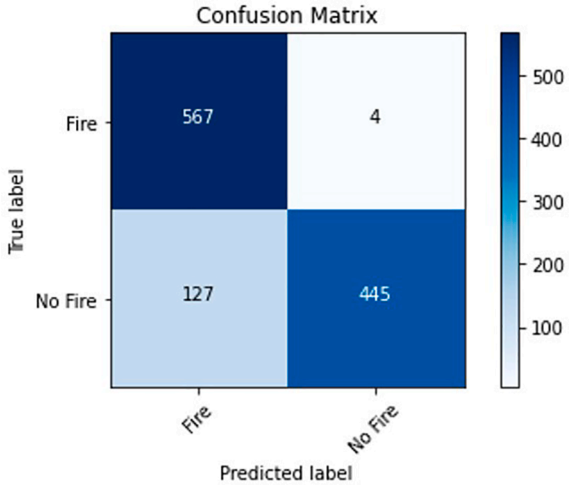

Fig. 7. Results of Confusion Matrix for the EfficientNetB0 based model.

that maps output in a range of 0 to 1.

3.2.4. Output Class

At the output phase, the architecture was reformed by defining a new top-level classifier on the base of the neural network, followed by a dropout layer to avoid overfitting. Additionally, the attention mechanism effectively enhanced the proposed network which showed a great improvement to explore the most informative features (Ba, Chen, Yuan, Song, & Lo, 2019). For testing the model, trained parameters with the best validation performance were used.

3.3. Transfer learning

Transfer learning is the popular concept where pre-trained models are reused to develop neural network models for a new problem (Saba et al., 2020). Various state-of-the-art pre-trained models for image based problems are publicly available that have been trained benchmark datasets with high efficiency. However, these models cannot be directly applied to any other images related problem as they might not be trained for the concerned task. Hence, instead of developing the model from scratch, these pre-trained models can be used where these can be further fine-tuned. For choosing an optimal architecture that would be effective for the concerned classification problem, a plethora of state-of-the-art models such as VGG-16, GoogLeNetV3, ResNet50, and EfficientNetB0, were explored. Though other model construction approaches including ensembling methods have been used in various studies (Kshatri et al.,

Fig. 8. GRAD-CAM visualization of test images showing the model's focus on exact sites of fire.

Fig. 9. Some misclassification of the proposed model.

Table4
Parameter Details of the Proposed Model.

Model	Total	Trainable	Non-Trainable
FireNet Model [6]	649,182	649,182	0
VGG16 [11]	134,342,526	134,342,526	0
VGG [12]	4,751,650	4,751,650	0
GoogLeNet [21]	22,826,846	22,792,414	34,432
MobileNet [17]	2,283,646	2,249,534	34,112
EfficientNetB0 (Proposed)	4,096,378	4,051,802	44,576

2021), but such networks tend to be computationally expensive, due to which these have not been employed in this work.

The VGG-16 network, proposed by Simonyan and Zisserman (2014), has achieved the top-5 accuracy of 92.3 % on ImageNet. This network architecture has been used for other applications as well (Cheah et al., 2021) But when trained for the fire classification problem, VGG-16 was extremely slow in training and didn't appear to be an accurate model as per the results obtained. Hence, was not used further for training because of inferior speed efficiency. Whereas, GoogLeNet, which was proposed by (Szegedy et al., 2015) showed better model training speed, but was an architecturally heavier model than VGG16. To assess the

Table 5
Comparison of proposed Attention-based CNN model with other methods on the fire dataset.

Model	Test Accuracy	Precision	Recall	F-score	TP	TN	FP	FN
VGG16 [11]	80.58	72.05	1.0	83.72	571	350	222	0
MobileNet [17]	87.66	80.80	98.77	88.88	564	438	134	7
GoogLeNet [21]	89.41	84.50	96.49	90.10	551	471	101	20
FireNet Model [6]	87.31	89.59	84.41	86.92	482	516	56	89
VGG12 [16]	75.15	67.06	98.77	79.88	564	295	277	7
Dilated CNNs [33]	73.12	65.02	83.11	75.01	550	426	121	46
MobileNet [34]	84.31	76.32	81.10	80.03	560	308	250	25
EfficientNetB0 (Proposed)	**95.40**	**91.77**	**97.61**	**94.76**	**557**	**522**	**50**	**14**

robustness, ResNet50, proposed by (He, Zhang, Ren, & Sun, 2016) a residual learning framework was used, which had shown slightly better performances for the concerned problem. Residual networks allow the training of deep networks by constructing the network through modules known as residual models. Among the CNNs used, the EfficientNetB0 (Tan, 2019) framework was dealt with strategically scaling deep neural networks. Employing this pre-trained neural network was an ideal choice as it had significantly better model efficiency and was also a lightweight model.

In this work, during the training phase, only the top 2 layers and newly stacked classification layer were trained, and the rest of the layers were made freeze. The drawn-out features from CNNs became an input to the attention model to get the attentionally-pooled feature depiction, after which the classification layer processes them for the final classification Moniruzzaman et al., 2019.

3.4. Attention mechanism

Attention mechanism has attained popularity newly and carries on to be an omnipresent module in state-of-the-art models. For the framework proposed, this technique focuses on the most important characteristics for fire categorization and has shown significant improvements in the performance. While processing the data, the mechanism focuses and pays more attention to different sections of the input. As a result, the proposed model concentrates and adds attention to the relevant regions of the image.

While the Attention mechanism has gone through various adaptations over recent years to suit multiple tasks, there are multiple types of attention applied. Hence, the proposed method focuses and sets additional "Attention" on the applicable parts of the image (Loye, 2021). Various researchers have initiated a study on the mechanism of attention that is implemented in the proposed framework to get a better idea of the discriminative features in the image. In (Cao et al., 2019), the paper proposed a network called a Bi-LSTM comprised of the spatial features' extraction network, a Bidirectional LSTM, along with a temporal attention subnetwork. It focuses on spatiotemporal features from image patch sequences and rewards contrasting levels of attention to separate patches.

Moniruzzaman et al., 2019 designed a framework comprising convolutional feature maps of deep CNN along with a mechanism of spatial attention for the classification of fire and traffic accident scenes. The attention model gives an insight into the most discriminative convolutional attributes. The model was implemented using videos and has shown promising performance. (Ba et al., 2019) designed a CNN framework for the detection of scenes of smoke by making use of satellite remote sensing. Moreover, a model named "SmokeNet" was developed which includes a mechanism of spatial and channel-wise attention in CNN. (Wu et al., 2019) explored a new method through hybrid deep-learning models, which includes detection of movement of fire and maximally stable extreme region in videos. Additionally, for classification, the SVM technique was used to eliminate false candidate fire regions. Further, the authors used a mechanism of channel-wise attention for detecting the rating of fire.

3.5. Model parameters

For training individual variant of the proposed model, the initial hyper-parameters, optimization method, and loss function were kept similar. The weights of the spatial attention network were learned using different optimizers, however the AdaGrad optimizer showed better performance for the selected problem. The network was trained with a learning rate of 0.01 (in some cases 0.001) and the batch size was set to 32. The model was fine-tuned by training it at different epochs. To obtain the best accuracy for these parameters, the final results of all the models were trained and compared at an epoch of 20. To make the model generalize better, slight modifications to the learning algorithm were done, which improved the model's performance on the test data as well. Some of the regularization techniques used were:

- A Dropout of 0.2 was added in between the Dense layers where a certain set of neurons are not considered during a particular forward or backward pass while training.
- L2 norms also known as Ridge Regression (least squares) have been used along with the dense layers. It minimizes the sum of the square of the differences between the target value and the estimated values.

The space of feasible values in which each respective hyper-parameter was fine-tuned is described in Table 2.

4. Results and discussions

In this section, evaluation and comparison results of the proposed model are presented. The performance of the model was calculated by some statistical measures like Precision, Accuracy, F-Score, and Recall. High results of these parameters show that the designed model was very efficient and highly accurate. Eventually, 2 models were used in this work, accordingly the results are presented.

4.1. Evaluation metrics

For a classification problem, the evaluation protocol includes various statistical metrics such as precision, recall, f-score, and accuracy. A confusion matrix is also used to assess the model performance. The Confusion Matrix shows an insight in which the classification module gets confused when it makes predictions. The predicted classes for the problem are displayed in the columns of the confusion matrix, whereas the actual classes are in the rows of the matrix. The confusion matrix can be divided into 4 groups of classification (Facts Statistics: Wildfires, 2021):

- True positives (TP) [sensitivity]: The model predicted the 'fire' image correctly.
- True negatives (TN): The model predicted a 'neutral' image correctly.
- False positives (FP): The model predicted the image incorrectly as 'fire'.
- False negatives (FN): The model predicted the image incorrectly as 'neutral'.

Other metrics are calculated using the following formula:

$$Precision = TP/(TP + FP) \quad (1)$$

$$Recall = TP/(TP + FN) \quad (2)$$

$$F - Measure = 2*((Precision*Recall)/(Precision + Recall)) \quad (3)$$

4.2. Experimental results

The proposed architecture has been explored with 4 state-of-the-art architectures, each trained several times to identify the best model. Parameters were fine-tuned till the time an accurate result was obtained. From the various combinations of models, an observation was made that ResNet50 and EfficinetNetB0 architectures results were accurate enough and the best validation accuracy was considered further with AdaGrad as an optimizer and a learning rate of 0.01. Each model combination that was obtained by fine-tuning hyperparameters was trained 5 times and the mean of the best results was taken into consideration. An observation was made that there was a slight improvement with a significant difference of 2% in the testing performance of the model when an attention layer was added. It can be concluded that EfficientNetB0 was a much better alternative because the total parameters of the network are considerably less, making it a lighter model as compared to ResNet50. These results of the two architectures with the best parameters are presented below in Table 3.

In Table 3 above, the red highlighted value indicates scores without attention layer, while the values in regular black color represent the score with attention layer. Fig. 8 presents a graph for the performance metric values shown in Table 3 for the all the explored network variants. Fig. 9.

Figs. 4 and 5 outline the training and validation accuracy and loss curves respectively. The consistency of the curves in these graphs highlight the stability of the model's performance for the selected problem. Though in Fig. 5, the loss may appear to be on a receding trend, but the model did not train further as no improvements in model's accuracy were observed beyond the certain epoch limit. Fig. 6 shows the results of the confusion matrix obtained during the test process. The curves are a result of a trained EfficientNetB0 model with 20 epochs with a batch size of 32 using the Adagrad optimizer. The model had shown a training accuracy of 99.90.

The predictions of the designed framework were presented by using a technique that created visual explanations from the test dataset of Convolutional Neural Network (CNN)-based models, making them more visually clear. The approach used was 'Gradient-weighted Class Activation Mapping' (Grad-CAM) (Selvaraju et al., 2017) that used the gradients of any target concept, such as in the proposed framework where the target was the detection of fire. This was used to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Some snapshots of images from the dataset used in this paper are shown.

Fig. 7 shows the attention on the feature map and thus representing true positives. The superimposed part represents the localization of the fire as predicted by the model showing correct results. While Fig. 8 shows the results of some misclassification using Grad-CAM on the dataset and thus representing false positives. The superimposed sites show that fire was being incorrectly localized by the model and thus predicting incorrect results.

4.3. Comparison of model performances

To compare the performance and to check the effectiveness of the proposed attentive-based convolutional neural network, the following methods were implemented to the proposed dataset for comparison. The parameters of the models are presented in Table 4 below.

The table above shows that the number of trainable parameters for the proposed model are definitely on the lower side compared to the other solutions for this problem. Though the fire net model has minor trainable parameters, it fails to detect fire in images with high efficacy. The results of the comparisons with accuracy details are given in Table 5.

As seen in Table 5, the overall performance of the proposed model using the prepared composite dataset was observed to be more than 95% with a very high recall of 97.61%. The performance of (Muhammad et al., 2018) was the next best with the least number of epochs. Their model was trained on GoogLeNet architecture and showed 89.41% test accuracy within 6 epochs. The work in (Khan et al., 2019), was an economic system, proposed using deep CNNs for early detection of smoke in normal as well as foggy IoT environments. Their process was applied to a self-made dataset and used VGG-16 architecture trained for 30 epochs with a batch size of 16 and an SGD optimizer. The model showed slightly underfit results on the selected dataset but had 0 false positives.

The model by (Jadon et al., 2019) has designed a lightweight neural network, named FireNet, that has shown good performance for 100 epochs. In (NAMOZOV and CHO, 2018) proposed a novel deep CNN model to attain high accuracy for detecting fire and smoke. The VGG-12 model didn't show good performance on the proposed dataset and lacked performance in detecting fire since the data size was insufficient. (Muhammad et al., 2019) designed a deployable model on mobile devices and created a lightweight deep neural networks system to detect fire. Their network was computationally inexpensive as it had no dense, fully connected layers. Mobile Net architecture was easy to implement and had good training speed compared to VGG models, and hence the results performance was superior.

The framework proposed by (Valikhujaev, Abdusalomov, & Cho, 2020) used a dilated convolutional neural network in their work. The model was tested on a custom-built dataset that included fire and smoke images gathered from the internet and labeled manually. Although the work was superior to previous approaches, false positives still account for 10.6% of the time, and there is still space for development in accuracy and false positives. To detect fire, (Dua, Kumar, Singh Charan, & Sagar Ravi, 2020) suggested a fire detection system based on transfer learning (deep CNN technique). For building a fire detection system, it employs pre-trained deep CNN architectures such as VGG and MobileNet. To simulate real-world conditions, these models were evaluated on unbalanced datasets. On the other hand, our dataset yielded a false positive rate of 21.9 percent on their proposed model. The thought to select the above methods for comparative analysis was that the methods were developed for fire detection. Moreover, all the networks have their applications and particular objective for the detection of fire.

5. Conclusion

This paper presents an attentive-based CNN model for the detection of fire using real-world images. An attention mechanism was also added to the model that yielded significant improvement in the performance of data. The introduced neural network had shown significantly good performance on the testing dataset with minimal false negatives. Several trials were executed for exploring a better performance than the baseline approach. From the several models explored for this problem, the EfficientNetB0 tuned out to be a highly efficient alternative with less trainable parameters. The model thus obtained yielded better results than most of the recently proposed solutions to this problem. Future work is devoted to expanding on the current work and develop a robust fire and smoke detection algorithm using videos. Future studies in this area can also explore the application of special generative networks such as GANs for this problem.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

References

"Wildfires, forest fires around world in 2020," *Anadolu Ajansı*. [Online]. Available: https://www.aa.com.tr/en/environment/wildfires-forest-fires-around-world-in-2020/2088198 [Accessed: 9-Mar-2021].

Ba, R., Chen, C., Yuan, J., Song, W., & Lo, S. (2019). SmokeNet: Satellite Smoke Scene Detection Using Convolutional Neural Network with Spatial and Channel-Wise Attention. *Remote Sensing, 11*(14), 1702.

P. Bansal, "Intel Image Classification," *Kaggle*, 30-Jan-2019. [Online]. Available: https://www.kaggle.com/puneet6060/intel-image-classification [Accessed: 10-Mar-2021].

Cao, Y., Yang, F., Tang, Q., & Lu, X. (2019). An Attention Enhanced Bidirectional LSTM for Early Forest Fire Smoke Recognition. *IEEE Access, 7*, 154732–154742.

Carlo, "fire_and_smoke.zip," *Kaggle*, 24-Jul-2019. [Online]. Available: https://www.kaggle.com/carlo946/fire-and-smokezip [Accessed: 10-Mar-2021].

Cheah, Kit Hwa, Nisar, Humaira, Yap, Vooi Voon, Lee, Chen-Yi, Sinha, G. R., & Maietta, Saverio (2021). Optimizing Residual Networks and VGG for Classification of EEG Signals: Identifying Ideal Channels for Emotion Recognition. *Journal of Healthcare Engineering, 2021*, 1–14. https://doi.org/10.1155/2021/5599615

DeepQuestAI,"DeepQuestAI/Fire-Smoke-Dataset," *GitHub*. [Online]. Available: https://github.com/DeepQuestAI/Fire-Smoke-Dataset . [Accessed: 10-Mar-2021].

Dua M., Kumar, M., Singh Charan, G., Sagar Ravi, P., "An improved approach for fire detection using deep learning models", 2020, International Conference on Industry 4.0 Technology (I4Tech). https://doi.org/10.1109/i4tech48345.2020.9102697.

Facts Statistics: Wildfires. (2021). Retrieved from https://www.iii.org/fact-statistic/facts-statistics-wildfires.

Fire Detection Dataset | Kaggle." [Online]. Available: https://www.kaggle.com/atulyakumar98/test-dataset. [Accessed: 10-Mar-2021].

K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Arpit Jadon, Omama, Mohd Varshney, Akshay Ansari, Samar Sharma, Rishabh, (2019), "FireNet: A Specialized Lightweight Fire & Smoke Detection Model for Real-Time IoT Applications", arXiv:1905.11922.

L.-W. Kang, I.-S. Wang, K.-L. Chou, S.-Y. Chen, and C.-Y. Chang, "Image-Based Real-Time Fire Detection using Deep Learning with Data Augmentation for Vision-Based Surveillance Applications," *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019.

Khan, S., Muhammad, K., Mumtaz, S., Baik, S. W., & de Albuquerque, V. H. C. (2019). Energy-Efficient Deep CNN for Smoke Detection in Foggy IoT Environment. *IEEE Internet of Things Journal, 6*(6), 9237–9245. https://doi.org/10.1109/JIoT.648890710.1109/JIOT.2019.2896120

Khare, Smith K., Bajaj, Varun, & Sinha, G. R. (2020). "Adaptive Tunable Q Wavelet Transform-Based Emotion Identification. *IEEE Transactions on Instrumentation and Measurement, 69*(12), 9609–9617. https://doi.org/10.1109/TIM.1910.1109/TIM.2020.3006611

Kim, B., & Lee, J. (2019). A Video-Based Fire Detection Using Deep Learning Models. *Applied Sciences, 9*(14), 2862.

Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. *IEEE Access, 9*, 67488–67500. https://doi.org/10.1109/access.2021.3075140

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Li, Xiuqing, Chen, Zhenxue, Wu, Q. M. Jonathan, & Liu, Chengyun (2020). 3D Parallel Fully Convolutional Networks for Real-Time Video Wildfire Smoke Detection. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(1), 89–103.

Li, P.u., & Zhao, W. (2020). Image fire detection algorithms based on convolutional neural networks. *Case Studies in Thermal Engineering, 19*, 100625. https://doi.org/10.1016/j.csite.2020.100625

G. Loye, "Attention Mechanism," *FloydHub Blog*, 17-Jan-2020. [Online]. Available: https://blog.floydhub.com/attention-mechanism/ [Accessed: 09-Mar-2021].

Luo, Y., Zhao, L., Liu, P., & Huang, D. (2017). Fire smoke detection algorithm based on motion characteristic and convolutional neural networks. *Multimedia Tools and Applications, 77*(12), 15075–15092. https://doi.org/10.1007/s11042-017-5090-2

O. Maksymiv, T. Rak, and D. Peleshko, "Real-time fire detection method combining AdaBoost, LBP and convolutional neural network in video sequence," *2017 14th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, 2017.

Mohdiwale, S., Sahu, M., Sinha, G. R., & Bhateja, V. (2021). Statistical Wavelets With Harmony Search- Based Optimal Feature Selection of EEG Signals for Motor Imagery Classification. *IEEE Sensors Journal, 21*(13), 14263–14271. https://doi.org/10.1109/jsen.2020.3026172

M. Moniruzzaman, Z. Yin, and R. Qin, "Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification," *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2019.

Muhammad, K., Ahmad, J., Mehmood, I., Rho, S., & Baik, S. W. (2018). Convolutional Neural Networks Based Fire Detection in Surveillance Videos. *IEEE Access, 6*, 18174–18183.

Muhammad, K., Khan, S., Elhoseny, M., Ahmed, S. H., & Baik, S. W. (2019). Efficient Fire Detection for Uncertain Surveillance Environment. *IEEE Transactions on Industrial Informatics, 15*(5), 3113–3122.

NAMOZOV, A., & CHO, Y. I. (2018). An Efficient Deep Learning Algorithm for Fire and Smoke Detection with Limited Data. *Advances in Electrical and Computer Engineering, 18*(4), 121–128.

N Prabhu Ram, R Gokul Kannan, V Gowdham, R Arul Vignesh," Fire Detection Using CNN Approach", *International Journal of Scientific & Technology Research* Volume 9, Issue 04, April 2020, ISSN 2277-8616.

Saba, L., Agarwal, M., Sanagala, S. S., Gupta, S. K., Sinha, G. R., Johri, A. M., … Suri, J. S. (2020). Brain MRI-based Wilson disease tissue classification: An optimised deep transfer learning approach. *Electronics Letters, 56*(25), 1395–1398. https://doi.org/10.1049/ell2.v56.2510.1049/el.2020.2102

Saeed, F., Paul, A., Karthigaikumar, P., & Nayyar, A. (2019). Convolutional neural network based early fire detection. *Multimedia Tools and Applications, 79*(13–14), 9083–9099.

A. Saied, "FIRE Dataset," *Kaggle*, 25-Feb-2020. [Online]. Available: https://www.kaggle.com/phylake1337/fire-dataset?select=fire_dataset [Accessed: 10-Mar-2021].

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv, 1409*, 1556.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

M. Tan V.L. Quoc "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". 2019, arXiv:1905.11946.

Valikhujaev, Y., Abdusalomov, A., & Cho, Y. I. (2020). Automatic fire and smoke detection method for surveillance systems based on Dilated CNNS". *Atmosphere, 11*(11), 1241. https://doi.org/10.3390/atmos11111241

Wu, Yirui, He, Yuechao, Shivakumara, Palaiahnakote, Li, Ziming, Guo, Hongxin, & Lu, Tong (2019). Channel-wise attention model-based fire and rating level detection in video. *CAAI Transactions on Intelligence Technology, 4*(2), 117–121.

Xu, G., Zhang, Y., Zhang, Q., Lin, G., & Wang, J. (2017). Deep domain adaptation-based video smoke detection using synthetic smoke images. *Fire Safety Journal, 93*, 53–59.

Yin, Z., Wan, B., Yuan, F., Xia, X., & Shi, J. (2017). A Deep Normalization and Convolutional Neural Network for Image Smoke Detection. *IEEE Access, 5*, 18429–18438.