



# Optimized IoT-Based Multimodal Fusion for Early Forest Fire Detection and Prediction

Abdi Muhaimin<sup>1</sup>, Edriyansyah<sup>2</sup>, Wahyat<sup>3</sup>, Yuda Irawan<sup>4</sup> and Refni Wahyuni<sup>5</sup>

## ABSTRACT

Forest and land fires are recurring ecological disasters that pose serious threats to environmental sustainability, particularly in vulnerable regions like Indonesia. Conventional fire detection methods using only visual or single-sensor data often suffer from low accuracy in poor lighting, thin smoke, or extreme weather. This study proposes an IoT-based multimodal system that combines visual imagery and real-time meteorological sensor data. Fire detection was conducted using the YOLOv11 model, trained for 50 epochs with the SGD optimizer. The model achieved a precision of 87.9%, recall of 79.7%, mAP@0.5 of 87.7%, and mAP@0.5:0.95 of 53.7%. Detected images are further classified using a hybrid ViT-GRU model, which achieves 99.97% accuracy by capturing spatial and temporal fire patterns. We performed fire detection using an LSTM model optimized with Optuna and SMOTE, yielding 92.66% accuracy and an AUC of 1.00. The decision-level fusion approach integrates visual and sensor outputs to improve the accuracy and contextual relevance of the final prediction. We deployed the system in a real-time Streamlit dashboard connected to cloud-based data acquisition. Results show that this multimodal approach significantly improves the reliability of early fire detection and risk prediction.

## Article information:

**Keywords:** YOLOv11, ViT-GRU, IoT, Multimodal Fusion, Fire

## Article history:

Received: July 1, 2025

Revised: August 26, 2025

Accepted: September 2, 2025

Published: September 20, 2025

(Online)

**DOI:** 10.37936/ecti-cit.2025194.262839

## 1. INTRODUCTION

Forest and land fires are a recurring ecological disaster in Indonesia, with more than 300,000 hectares of land burning throughout 2023, causing environmental damage, economic losses, and health threats due to haze[1]. This condition is exacerbated by climate change, prolonged dry seasons, and weak early detection systems that are fast and accurate. One of the main obstacles in environmental data-based fire prediction is the quality and continuity of meteorological data, which often experience missing values due to device interference or extreme conditions in the field[2], [3]. In this context, early detection refers to the capability of a system to identify fire incidents at the very initial stage, such as the emergence of thin smoke, small flame spots, or abnormal dryness

conditions, before the fire spreads uncontrollably and causes large-scale damage. Previous systems that relied only on meteorological data often failed to capture these subtle early indicators in real-time, since weather variables primarily reflect broader environmental conditions rather than the immediate presence of ignition. Therefore, integrating meteorological data with real-time visual detection is essential to achieve a truly effective early warning system. Therefore, an IoT-based system is needed that is capable of acquiring meteorological data in real-time and continuously in fire-prone areas[4]. IoT connects physical devices to exchange data and enable intelligent automation[5]–[10].

Previous research has used methods such as Support Vector Regression[11], [12] and Fuzzy Logic[13], [14] in fire prediction. However, this approach is still

<sup>1</sup>The author is with the Department of Information System, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Indonesia, Email: [abdi.muhamin86@gmail.com](mailto:abdi.muhamin86@gmail.com)

<sup>2,4,5</sup>The authors are with the Department of Computer Science, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Indonesia, Email: [sinksonk@gmail.com](mailto:sinksonk@gmail.com), [yudairawan89@gmail.com](mailto:yudairawan89@gmail.com) and [refniabid@gmail.com](mailto:refniabid@gmail.com)

<sup>3</sup>The author is with the Department of Computer Network Administration, Politeknik Negeri Bengkalis, Indonesia, Indonesia, Email: [wahyat@polbeng.ac.id](mailto:wahyat@polbeng.ac.id)

limited because it does not fully utilize the temporal relationship between sensor variables, so that the resulting fire risk prediction is less than optimal. Meanwhile, Computer Vision technology such as YOLO has been widely used in automatic visual detection of fire and smoke[15]–[17]. The YOLOv9 and YOLOv10 models offer improved accuracy and detection speed, but still rely on CNN architectures that are less effective at capturing global spatial context[18]–[20]. Both still have limitations in detecting thin smoke, fire objects in low lighting conditions, at night, or in thick fog[21]–[23]. Many of these approaches still have limitations, especially in terms of multimodal integration and detection accuracy in complex environmental conditions. IoT-based fire detection systems have been widely developed to monitor environmental parameters in real-time. Previous studies have provided early warnings using temperature and fire sensors, but their scope is limited and they have not utilized intelligent machine learning[24].

The weakness of this previous approach is the basis for the development of this research, namely the integration of the YOLOv11 model with the Vision Transformer (ViT), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models. The integration of ViT, GRU with YOLOv11 is done to overcome the weakness of Convolutional Neural Networks (CNN) in capturing long-range dependencies and global spatial context in fire images[25]–[27]. ViT has been introduced with superior self-attention capabilities in recognizing complex visual patterns compared to CNN approaches[28]–[30]. The ViT extracted feature vectors are fed into the GRU to capture temporal or sequential patterns that emerge from the visual representation [31], [32]. The integration of ViT and GRU forms a hybrid model that is able to combine the visual representation capabilities of ViT with the sequence modeling capabilities of GRU[33], thus improving the classification accuracy between fire and non-fire images. YOLOv11 offers high speed, better detection capability of small or faint objects, and improved detection accuracy in various complex visual conditions[34], [35]. LSTM model optimized with Synthetic Minority Oversampling Technique (SMOTE) technique and hyperparameter optimization using Optuna will excel in understanding the temporal pattern of environmental data from IoT sensors. Through multimodal integration, the system is not only able to visually detect the presence of fire or smoke, but can also perform early fire prediction based on real-time environmental conditions.

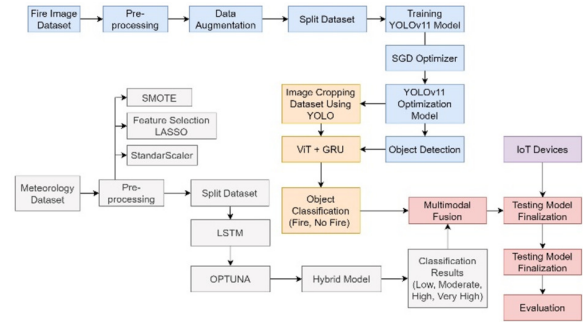
Although IoT-based sensing and vision-based detection can work independently, both approaches have limitations when used in isolation. IoT sensors effectively capture environmental variables such as temperature, humidity, or soil moisture, but they cannot directly confirm the visual presence of

fire. Conversely, vision-based systems such as YOLO are able to detect flames or smoke visually, but their performance can degrade under poor lighting, haze, or thin smoke conditions. By combining both modalities, the system leverages the complementary strengths of each approach, enabling more robust, accurate, and reliable early fire detection.

## 2. MATERIALS AND METHODS

### 2.1 Overview of the proposed methodology

This research uses a combination approach between computer vision and meteorological sensor data to build an intelligent and real-time forest fire detection and prediction system. Model integration is carried out gradually through the stages of preprocessing, training, classification, and multimodal fusion based on deep learning. The flow of model development can be seen in the following figure:



**Fig.1:** Flow of Model Development.

This research methodology proposes a multimodal hybrid deep learning approach based on the integration of visual data and environmental sensor data for real-time forest fire early prediction. The workflow begins with processing the fire image dataset through the stages of data augmentation, training the YOLOv11 model using Stochastic Gradient Descent (SGD) optimization, to produce precise detection of fire objects. The detected images are then cropped and further classified using a combination of ViT and GRU to capture spatial and sequential relationships simultaneously. In parallel, meteorological data from IoT devices is processed through standardization, Least Absolute Shrinkage and Selection Operator (LASSO) based feature selection, and data balancing using SMOTE, and then trained with LSTM optimized by Optuna. The classification results from ViT-GRU (fire/no fire) and LSTM (fire risk: low to very high) are then combined through a multimodal fusion process, forming a hybrid prediction system capable of adaptively integrating visual perception and environmental conditions. The main novelty of this methodology lies in the utilization of ViT-GRU for YOLO detection image classification, as well as the integration of image prediction and IoT data in an optimized deep learning-based multimodal framework,

which has not been explored in previous studies. The fusion stage combines probability distributions from both models using weighted voting, where the final decision corresponds to the class with the maximum weighted probability (arg max).

## 2.2 Fire image dataset processing

The fire image dataset used in this study consists of 7,244 original images obtained through a collection process from various open sources and manual collections. To increase the variation and generalization of the model, a data augmentation process was carried out using the Roboflow platform. The preprocessing process includes auto-orienting and resizing the image to a size of 640×640 pixels. Augmentation is carried out by producing two outputs per image sample through a small rotation transformation (between  $-2^\circ$  to  $+2^\circ$ ), shear ( $\pm 2^\circ$  horizontal and  $\pm 1^\circ$  vertical), blur up to 0.4 pixels, and noise up to 0.02% of the number of pixels. The results of this augmentation significantly increase the amount of training data to 10,380 images (83%), validation of 1,329 images (11%), and testing of 725 images (6%) which have been divided stratified. To strictly prevent data leakage, the dataset was first divided into training (83%), validation (11%), and testing (6%) subsets using stratified sampling. Only after this splitting was performed, the augmentation process was applied within each subset. This ensures that each original image and its augmented variants are confined exclusively to a single subset, thereby eliminating the possibility of the same image instance (or its transformation) appearing in both training and testing sets. The following is a display of the fire image dataset:



**Fig.2:** Fire Image Dataset View.

The training process is then carried out using the YOLOv11 architecture based on the Ultralytics framework, which is specifically configured to detect visual objects in the form of fire in images. The model is trained for 50 epochs using the SGD optimization algorithm, with a momentum of 0.937 and a weight decay of 0.0005 to avoid overfitting and improve convergence. During the training process, adaptive learning rate adjustments are applied through the cosine learning rate scheduler, with an initial value of 0.01. In addition, fine-tuning is carried out on several important components in the YOLOv11 architecture, such as Backbone (CSPDarknet), Neck (PANet), and

Head Detection (Anchor-Free), to be more sensitive to visual characteristics of fires such as small flames, thin smoke, or extreme lighting. This training stage produces an initial detection model with high precision, which will be used as a basis for the advanced classification stage using ViT-GRU. The training process of YOLOv11 using an NVIDIA RTX 3060 GPU with 12GB VRAM, taking approximately 3 hours. The model was optimized using Stochastic Gradient Descent (SGD) with a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The cosine learning rate scheduler was used to adaptively adjust learning rates during training.

## 2.3 Visual extraction and classification with ViT-GRU

Once the YOLOv11 model was properly trained to detect the presence of fire in the image, the next step was to perform image cropping of the detected objects. This process resulted in a total of 10,380 cropped images that specifically represent the visual parts that indicate the presence of fire or non-fire. This dataset was then used for training a classification model based on a combination of ViT and GRU. The ViT architecture is used as a feature extractor with a multi-head self-attention mechanism that is able to capture long-range dependencies in the spatial representation of images. This mechanism can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

Where  $Q, K, V$  are the query, key, and value matrices respectively, and  $d_k$  is the dimension of the key. The output of ViT in the form of a fixed-dimensional embedding vector (misal:  $R^{768}$ ) is then further processed by the GRU unit to model the spatial sequence in the time dimension, especially if there are variations in the visual sequence order (e.g. the sequence of flames from small to large flames). GRU is used because it has an efficient architecture to capture temporal patterns without causing vanishing gradient problems as in conventional RNNs. The internal process of GRU can be described by the formula:

$$h_t = \text{GRU}(f_{ViT}(x_t)) \quad (2)$$

Where  $f_{ViT}(x_t)$  is a feature representation of the image extracted by Vision Transformer  $h_t$  is the hidden state output of GRU which is then fed to the final classification layer. The combination of ViT and GRU forms a hybrid model that combines the advantages of visual global representation with temporal dynamics, thereby improving the classification ability between fire and non-fire categories more accurately and contextually.

## 2.4 Meteorological sensor data processing

The meteorological dataset used in this study consists of 5000 data entries collected from the Meteorology, Climatology and Geophysics Agency (BMKG) of Riau Province, Indonesia, covering the period 2011 to 2024. This dataset includes various environmental parameters such as minimum and maximum temperature, average temperature ( $T_{avg}$ ), average air humidity ( $RH_{avg}$ ), rainfall ( $RR$ ), sunshine duration ( $ss$ ), maximum wind speed and direction ( $ff_x$ ,  $ddd_x$ ), and soil surface moisture. The appearance of the dataset can be seen in the following image:

Tanggal	Tn: Temperatur minimum (°C)	Tx: Temperatur maksimum (°C)	Tavg: Temperatur rata-rata (°C)	RH_avg: Kelembapan rata-rata (%)	RR: Curah hujan (mm)	ss: Lamanya penyinaran matahari (jam)	ff_x: Kecepatan angin maksimum (m/s)	ddd_x: Arah angin saat kecepatan maksimum (°)	ff_avg: Kecepatan angin rata-rata (m/s)	ddd_car: Arah angin terbanyak (°)	Kelembaban Permukaan Tanah
01-01-2011	22.5	32.5	26.4	79	8	0	4	45	3	NE	50
02-01-2011	22.3	31.1	26	78	1	3.1	3		0	W	65
03-01-2011	22.4	32.5	26.1	81	0	0.8	4	45	3	NE	60
04-01-2011	23.5	33.2	27.6	74	18	0	25	45	3	NW	63
05-01-2011	22.6	33.5	27.3	75	7	0	10	315	4	W	57
06-01-2011	22.8	32.3	26.8	76	1	0	15	315	3	W	51
07-01-2011	22.3	30.2	25.9	80	0	5.6	3	270	3	W	69
08-01-2011	22.4	30.4	25	85	4	0	9	315	3	W	67
09-01-2011	23.5	29.5	25.8	82	3	1.1	4	360	3	N	69

**Fig.3:** Meteorology Dataset View.

The initial process begins with cleaning and standardizing data types, then SMOTE is performed to handle class imbalance in fire risk labels. Next, feature selection is performed using LASSO to select the most relevant features to the prediction target[36]. The selected features are then normalized using StandardScaler to ensure a uniform distribution on each model input. The dataset is then divided into training and testing data before being used for training the predictive model.

The training stage is carried out by building an optimized LSTM model using the Optuna algorithm to search for the best hyperparameters[37]. This process aims to optimize parameters such as the number of LSTM neuron units, learning rate, and number of epochs, so that the most efficient model is obtained in capturing sequential patterns between weather variables. The loss function used is categorical crossentropy and the model is compiled using the Adam optimizer. The training process is formulated as follows:

$$\hat{y} = \text{LSTM}(X_{scaled}; \theta^*) \quad (3)$$

Where  $\theta^*$  is the best set of search results parameters with Optuna, and  $X_{scaled}$  is the normalized meteorological feature matrix. Output  $\hat{y}$  in the form of predictions of fire risk levels in four classes: Low, Moderate, High, and Very High. This model is then used in the multimodal integration stage with visual imagery as part of a real-time fire detection and prediction system. The meteorological dataset was also divided using stratified splitting into 80% training and 20% testing.

To ensure robust model evaluation and prevent overfitting, the dataset was partitioned into training and testing subsets using stratified sampling, with

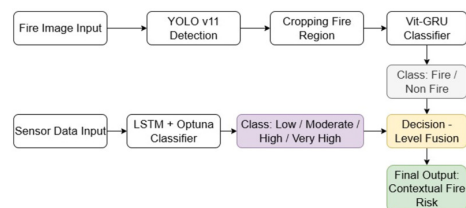
the testing data completely excluded from all preprocessing stages, including SMOTE oversampling and feature selection. Only the training subset underwent class balancing and hyperparameter optimization. This procedure guarantees that the model's performance metrics, such as accuracy and AUC, are measured strictly on unseen data. Furthermore, dropout regularization and early stopping strategies were employed during training to improve the model's generalization capability.

## 2.5 Multimodal fusion integration

The multimodal fusion integration stage was conducted to combine two different but complementary sources of information, namely the visual image classification results from the ViT-GRU model and the fire risk prediction results from meteorological sensor data processed through the LSTM-Optuna model. The integration is done at the decision-level fusion, where both model outputs are combined to produce a more accurate and contextualized final prediction. This fusion uses a weighted voting approach, with weights assigned based on the accuracy of each model on the validation data. Mathematically, the final classification result  $y_{final}$  can be expressed as:

$$y_{final} = \arg \max_{c \in C} (\alpha \cdot P_{ViT-GRU}(c) + \beta \cdot P_{LSTM}(c)) \quad (4)$$

Here,  $c$  represents a candidate class label from the set of all possible classes  $C$ . The probabilities  $P_{ViT-GRU}(c)$  and  $P_{LSTM}(c)$  denote the confidence scores assigned by the visual and sensor-based models, respectively. The final decision  $y_{final}$  is obtained by selecting the class  $c$  that maximizes the weighted probability combination, as indicated by the  $\arg \max$  operator, and  $\alpha$  and  $\beta$  are weights that satisfy  $\alpha + \beta = 1$ . Here,  $\arg \max$  refers to the argument of the maximum, which indicates the index or class label corresponding to the highest prediction probability among the outputs. This strategy allows the system to provide a final prediction based on a combination of visual and sensor information in an adaptive manner to the conditions on the ground. To clarify the multimodal integration process in a real-time forest fire detection and prediction system, the workflow of the proposed pipeline can be seen in the following figure:



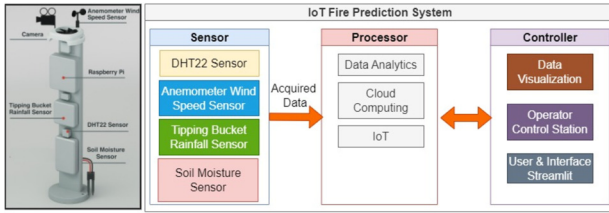
**Fig.4:** Flowchart of the Multimodal Fusion.



This figure shows the flow of a hybrid system that combines visual detection using YOLOv11 and ViT-GRU with IoT sensor data-based predictions through an LSTM model optimized using Optuna. Fire images are processed through detection and cropping stages before being classified as fire or non-fire, while environmental sensor data is classified into four fire risk levels. These two outputs are combined through a weighted decision-level fusion process to produce a final prediction of fire risk that is contextual and adaptive to real-time environmental conditions.

## 2.6 IoT Sensor Data Acquisition

IoT devices are designed to integrate various types of sensors relevant to fire risk indicators. The design of the IoT system used in this study can be seen in the image below.



**Fig.5:** IoT Device Design.

The figure above illustrates the architecture of the IoT sensor data acquisition system for real-time forest and land fire prediction. This system consists of four main sensors, namely DHT22 to measure air temperature and humidity, Anemometer Wind Speed Sensor to detect wind speed, Tipping Bucket Rainfall Sensor to measure rainfall, and Soil Moisture Sensor to detect surface soil moisture. All sensors are connected to Raspberry Pi as the main processing unit that performs data acquisition and delivery to the cloud via IoT infrastructure. The collected data is processed in an integrated manner using analytical modules and cloud computing, then the results are visualized through a Streamlit-based user interface. This system also allows operator interaction through station control for adaptive and real-time fire prediction monitoring.

## 2.7 Finalization and evaluation of the model

Model finalization and evaluation are performed to assess the overall system performance after the training and integration phases are completed. Evaluation is performed using various metrics for each model, including classification report, confusion matrix to see the distribution of correct and incorrect predictions[38]–[40], and ROC curve to measure the trade-off between true positive rate and false positive rate[41]–[43]. In the YOLOv11 model, the evaluation metric used is the mean Average Precision (mAP) at two levels, namely mAP@0.5 and mAP@0.5:0.95

which reflects the detection accuracy at various IOU thresholds. In addition, training loss and validation loss graphs are used to monitor the convergence and stability of the model during training. This evaluation ensures that the model is not only accurate, but also generalizable when applied to new data in a real environment.

## 3. RESULT AND DISCUSSION

### 3.1 Performance of fire object detection with YOLOv11

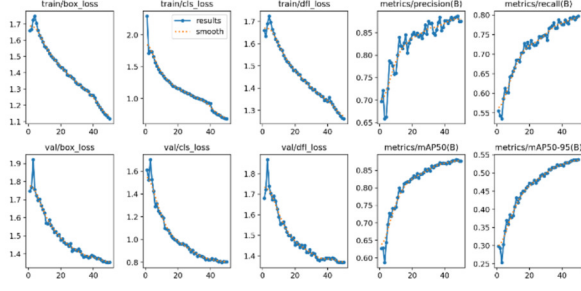
The performance evaluation of the object detection model was carried out using the mean Average Precision (mAP) metric as the main indicator of detection success at various Intersection over Union (IoU) thresholds. The mAP@0.5 and mAP@0.5:0.95 values were used to measure the accuracy of the model in detecting fire objects in the test image. The evaluation results of the YOLOv11 training process with SGD optimization for 50 epochs can be seen in the following figure:

Epoch	GPU_mem	box_loss	cls_loss	dfl_loss	Instances	Size	640	1280	1298/1298	[03:21:08:09]
48/50	1.92G	1.118	0.7816	1.275	6	mAP50	0.879	0.535	100%	04/04 [00:11]
Class		Images	Instances	Box(P)	R					
all		1329	2183	0.887	0.788					
Epoch	GPU_mem	box_loss	cls_loss	dfl_loss	Instances	Size	640	1280	1298/1298	[03:18:08:09]
49/50	1.93G	1.127	0.6878	1.264	7	mAP50	0.876	0.536	100%	04/04 [00:11]
Class		Images	Instances	Box(P)	R					
all		1329	2183	0.875	0.794					
Epoch	GPU_mem	box_loss	cls_loss	dfl_loss	Instances	Size	640	1280	1298/1298	[03:20:08:09]
50/50	1.93G	1.117	0.6837	1.262	4	mAP50	0.876	0.537	100%	04/04 [00:11]
Class		Images	Instances	Box(P)	R					
all		1329	2183	0.876	0.798					
val: Caching Images (1.562 Disks): 100%					1329/1329	[03:08:08:08]	21258	161143		
Class		Images	Instances	Box(P)	R					
all		1329	2183	0.877	0.797					
Speed: 0.1ms preprocess, 3.1ms inference, 0.1ms loss, 1.1ms postprocess per image										

**Fig.6:** YOLOv11 Training Evaluation Results Using mAP and Loss Function.

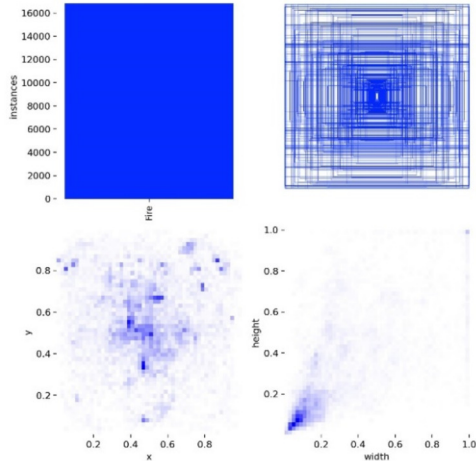
Based on the training results, the YOLOv11 model shows excellent visual detection performance, with mAP@0.5 reaching 0.877 and mAP@0.5:0.95 reaching 0.537 at the end of the 50th epoch. During the training process, the loss values of components such as box\_loss and cls\_loss consistently decreased, indicating an increase in accuracy in predicting bounding box locations and object classification. In addition, the stable dfl\_loss value in the range of 1.26 indicates the model's ability to estimate the bounding box distribution with high precision. Precision and recall were recorded at 0.876 and 0.797, respectively, indicating that the model is able to detect fire objects with a low error rate and high prediction consistency. Overall, these results indicate that the YOLOv11 model has been optimally trained and is ready to be used in the integration stage with ViT-GRU for advanced classification. The following are the evaluation results of the development of loss metrics (box\_loss, cls\_loss, dfl\_loss) and detection metrics (precision, recall, mAP50, and mAP50-95) during 50 training epochs which can be seen in the following image:

Evaluation of training data characteristics is essential to understand the spatial distribution pattern and size of objects to be detected by the model. Vi-



**Fig.7:** Loss Evaluation Curve and Detection Performance of YOLOv11 Model During Training.

sualization of the bounding box distribution and the relative position of objects in the image are used to ensure the uniformity of annotation and the adequacy of data representation against variations in the shape and location of fires. The results of the evaluation of object distribution in the training dataset can be seen in the following figure:



**Fig.8:** Distribution and Characteristics of Fire Object Bounding Boxes in the Training Dataset.

Based on the visualization in Figure 8, it can be seen that the fire objects in the dataset have a very large number of instances and only consist of one class, namely fire. The upper right image shows a fairly even distribution of bounding boxes with a concentration in the central area of the image, indicating that most of the fire objects are in the central area. The distribution of the object's center coordinates ( $x, y$ ) shows a high concentration around the middle value, while the distribution of the bounding box dimensions (width and height) indicates that most objects are relatively small to medium in size. This information indicates that the training dataset has sufficient diversity of object positions and sizes, which is very important in helping the YOLOv11 model learn to detect fire in various spatial conditions, here are the results of fire object detection:

Based on Figure 9, it can be seen that the optimized YOLOv11 model is able to detect fire ob-

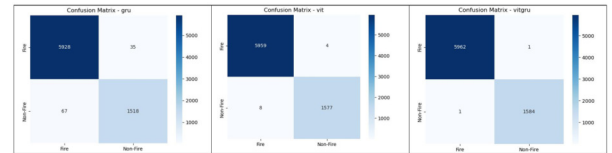


**Fig.9:** Fire Object Detection Results by Model.

jects accurately in various backgrounds, light intensities, and environmental conditions, including large flames and sparks. The resulting bounding box shows the model's ability to identify the location and size of objects with high precision. In addition, detection remains stable even in images with many fire sources or high visual complexity. This strengthens the results of the previous quantitative evaluation that YOLOv11 is effective for detecting fire objects in real-time fire monitoring systems.

### 3.2 Classification results using ViT-GRU

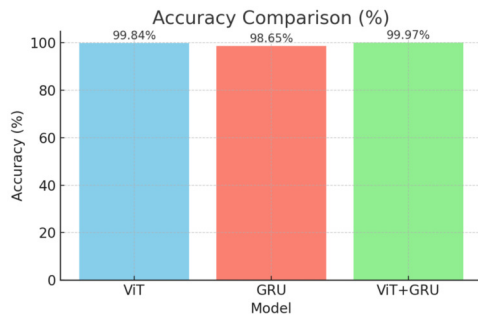
The evaluation of image classification performance between fire or non-fire was conducted to measure the accuracy and precision of detection of each model. The three models compared were GRU, ViT, and ViT-GRU integration. The evaluation results of the three models can be seen in the following figure:



**Fig.10:** Comparison of Confusion Matrix in GRU, ViT, and ViT-GRU Models.

Analysis of Figure 10 shows that the GRU model, which only relies on the temporal feature data sequence from the CNN-extracted image, produces a misclassification of 102 instances (67 false positives and 35 false negatives), indicating its limitations in accurately capturing spatial features from visual images of fires. Meanwhile, the ViT model, which utilizes the self-attention mechanism to capture global dependencies between image patches, managed to drastically reduce the misclassification to only 12 instances, demonstrating its effectiveness in recognizing complex spatial patterns such as flame shapes and smoke plumes. The ViT-GRU model, which combines the spatial representation of ViT with the sequential capabilities of GRU, managed to provide the best re-

sults, only making 2 misclassifications from a total of more than 7500 samples. This shows that the integration of GRU after ViT is able to capture sequential information between patch features formed from the ViT visual representation, such as the dynamic patterns of flames and smoke contours that repeat in the image. With very low false positives and false negatives, the ViT-GRU model is proven to be optimal in detecting fire objects precisely and can be relied on to support real-time visual early warning systems in extreme environmental conditions. Here is a comparison graph of the accuracy levels of the three models:



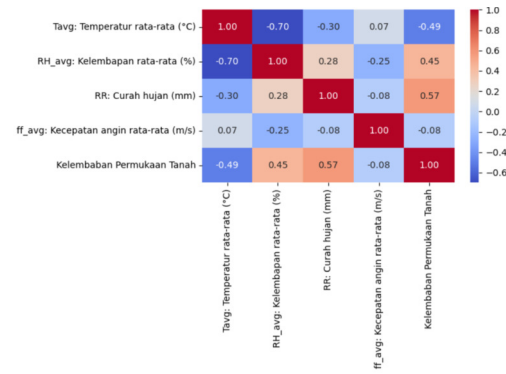
**Fig.11:** Model Accuracy Level Comparison Chart.

The figure compares the accuracy of the ViT, GRU, and ViT-GRU models in classifying fire images. The results show that the ViT-GRU model has the highest accuracy of 99.97%, followed by ViT with 99.84%, and GRU with 98.65%. This indicates that the combination of ViT and GRU is able to utilize the power of spatial representation of ViT and sequential processing of GRU synergistically, resulting in a more accurate and reliable classification than a single model.

### 3.3 Fire risk classification based on IoT sensor data using LSTM-Optuna

The meteorological dataset used in this study comes from the acquisition of IoT sensors that record temperature, humidity, rainfall, wind speed, and soil moisture variables periodically. To improve the effectiveness of fire risk prediction, a feature selection process is carried out using the LASSO method to eliminate variables that are less relevant to the target. The results of feature selection and correlation analysis between meteorological variables are visualized in the form of a heatmap, which can be seen in the following figure:

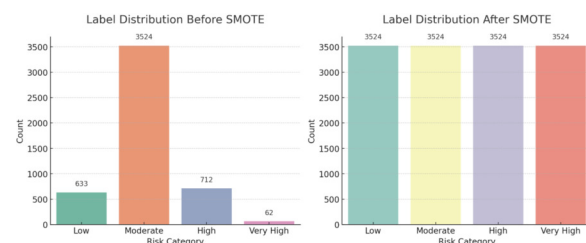
Based on the results of the correlation visualization between features using heatmaps, several statistical relationships were obtained that showed significant relationships between input variables. Tavg (average temperature) has a strong negative correlation with RH\_avg (average air humidity) of -0.70, indicating that increasing temperature is closely correlated with decreasing air humidity, a condition that is scientifically known to increase the risk of vegetation



**Fig.12:** Correlation Between Meteorological Features Based on Heatmap.

fires. In addition, the soil surface moisture variable has a positive correlation of 0.57 with rainfall (RR), indicating that rain accumulation consistently contributes to maintaining soil water content, which is an important indicator in predicting extreme drought. Another positive correlation of 0.45 between RH\_avg and soil moisture also strengthens the understanding that humid atmospheric conditions will be reflected in soil conditions. By combining this correlation analysis and the results of feature selection from LASSO (which automatically eliminates features with regression contributions close to zero), relevant features will be trained into an LSTM model optimized through the Optuna algorithm. This strategy not only minimizes overfitting due to multivariate noise, but also improves the model's generalization ability in identifying fire risk patterns based on complex weather and environmental dynamics.

The initial distribution of labels in the fire risk classification dataset showed significant class imbalance, with the Moderate category dominating with 3,524 samples, while the Very High category only had 62 samples. To overcome this problem and improve model performance on minority classes, data balancing was carried out using the SMOTE method. The distribution results before and after applying SMOTE can be seen in the following figure:

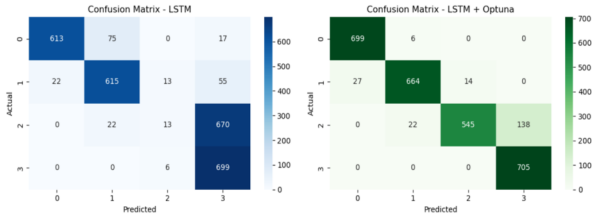


**Fig.13:** Label Distribution Before and After SMOTE Implementation.

The figure shows that before the application of SMOTE, there was a label imbalance that could cause model bias towards the majority class (Moderate). This imbalance has the potential to reduce

the model's sensitivity to the high and very high fire risk classes which are important for early detection. After SMOTE was applied, all classes were equalized into 3,524 samples, resulting in a balanced distribution between categories (Low, Moderate, High, Very High). This oversampling process not only significantly increased the proportion of the minority class but also helped the LSTM model learn more representative patterns from the entire fire risk spectrum, leading to more accurate and fair predictions.

Model performance evaluation is performed using a confusion matrix to compare the classification accuracy of the basic LSTM model with the LSTM model that has been optimized using the Optuna algorithm. The following figure presents a comparison of classification between four fire risk classes: Low (0), Moderate (1), High (2), and Very High (3):

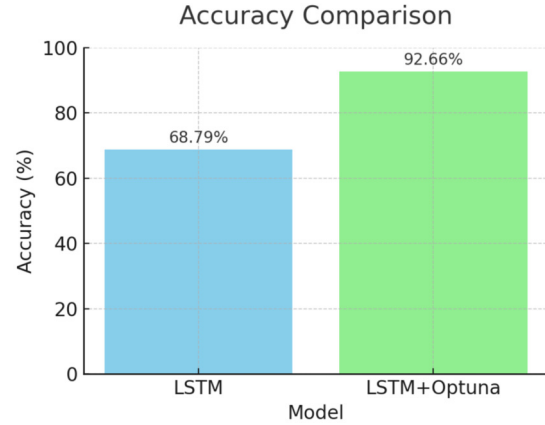


**Fig.14:** Comparison of Confusion Matrix of LSTM and LSTM + Optuna Models.

Based on the confusion matrix, the baseline LSTM model showed quite good classification performance in the Very High and High classes, but produced many misclassifications in the Low and Moderate classes, as seen from the misclassifications between classes 0 to 1 (75 cases) and 1 to 3 (55 cases). After hyperparameter optimization using Optuna, there was a significant increase in the accuracy of the Low and Moderate classes, with correct predictions reaching 699 and 664 respectively, as well as a drastic decrease in inter-class errors. Although there was an increase in misclasses in the High class (from 13 to 138), this indicates a more aggressive classification trade-off in classes with narrow distributions. Overall, LSTM+Optuna managed to improve the stability of predictions in most classes, demonstrating the effectiveness of tuning in improving model generalization to variations in IoT sensor data in fire risk classification. Here is a comparison of the accuracy levels of the two models:

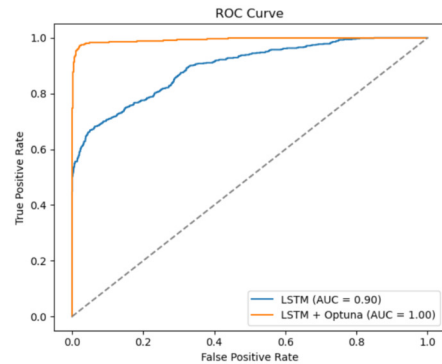
The image above shows a significant increase in accuracy of the LSTM model after optimization using Optuna, from 68.79% to 92.66%. This proves that systematic hyperparameter tuning with Optuna can substantially improve the predictive performance of the model in classifying fire risk levels based on IoT sensor data. This striking difference in accuracy emphasizes the importance of the optimization process to obtain optimal model performance.

Further evaluation of classification performance



**Fig.15:** Comparison of LSTM and LSTM+Optuna Model Accuracy.

was carried out using the Receiver Operating Characteristic (ROC) Curve to compare the ability of the LSTM and LSTM+Optuna models in distinguishing fire risk classes. The main objective of this analysis is to determine the level of model generalization, as well as detect potential overfitting in the optimized model. The results of the ROC Curve performance evaluation can be seen in the following figure:



**Fig.16:** Comparison of ROC Curves for LSTM and LSTM+Optuna Models.


Based on the ROC curve, the LSTM model produces an Area Under Curve (AUC) of 0.90, indicating a fairly good but not perfect classification performance. Meanwhile, the LSTM model that has been optimized with Optuna achieves a perfect AUC of 1.00, indicating that the model is able to separate classes very well without errors in the test data. Although  $AUC = 1.00$  is an ideal indicator, this value also needs to be further analyzed to avoid overfitting. However, because this evaluation was carried out on separate test data, and the recall and precision values were also high and consistent, it can be concluded that LSTM + Optuna did not experience significant overfitting and actually strengthened the model's generalization ability to IoT sensor-based fire risk data.

Although the optimized LSTM model achieved an



AUC score of 1.00, this result was obtained from a completely isolated test set that was not involved during training, feature selection, or oversampling. The combination of stratified hold-out validation, regularization techniques, and Optuna-based hyperparameter tuning effectively mitigated the risk of overfitting. Additionally, the consistent performance across all metrics such as accuracy, precision, recall, and ROC curve further confirms the model's ability to generalize well to unseen environmental data.

The following image shows the results of the IoT system trial developed to measure meteorological parameters directly in the field. Data from sensors such as air temperature, humidity, rainfall, wind speed, and soil moisture were successfully sent and displayed in real-time to Google Sheet. The results of the data acquisition can be seen in the following image:



	B	C	D	E	F
	Suhu Udara	Ketebapan Udara	Curah Hujan/jam	Kecapatan Angin (m/s)	Ketebapan Tanah
28	78	3	3	51	
29	70	0	3	46	
20.9	89	0	3	52	
28.4	81	0	1	30	
28.3	76	0	2	38	
26.9	83	2.3	1	35	
27.2	82	2.3	2	42	
26.5	84	4.2	1	48	
27.3	81	21.9	1	71	
28.4	73	0	1	64	
26.4	86	53	4	86	
29.1	71.1	0	2	70	
25	80	50	2	70	

**Fig.17:** Field Testing and Real-Time Meteorological Data Acquisition Through IoT.

This test proves the successful integration between IoT hardware and Google Sheets-based cloud system for real-time meteorological data acquisition. The transmitted data includes five main variables that are highly relevant for fire risk prediction: air temperature, air humidity, hourly rainfall, wind speed, and soil surface moisture. The variability of the data recorded at each time shows that the sensor is able to capture environmental dynamics accurately and continuously.

### 3.4 Multimodal fusion results and final decision

The following figure shows the results of the development of a Streamlit-based interface system that integrates the LSTM-Optuna model with real-time IoT sensor data. Data is automatically sent from IoT devices to Google Sheets and accessed directly by the system to generate fire risk predictions. The appearance of the developed system can be seen in the following figure:

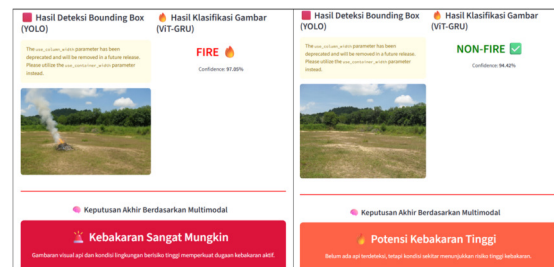
This dashboard is the result of the implementation of a fire risk prediction model based on meteorological data using an LSTM architecture that has been optimized with the Optuna algorithm. Data such as temperature, air humidity, rainfall, wind speed, and soil moisture are sent in real-time from IoT devices to Google Sheets, then used as prediction input. The classification results are displayed in four risk categories: Low, Moderate, High, and Very High, which



**Fig.18:** Streamlit Interface for LSTM and IoT-Based Fire Risk Prediction.

are displayed dynamically along with the latest sensor values. With this interface, the system is able to provide risk notifications directly to users without the need for manual intervention, indicating the system's readiness to be used in preventive and real-time field monitoring scenarios.

The following shows the interface of the results of the implementation of the fire detection and classification model based on IoT-based visual and environmental data. The system uses YOLO to detect the presence of fire objects spatially, and ViT-GRU to classify images into fire or non-fire based on the extracted visual features. The final result is determined based on the fusion with fire risk predictions from IoT sensors using the LSTM model, as can be seen in the following figure:



**Fig.19:** Visual Detection System Interface (YOLO + ViT-GRU) and Multimodal-Based Decision Finalization.

In the left view, YOLO successfully detected the presence of fire in the field image, and the ViT-GRU classification result shows the fire label with a confidence level of 97.06%. On the other hand, in the right view, the system did not detect the presence of fire, and the ViT-GRU classification result indicates non-fire with a confidence level of 94.42%. Both of these processes show that ViT-GRU is effective in distinguishing fire objects based on visual features, and is able to maintain high accuracy in both positive and negative cases. The combination of YOLO and ViT-GRU provides a fast and reliable image-based classification basis.

Next, the system performs multimodal fusion with the results of the fire risk level prediction from IoT data using LSTM-Optuna. In the test shown, the sensor prediction shows a High risk level, which is

then combined with the visual classification results. The final fusion result is displayed in the form of a more contextual decision, such as Fire is Very Likely if visual detection and sensor risk are mutually reinforcing, or High Fire Potential if only one component is active. This approach increases the reliability of the system in providing warnings that are not only based on visual detection alone, but also consider the dynamics of the microenvironment in real-time.

A performance comparison of YOLO-based object detection models was conducted to evaluate the superiority of the YOLOv11 model developed in this study over previous versions, namely YOLOv6 to YOLOv10. All models were tested using the same fire dataset with similar training parameters to ensure test objectivity. The evaluation was conducted based on four main metrics: Precision, Recall, mAP@0.5, and mAP@0.5:0.95, which represent detection accuracy at various Intersection over Union (IoU) thresholds. The following table contains the results of testing the YOLO version with the same dataset:

**Table 1:** Comparison of YOLO Model Performance With the Same Dataset.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv6	72.3%	68.5%	85.6%	0.480
YOLOv7	75.8%	70.4%	86.4%	0.487
YOLOv8	80.6%	74.1%	87.1%	0.495
YOLOv9	84.7%	77.5%	87.4%	0.509
YOLOv10	83.2%	76.8%	86.9%	0.531
Research Model (YOLOv11)	87.9%	79.7%	87.7%	0.537

Based on the evaluation results in Table 1, the YOLOv11 model shows the best performance with a precision of 87.9%, recall of 79.7%, mAP@0.5 of 87.7%, and mAP@0.5:0.95 of 0.537. All of these metric values are higher than the previous version of YOLO. YOLOv10, which is the latest version before YOLOv11, only achieved mAP@0.5 of 86.9% and mAP@0.5:0.95 of 0.531. This indicates that this research model is not only superior in detecting fire objects precisely, but also more stable in recognizing variations in the size and position of objects in the image. These advantages strengthen the validity of YOLOv11 as the foundation for early detection in the multimodal system proposed in this study.

Table 2 presents a performance comparison of different ViT models combined with different deep learning algorithms for visual classification.

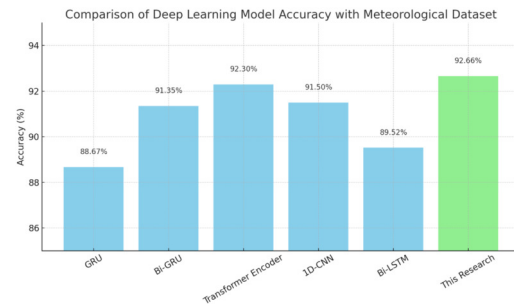
Based on Table 2, the ViT+GRU model proposed in this study shows the highest accuracy of 99.97%, outperforming all other comparison models such as ViT-CNN (99.61%), ViT-UNet (93.50%), and ViT-LSTM (91.15%). This shows that the integration of ViT architecture with GRU is able to capture the spatial representation of visuals in depth through self-attention and strengthen it with GRU's ability to understand temporal relationships between

**Table 2:** Performance Comparison of ViT Model with Deep Learning Model.

Researcher	Year	Model	Accuracy
[44]	2022	VIT-Bi-LSTM	86.67%
[45]	2024	ViT-LSTM	91.15%
[46]	2025	ViT-RNN	92.10%
[47]	2024	ViT-UNet	93.50%
[48]	2025	ViT-CNN	99.61%
(Research)	2025	ViT+GRU	99.97%

extracted visual features. Compared to other approaches that rely solely on feed-forward networks or convolutional networks after ViT, the use of GRU provides significant advantages in capturing the dynamics of information sequences in images that may contain complex spatial variations such as smoke, sparks, and unstable lighting. This combination makes ViT+GRU highly reliable for image-based fire classification tasks.

Figure 19 presents the results of the performance comparison of various deep learning models in the classification of fire risk levels based on meteorological data. Testing was carried out on the same dataset using the accuracy metric as the main indicator of model performance.



**Fig.20:** Comparison of Deep Learning Model.

Based on the results in Figure 20, the LSTM model optimized using Optuna in this study achieved the highest accuracy of 92.66%, outperforming other models such as Transformer Encoder (92.30%) and 1D-CNN (91.50%). The GRU and Bi-LSTM models performed quite well with accuracies of 88.67% and 89.52% respectively, but were still below the research model. The superiority of this model lies not only in the LSTM architecture itself, but also in the application of hyperparameter optimization process using Optuna and data balancing technique (SMOTE), which significantly improves the generalization ability to unbalanced data. These results confirm that the combination of feature engineering, auto-tuning, and LSTM-based temporal modeling provides more accurate classification results for meteorological data in the context of fire risk prediction.

### 3.5 Practical Implementation Considerations

In addition to accuracy metrics, practical deployment aspects of the proposed system were also evaluated. The real-time Streamlit dashboard demonstrated an average response time of approximately 3–5 seconds from sensor data acquisition to the final risk prediction output, which is sufficient for early warning purposes in forest fire monitoring. The system architecture is designed to be scalable, allowing integration of additional IoT nodes across larger geographic areas and extension with drone or satellite imagery to further enhance coverage. From a cost perspective, each IoT unit including Raspberry Pi, DHT22 temperature and humidity sensor, soil moisture sensor, tipping bucket rainfall sensor, and anemometer was estimated to cost approximately USD 800 per unit. This relatively low hardware cost highlights the feasibility of deploying multiple units in parallel to establish a distributed fire monitoring network in vulnerable regions.

## 4. CONCLUSIONS

This research successfully developed a multimodal forest fire detection and prediction system that integrates Computer Vision and Internet of Things (IoT) technologies. The YOLOv11 model is used to detect the presence of fire in the image, followed by a visual classification process using a combination of Vision Transformer (ViT) and Gated Recurrent Unit (GRU), which is proven to produce the highest accuracy of 99.97%. Meanwhile, the Long Short-Term Memory (LSTM) model optimized with Optuna showed superior performance in classifying fire risk levels based on meteorological sensor data, with an accuracy of 92.66% and an AUC of 1.00. Multimodal fusion is performed at the decision-level to obtain adaptive and contextualized final predictions, enabling the system to provide more precise and real-time fire warnings. The system has also been implemented in the form of a Streamlit dashboard integrated with Google Sheets to display IoT sensor data in real-time, proving its feasibility to be applied in the field as an automatic monitoring system and responsive to environmental conditions.

Future research can focus on developing a stacked ensemble method to improve the performance of fire risk level prediction. By combining various deep learning and machine learning models in a stacked ensemble architecture, the system is expected to produce more robust and generalizable predictions. In addition, exploration of feature-level fusion and utilization of satellite or drone imagery data can further expand the scope and precision of the system in real scenarios.

## ACKNOWLEDGEMENT

This research is funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology with the Regular Fundamental Research Scheme in 2025, in accordance with Research Contract Number: 010/LL17/DT.05.00/PL/2025.

## AUTHOR CONTRIBUTIONS

Conceptualization, A.M.; methodology, E.; image processing, R.W.; IoT integration, Y.I.; formal analysis, E.; investigation, W.; data curation, A.M.; writing—original draft preparation, A.M. and Y.I.; writing—review and editing, E., W., and R.W. All authors have read and agreed to the published version of the manuscript.

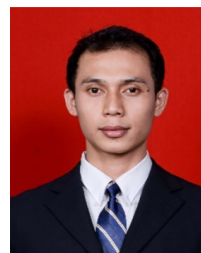
## References

- [1] Y. Irawan *et al.*, “Realtime Monitoring and Analysis Based on Cloud Computing Internet of Things (CC-IoT) Technology in Detecting Forest and Land Fires in Riau Province,” *Ilk. J. Ilm.*, vol. 15, no. 3, pp. 445–454, 2023.
- [2] B. Setya, R. A. Nurhidayatullah, M. B. Hewen and K. Kusriani, “Comparative Analysis Of Rainfall Value Prediction In Semarang Using Linear And K-Nearest Neighbor Algorithms,” in *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1–5, 2023.
- [3] A. L. R. Putri, B. Surarso and T. U. SRRM, “MICE Implementation to Handle Missing Values in Rain Potential Prediction Using Support Vector Machine Algorithm,” *JTAM (Jurnal Teor. dan Apl. Mat.)*, vol. 7, no. 4, p. 1167, 2023.
- [4] Y. Irawan, R. Wahyuni, M. Muhandi, H. Fonda, M. L. Hamzah and R. Muzawi, “Real Time System Monitoring and Analysis-Based Internet of Things (IoT) Technology in Measuring Outdoor Air Quality,” *Int. J. Interact. Mob. Technol.*, vol. 15, no. 10, pp. 224–240, 2021.
- [5] R. Wahyuni, Herianto, Ikhtiyaruddin, and Y. Irawan, “IoT-Based Pulse Oximetry Design as Early Detection of Covid-19 Symptoms,” *Int. J. Interact. Mob. Technol.*, vol. 17, no. 3, pp. 177–187, 2023.
- [6] Y. Irawan, E. Sabna, A. F. Azim, R. Wahyuni, N. Belarbi and M. M. Josephine, “Automatic Chili Plant Watering Based on Internet of Things (IoT),” *J. Appl. Eng. Technol. Sci.*, vol. 3, no. 2, pp. 77–83, 2022.
- [7] Y. Irawan, A. W. Novrianto and H. Sallam, “Cigarette Smoke Detection and Cleaner Based on Internet of Things (IoT) Using Arduino Microcontroller and MQ-2 Sensor,” *J. Appl. Eng. Technol. Sci.*, vol. 2, no. 2, pp. 85–93, 2021.
- [8] A. R. Abidin, Y. Irawan and Y. Devis, “Smart

- Trash Bin for Management of Garbage Problem in Society,” *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 202–208, 2022.
- [9] S. Purwanti, A. Febriani, M. Mardeni and Y. Irawan, “Temperature Monitoring System for Egg Incubators Using Raspberry Pi3 Based on Internet of Things (IoT),” *J. Robot. Control*, vol. 2, no. 5, 2021.
- [10] R. M. Sari, E. Sabna, R. Wahyuni and Y. Irawan, “Implementation of open and close a housing gate portal using RFID card,” *J. Robot. Control*, vol. 2, no. 5, pp. 363–367, Sep. 2021.
- [11] A. N. Mabdeh, A. Al-Fugara, K. M. Khedher, M. Mabdeh, A. R. Al-Shabeeb and R. Al-Adamat, “Forest Fire Susceptibility Assessment and Mapping Using Support Vector Regression and Adaptive Neuro-Fuzzy Inference System-Based Evolutionary Algorithms,” *Sustainability*, vol. 14, no. 15, 2022.
- [12] A. S. Nur, Y. J. Kim, J. H. Lee and C.-W. Lee, “Spatial Prediction of Wildfire Susceptibility Using Hybrid Machine Learning Models Based on Support Vector Regression in Sydney, Australia,” *Remote Sens.*, vol. 15, no. 3, 2023.
- [13] À. Nebot and F. Mugica, “Forest Fire Forecasting Using Fuzzy Logic Models,” *Forests*, vol. 12, no. 8, 2021.
- [14] L. Soualah, A. Bouzekri, and H. Chenchouni, “Hoping the best, expecting the worst: Forecasting forest fire risk in Algeria using fuzzy logic and GIS,” *Trees, For. People*, vol. 17, p. 100614, 2024.
- [15] S. Madkar, D. Y. Sakhare, K. A. Phutane, A. P. Haral, K. B. Nikam and S. Tharunya, “Video Based Forest Fire and Smoke Detection Using YoLo and CNN,” *3rd Int. Conf. Power, Energy, Control Transm. Syst. ICPECTS 2022 - Proc.*, pp. 1–5, 2022.
- [16] A. Bayegizova *et al.*, “Fire detection using deep learning methods,” *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, p. 547, 2024.
- [17] S. N. Saydirasulovich, M. Mukhiddinov, O. Djuraev, A. Abdusalomov and Y. I. Cho, “An Improved Wildfire Smoke Detection Based on YOLOv8 and UAV Images,” *Sensors (Basel)*, vol. 23, no. 20, 2023.
- [18] A. Sharma, V. Kumar and L. Longchamps, “Comparative performance of YOLOv8, YOLOv9, YOLOv10, YOLOv11 and Faster R-CNN models for detection of multiple weed species,” *Smart Agric. Technol.*, vol. 9, p. 100648, 2024.
- [19] M. A. Şimşek and A. Sertbaş, “Automatic Detection Of Meniscus Tears From Knee Mri Images Using Deep Learning: Yolo V8, V9, And V10 Series,” *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilim. Derg.*, vol. 28, no. 1, pp. 292–308, 2025.
- [20] P. Mehta *et al.*, “Benchmarking YOLO Variants for Enhanced Blood Cell Detection,” *Int. J. Imaging Syst. Technol.*, vol. 35, no. 1, p. e70037, 2025.
- [21] S. N. R. Rajakrishnamoorthi, S. G. V. S. S. M. A. R. and C. H. Madhavan, “Smoke Sentry: Employing YOLO-based Deep Learning for Sophisticated Real-Time Smoking (Cigarette smoking) Detection in Surveillance Networks,” in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 892–898, 2024.
- [22] C. Catargiu, N. Cleju and I. B. Ciocoiu, “A Comparative Performance Evaluation of YOLO-Type Detectors on a New Open Fire and Smoke Dataset,” *Sensors*, vol. 24, no. 17, 2024.
- [23] F. Akhmedov, R. Nasimov and A. Abdusalomov, “Dehazing Algorithm Integration with YOLOv10 for Ship Fire Detection,” *Fire*, vol. 7, no. 9, pp. 1–20, 2024.
- [24] A. Morchid, Z. Oughannou, R. El Alami, H. Qjidaa, M. O. Jamil, and H. M. Khalid, “Integrated internet of things (IoT) solutions for early fire detection in smart agriculture,” *Results Eng.*, vol. 24, p. 103392, 2024.
- [25] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov and N. Pacal, “A novel CNN-ViT-based deep learning model for early skin cancer diagnosis,” *Biomed. Signal Process. Control*, vol. 104, p. 107627, 2025.
- [26] Y. Pan, Y. Li, T. Yao, C. W. Ngo and T. Mei, “Stream-ViT: Learning Streamlined Convolutions in Vision Transformer,” *IEEE Trans. Multimed.*, vol. PP, pp. 1–11, 2025.
- [27] G. Wang, D. Bai, H. Lin, H. Zhou and J. Qian, “FireViTNet: A hybrid model integrating ViT and CNNs for forest fire segmentation,” *Comput. Electron. Agric.*, vol. 218, p. 108722, 2024.
- [28] T. Chen *et al.*, “A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images,” *Healthc. Anal.*, vol. 5, p. 100332, 2024.
- [29] J. Li and Y. Zhang, “Regressive vision transformer for dog cardiomegaly assessment,” *Sci. Rep.*, vol. 14, no. 1, p. 1539, 2024.
- [30] M. Shahid, H.-C. Wang, Y.-Y. Chen, and K.-L. Hua, “Hybrid CNN-ViT architecture to exploit spatio-temporal feature for fire recognition trained through transfer learning,” *Multimed. Tools Appl.*, vol. 84, no. 8, pp. 4703–4732, 2025.
- [31] S. S. Rajashree, S. C. M. A. S. M. M. K. and S. K. S., “Hybrid CNN-GRU based Machine Vision Multimodal Deep Learning Model for Glaucoma Detection in Retinal Fundus Image,” in *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 913–919, 2025.
- [32] U. S. Kumar *et al.*, “Fusion of MobileNet and



- GRU: Enhancing Remote Sensing Applications for Sustainable Agriculture and Food Security,” *Remote Sens. Earth Syst. Sci.*, vol. 8, no. 1, pp. 118–131, 2025.
- [33] M. J. Zobair, M. A. Rahman, M. S. Hossain, N. Khan, M. A. A. K. Akash and M. H. I. Bijoy, “A Hybrid ViT-GRU Model for Breast Cancer Detection: Addressing Class Imbalance Challenges,” in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2025.
- [34] Y. Lee and H. Kim, “Comparative Analysis of YOLO Series ( from V1 to V11 ) and Their Application in Computer Vision,” *J. Semicond. Disp. Technol.*, vol. 23, no. 4, pp. 190–198, 2024.
- [35] A. Ghahremani, S. D. Adams, M. Norton, S. Y. Khoo and A. Z. Kouzani, “Detecting Defects in Solar Panels Using the YOLO v10 and v11 Algorithms,” *Electron.*, vol. 14, no. 2, 2025.
- [36] S. Wang, Y. Chen, Z. Cui, L. Lin and Y. Zong, “Diabetes Risk Analysis based on Machine Learning LASSO Regression Model,” *J. Theory Pract. Eng. Sci.*, vol. 4, no. 1, pp. 58–64, 2024.
- [37] V. A. Boateng and B. Yang, “A Global Modeling Pruning Ensemble Stacking With Deep Learning and Neural Network Meta-Learner for Passenger Train Delay Prediction,” *IEEE Access*, vol. 11, no. May, pp. 62605–62615, 2023.
- [38] H. Fonda, Y. Irawan, R. Melyanti, R. Wahyuni and A. Muhaimin, “A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1701–1714, 2024.
- [39] A. Lubis, Y. Irawan, Junadhi and S. Defit, “Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1625–1638, 2024.
- [40] M. K. Anam *et al.*, “Sara Detection on Social Media Using Deep Learning Algorithm Development,” *J. Appl. Eng. Technol. Sci.*, vol. 6, no. 1, pp. 225–237, Dec. 2024.
- [41] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, “Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction,” *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1272–1285, 2024.
- [42] A. Febriani, R. Wahyuni, Y. Irawan, and R. Melyanti, “Improved Hybrid Machine and Deep Learning Model for Optimization of Smart Egg Incubator,” *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1052–1068, 2024.
- [43] D. Setiawan, R. N. Putri, I. Fitri, A. N. Hidayanto, Y. Irawan, and N. Hohashi, “Improved Deep Learning Model for Prediction of Dermatitis in Infants,” *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 871–884, 2025.
- [44] H. Chen, J. Cui, Y. Zhang, and Y. Zhang, “ViT and Bi-LSTM for Micro-Expressions Recognition,” in *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 946–951, 2022.
- [45] A. B. Nassif, I. Shahin, M. Bader, A. Ahmed, and N. Werghi, “ViT-LSTM synergy: a multi-feature approach for speaker identification and mask detection,” *Neural Comput. Appl.*, vol. 36, no. 35, pp. 22569–22586, 2024.
- [46] A. R. Borah, A. A. Hameed, H. P. Thethi, J. L. Prasanna, A. Sangeetha and D. D. Gautam, “ViT and RNN for Temporal and Spatial Analysis in Video Sequences,” in *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*, pp. 651–656, 2025.
- [47] N. Zhou *et al.*, “ViT-UNet: A Vision Transformer Based UNet Model for Coastal Wetland Classification Based on High Spatial Resolution Imagery,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 19575–19587, 2024.
- [48] M. S. Wasif, M. P. Miah, M. S. Hossain, M. J. F. Alenazi and M. Atiquzzaman, “CNN-ViT synergy: An efficient Android malware detection approach through deep learning,” *Comput. Electr. Eng.*, vol. 123, p. 110039, 2025.



**Abdi Muhaimi** is a lecturer at the Information Systems undergraduate program (S1) at Universitas Hang Tuah Pekanbaru, Indonesia. He earned his Bachelor's degree in Computer Science from STMIK Hang Tuah Pekanbaru and his Master's degree in Computer Science from Universitas Putra Indonesia YPTK Padang. His research interests focus on the application of Artificial Intelligence in business and enterprise.



**Edriyansyah** is a lecturer at the Informatics Engineering undergraduate program (S1) at Universitas Hang Tuah Pekanbaru, Indonesia. He obtained both his Bachelor's and Master's degrees in Computer Science from Universitas Putra Indonesia YPTK Padang. His research interests are centered on computer networks.



**Wahyat** is a lecturer at the Computer Network Administration program at Politeknik Negeri Bengkalis, Indonesia. He earned his Bachelor's degree in Computer Science from STMIK Amik Riau and his Master's degree in Computer Science from Universitas Putra Indonesia YPTK Padang. His research interests revolve around computer network administration.



**Yuda Irawan** is a lecturer at the Informatics Engineering undergraduate program (S1) at Universitas Hang Tuah Pekanbaru, Indonesia. He completed his Bachelor's degree in Computer Science at STMIK Amik Riau and his Master's degree in Informatics Engineering at Bina Nusantara University. He is currently pursuing a doctoral degree (Ph.D.) in Information Technology at Universitas Putra Indonesia YPTK

Padang. His research interests include the Internet of Things (IoT), Machine Learning, and Deep Learning.



**Refni Wahyuni** is a lecturer at the Informatics Engineering undergraduate program (S1) at Universitas Hang Tuah Pekanbaru, Indonesia. She obtained her Bachelor's degree in Computer Science from STMIK Amik Riau and her Master's degree in Informatics Engineering from Bina Nusantara University. Her research interests include the Internet of Things (IoT) and Image Processing.