Dual-Stream MobileNetV2 and Light Mixer Fusion for Robust Weather Classification

Muhammad Reza Al Fatah [1], Hadi Santoso [2,*]
[1, 2] Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
Email: [1] muhammadrezaalfatah09@gmail.com, [2] hadi.santoso@mercubuana.ac.id
*Corresponding Author

***Abstract*—This study proposes an image-based weather classification model designed to accurately recognize various sky conditions, with the aim of providing accessible weather information for elderly individuals and users with visual impairments. While existing lightweight models often struggle to effectively capture both fine-grained local textures and global contextual patterns, this study bridges the gap by proposing a hybrid dual-branch architecture. Specifically, the proposed model utilizes an early spatial feature-level fusion dual-branch architecture. The first branch combines MobileNetV2 with a Feature Pyramid Network (FPN) and incorporates selective attention mechanisms to capture multi-scale features. The second branch, referred to as the Mixer branch, improves visual feature representation through patch embedding and feature mixing techniques. Outputs from both branches are integrated using a fusion layer before being processed by a softmax classifier. The dataset includes five weather categories: cloudy, foggy, rainy, sunny, and sunrise, and is preprocessed through normalization, data augmentation, and partitioning into training, validation, and testing sets. Model training is conducted using TensorFlow and Keras with the Adam optimizer over a two-phase training schedule of 60 epochs (20 epochs for head-only pre-training and 40 epochs for whole-model fine-tuning). The experimental evaluation achieves a test accuracy of 0.975 (97.50%), with precision, recall, and F1-score reaching 0.976, 0.975, and 0.975, respectively, reflecting consistent and reliable classification performance. These results indicate that the proposed model has strong potential for integration with text-to-speech systems to improve accessibility of weather information for users with special needs.**

Keywords*—*MobileNetV2; Weather Classification; Mixer Branch; Two-Branch Fusion; FPN; Selective Attention

I. INTRODUCTION

Weather represents a fundamental environmental factor that plays a crucial role in shaping daily human activities. Variations in temperature, precipitation, and humidity can significantly influence productivity, health, and overall safety [1]. In tropical regions, where weather patterns tend to fluctuate rapidly, communities face unique challenges, particularly in sectors such as agriculture, transportation, and outdoor activities [1]. The 2023 Lancet Countdown report highlights the economic and societal impact of such conditions, noting that heat exposure alone caused global income losses estimated at US$863 billion in 2022, with tropical regions experiencing a disproportionate decline in labor capacity [2]. Vulnerable populations, including older adults and individuals with visual impairments, are especially at risk, as their limited ability to visually perceive environmental changes reduces their capacity to respond effectively to weather conditions [3]. Moreover, the increasing frequency of extreme weather events due to global climate change has intensified these risks. Sudden heavy rainfall, extreme heat, and significant daily temperature variations have become more prevalent in tropical areas [4]. These environmental changes adversely affect comfort and safety, particularly for individuals who face difficulties adapting to rapidly changing conditions [4].

Exposure to extreme weather also has direct consequences for human health. Prolonged heat in tropical climates has been associated with increased incidences of respiratory illnesses, dehydration, and cardiovascular diseases, particularly among elderly populations [5]. Similarly, outdoor workers are more susceptible to heat stress and occupational accidents due to unpredictable weather patterns [6]. Therefore, access to timely and accurate weather information is essential, especially for vulnerable groups. Inclusive environmental design offers a potential solution by reducing barriers for individuals with disabilities in adapting to environmental conditions. In this context, sensor-based and Artificial Intelligence (AI) technologies can be leveraged to automatically detect weather conditions and deliver information through audio or haptic feedback, thereby enhancing accessibility and promoting human-centered technological solutions.

Recent advancements in artificial intelligence, particularly in deep learning, have significantly accelerated progress in computer vision, enabling the development of such assistive systems. Among these approaches, Convolutional Neural Networks (CNNs) have emerged as a dominant method due to their ability to perform hierarchical and automated feature extraction directly from raw image data [7]. Early studies utilized conventional CNN architectures such as LeNet-5 and VGG-16. For instance, research by [8] demonstrated that a modified LeNet-5 model achieved an accuracy of up to 94% in weather classification tasks. Similarly, [9] employed the VGG-16 architecture and reported an accuracy of 96.87%. Despite their strong performance, models such as VGG-16 are computationally intensive due to their large number of parameters, making them less suitable for deployment on resource-constrained devices.

To improve efficiency, subsequent studies have explored lightweight CNN architectures. MobileNetV2, in particular, has gained popularity due to its use of depthwise separable convolutions, which significantly reduce parameter size while maintaining performance. This architecture has been successfully applied in domains such as remote sensing [10] and medical image classification [11]. In weather

classification tasks, comparative studies by [12] and [13] found that EfficientNet achieved superior performance, with accuracy reaching 96.75%. Additionally, [14] demonstrated the effectiveness of transfer learning using EfficientNetV2S, achieving an accuracy of 92.35%. However, several studies have highlighted limitations in the generalization capability of CNN-based models when applied to real-world datasets with high visual variability [15].

To address these limitations, various studies have incorporated Feature Pyramid Networks (FPN) to enhance multi-scale feature representation. For example, [12] showed that combining MobileNetV2 with FPN significantly improved precision and recall. Similar strategies have also been applied to non-CNN architectures, such as the Swin Transformer [16], where the integration of FPN with a Selective Attention Module enabled the model to focus on the most relevant visual features. The use of attention mechanisms has proven effective in improving feature selection, as demonstrated in recent studies employing multi-head attention for enhanced pattern recognition in medical imaging tasks.

In addition, recent research has explored alternative architectures such as the MLP-Mixer introduced by [17], which enables the modeling of global dependencies across image patches by facilitating information exchange through token-mixing and channel-mixing operations, alongside Transformer-based architectures. Fusion strategies, particularly early spatial feature-level fusion techniques, have also been proposed to combine complementary feature representations and improve classification performance.

Overall, previous studies demonstrate the strong potential of CNNs and modern deep learning architectures in image-based classification tasks. However, most existing approaches rely on a single architectural paradigm and do not fully integrate the strengths of CNN-based spatial feature extraction with the global contextual modeling capabilities of non-CNN architectures. Addressing this gap and in response to the growing demand for inclusive technologies, this study proposes an image-based weather classification system utilizing a two-branch fusion architecture. The system is designed to deliver accurate and efficient weather recognition and to be deployable on portable devices for visually impaired and elderly users.

The novelty of this research is reflected in three primary contributions. First, it introduces a hybrid two-branch fusion architecture that combines the efficiency of MobileNetV2 with a Light Mixer mechanism, enabling the model to capture both fine-grained local features (such as rain textures) and broader global patterns (such as lighting conditions) while maintaining computational efficiency compared to conventional Vision Transformers. Second, the integration of a Feature Pyramid Network (FPN) with Selective Attention is proposed to address the challenge of distinguishing visually similar weather categories, such as cloudy and foggy conditions, which remains a common limitation in lightweight models.

Finally, this study bridges the gap between computer vision and inclusive, human-centered design. By translating classification outputs into audio-based feedback, the proposed system provides a practical, real-time solution to enhance safety and accessibility for visually impaired and elderly individuals, directly responding to the increasing need for adaptive technologies in the context of global climate change.

II. Conceptual Framework

The proposed conceptual framework is designed to address the challenges of accurate and efficient weather classification from visual data, particularly for deployment in real-time and resource-constrained environments. The framework is structured into three main stages: data processing, feature learning, and decision-level integration, as illustrated conceptually through a multi-stage learning pipeline.

In the first stage, input images representing various weather conditions undergo preprocessing to ensure data consistency and improve model generalization. This includes normalization, resizing, and data augmentation techniques such as rotation, flipping, and brightness adjustment. These steps aim to reduce overfitting and enhance the robustness of the model when exposed to diverse real-world scenarios.

The second stage focuses on feature learning through a dual-representation strategy. Instead of relying on a single model architecture, this framework separates feature extraction into two complementary pathways. The first pathway emphasizes local spatial feature extraction using a lightweight Convolutional Neural Network (CNN), specifically MobileNetV2. This branch is further enhanced by incorporating a Feature Pyramid Network (FPN), which enables the extraction of multi-scale features, and a selective attention mechanism that prioritizes the most informative regions within the image [10]-[16]. This pathway is particularly effective in capturing fine-grained visual details such as cloud textures, raindrop patterns, and lighting variations.

The second pathway focuses on capturing global contextual information. This is achieved through a lightweight Mixer-based architecture that models long-range dependencies across image patches. By utilizing token-mixing and channel-mixing operations, this branch facilitates interaction between spatial regions and feature channels, enabling the model to understand overall scene composition and atmospheric conditions [18]. This global perspective complements the local feature extraction performed by the CNN-based branch.

In the final stage, the framework integrates the outputs from both pathways using an early spatial feature-level fusion mechanism. An early spatial feature-level fusion strategy is employed to concatenate the 2D feature representations from each branch before Global Average Pooling and classification. This approach allows the model to leverage complementary information from both local spatial representations and global context, resulting in improved robustness and classification accuracy. The fused and pooled features are then passed through a batch normalization layer, followed by a fully connected layer and a softmax classifier to produce the final weather class prediction.

Overall, this conceptual framework highlights the integration of efficient local feature extraction and global contextual modeling within a unified architecture. By combining these complementary approaches, the proposed system aims to achieve a balance between computational efficiency and classification performance. Additionally, the

framework is designed with practical deployment in mind, enabling its integration into portable assistive devices that can provide real-time weather information through accessible interfaces such as audio feedback.

III. Literature Review

Recent advancements in computer vision and deep learning have significantly improved the capability of automatic weather recognition systems based on visual data. Image-based

Classification approaches are particularly effective due to their ability to extract meaningful patterns from environmental scenes. Among these approaches, Convolutional Neural Networks (CNNs) have been widely adopted due to their capability to learn hierarchical feature representations directly from raw image inputs [7].

Despite these advancements, most existing studies focus on single-architecture models or employ fusion strategies without considering coHowever, conventional CNN architectures often face challenges in balancing accuracy and computational efficiency, particularly when deployed in real-time systems or resource-constrained devices. Lightweight architectures such as MobileNetV2 have been introduced to address this issue by utilizing depthwise separable convolutions and inverted residual blocks, allowing efficient feature extraction with reduced computational cost. Despite their efficiency, CNN-based models primarily focus on local spatial features and may struggle to capture global contextual relationships present in complex weather scenes. Imputational efficiency and deployment constraints. There remains a research gap in designing a unified architecture that integrates lightweight CNNs with global-context modeling mechanisms while maintaining low computational overhead. This study addresses this gap by proposing a two-branch fusion model that combines MobileNetV2 with FPN and selective attention in one branch and a lightweight Mixer mechanism in the other branch, integrated through an early spatial feature-level fusion strategy to enhance weather image classification performance.

To overcome these limitations, this study incorporates a Feature Pyramid Network (FPN) to enhance multi-scale feature representation. FPN enables the aggregation of feature maps from different levels of the network, allowing the model to capture both low-level details and high-level semantic information simultaneously [12]. Additionally, a selective attention mechanism is integrated to improve the model's ability to focus on the most relevant visual regions, thereby enhancing discriminative performance in challenging classification scenarios [16].

In parallel, recent research has explored alternative architectures that emphasize global context modeling. Approaches such as the MLP-Mixer and Transformer-based models are capable of learning long-range dependencies by enabling interactions across image patches through token-mixing and channel-mixing mechanisms [18]. These methods complement CNN-based approaches by providing a broader understanding of global visual patterns. However, their high computational requirements often limit their practical deployment in mobile or embedded systems.

To leverage the strengths of both local and global feature representations, this study adopts a two-branch fusion framework. The first branch is based on MobileNetV2 enhanced with FPN and selective attention, focusing on efficient extraction of local and multi-scale features. The second branch employs a lightweight Mixer-based mechanism to capture global contextual relationships across image patches. The outputs from both branches are then combined using an early spatial feature-level fusion strategy, which concatenates complementary 2D feature representations spatially before Global Average Pooling and classification.

This conceptual framework is designed to address the limitations of single-architecture models by providing a balanced approach between computational efficiency and representational capability. By integrating local feature extraction with global context modeling, the proposed framework aims to improve classification performance in diverse and complex weather conditions. Furthermore, the lightweight design ensures compatibility with portable devices, making the system suitable for real-world applications, particularly for assisting visually impaired and elderly users in accessing weather information.

A. Dataset and Data Preprocessing

The dataset in Fig. 1 used in this study was obtained from the Kaggle platform and contains weather images categorized into five classes: cloudy, foggy, rainy, sunny, and sunrise. The dataset (consisting of 1,500 images) was partitioned into training (1,080 images, 72%), validation (120 images, 8%), and testing (300 images, 20%) subsets. All images underwent a preprocessing step that included resizing to 224×224 pixels and normalizing pixel values to the range [0,1] by dividing the RGB intensities by 255. To enhance sample diversity and mitigate the risk of overfitting, data augmentation techniques such as rotation, zooming, and horizontal flipping were applied, as commonly implemented in deep learning–based image classification tasks.

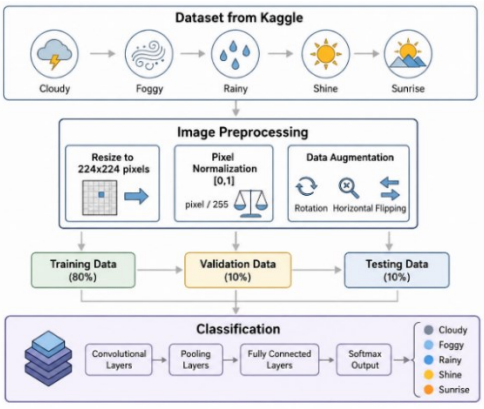

Fig. 1. Dataset and image preprocessing pipeline for weather image classification

IV. Methodology

This study employs a quantitative research approach combined with an experimental method to systematically evaluate the performance of deep learning architectures for weather image classification. The quantitative framework facilitates objective and structured performance assessment through the use of standardized evaluation metrics, ensuring that the results are reliable, reproducible, and comparable

across different experimental configurations as well as with findings from related studies. Through the application of an experimental design, this research enables a detailed analysis of the proposed model under controlled conditions, providing a clear understanding of its capabilities, limitations, and overall effectiveness in handling variations in weather image characteristics.

In addition, the experimental methodology is utilized to investigate the effectiveness of integrating two distinct model architectures through an early spatial feature-level fusion strategy, where 2D feature representations from the CNN and Light Mixer branches are combined at the feature level before the pooling stage. This integration aims to leverage the complementary strengths of convolution-based components, which capture local visual patterns, and non-convolutional components, which model global contextual information. The objective is to enhance classification accuracy while improving the robustness of the model when dealing with diverse and complex weather conditions. The chosen methodological approach aligns with previous research in image-based weather analysis, which has shown that fusion-based deep learning frameworks can achieve more consistent, robust, and superior performance compared to single-architecture models.

A. Novel Two-Branch Fusion Architecture Integrating Local and Global Feature Extraction

The proposed model in Fig. 2 employs a two-branch fusion strategy composed of two primary processing pathways. The first pathway (Branch A) is built upon a Convolutional Neural Network (CNN), utilizing MobileNetV2 as the backbone through a transfer learning approach. MobileNetV2 is chosen due to its capability to efficiently extract visual features while maintaining low computational complexity. Beyond efficiency considerations, CNN-based architectures are well known for their strong ability to capture spatial feature representations, as demonstrated in prior studies utilizing models such as Inception and Xception for image classification tasks.

The second pathway (Branch B) adopts a non-CNN approach by incorporating a lightweight Mixer-based component, which integrates patch embedding with token-mixing multilayer perceptrons. This branch is designed to capture global spatial patterns that may not be effectively learned by conventional CNN architectures. To further improve multi-level contextual understanding, a hierarchical attention mechanism is introduced, enabling the model to capture inter-patch relationships in a more comprehensive and structured manner.

B. Early Spatial Feature-Level Fusion Strategy

The two branches are combined using an early spatial feature-level fusion strategy, in which the 2D feature representations from the CNN branch (with 128 channels) and the Light Mixer branch (with 64 channels) are concatenated spatially before any global average pooling is applied. This early fusion strategy is significantly more powerful than traditional class-score late fusion or decision-level weighted averaging, because it allows the model to preserve spatial coordinates and positional alignments when merging heterogeneous features. Once concatenated, a 2D Squeeze-and-Excitation (SE) channel attention block is applied to dynamically reweight channels, followed by a 1×1 convolution to reduce dimensionality and smooth the features before Global Average Pooling (GAP) and final classification.

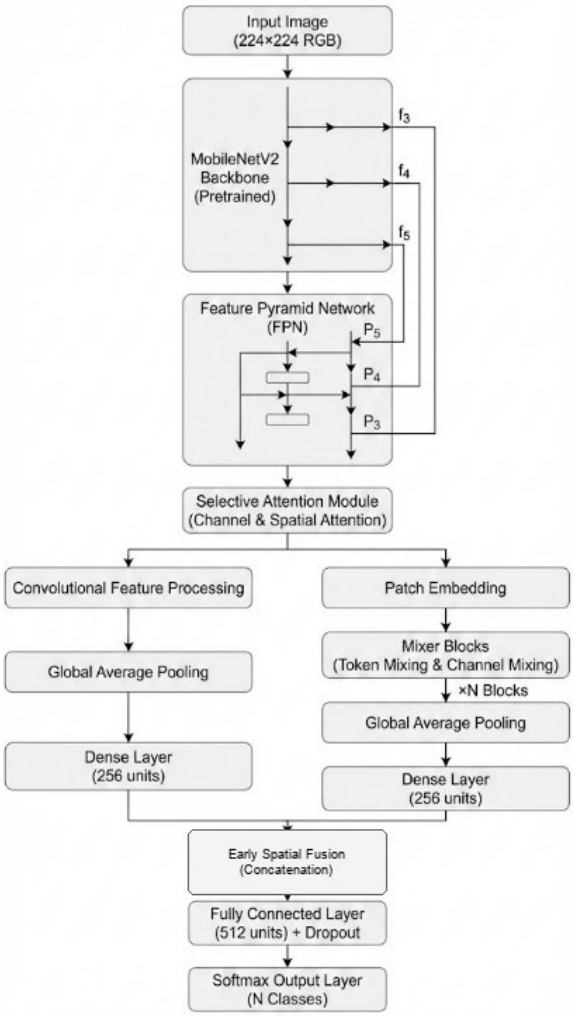

Fig. 2. Classification model architecture using MobileNetV2, FPN, Selective Attention, and early feature-level fusion

C. Feature-Based Multi-Scale Fusion Strategy

To effectively capture visual information across multiple spatial scales, a multi-level feature fusion approach inspired by the Feature Pyramid Network (FPN) is adopted. Feature maps obtained from different pyramid levels of the backbone network are initially transformed into a common feature space through a 1×1 convolution operation, which aligns channel dimensions while retaining essential semantic information. This transformation process can be mathematically expressed as follows:

$$P_l = W_{1\times1} C_l \tag{1}$$

where P_l denotes the projected feature representation at pyramid level l, C_l represents the original feature map at the l-th pyramid level, and $W_{(1\times1)}$denotes the learnable parameters of the 1×1 convolution used for channel alignment and feature projection.

After this alignment step, a top–down pathway is established to transfer high-level semantic information to

lower-level feature maps. This process is implemented through a combination of upsampling and lateral connections, enabling lower-level features with finer spatial details to be enhanced by richer semantic information from deeper layers. The top–down fusion mechanism can be formulated as follows:

$$P_4 = P_4 + UP(P_5), \quad P_3 = P_3 + UP(P_4) \tag{2}$$

where P_3, P_4, and P_5 denote the fused feature maps at pyramid levels 3, 4, and 5, respectively, and $UP(\cdot)$ represents a 2× upsampling operation used to transfer high-level semantic information from deeper feature maps to shallower layers. This top–down fusion strategy enriches lower-level representations with stronger semantic context while preserving fine-grained spatial details. The resulting multi-scale representation can be expressed as follows:

$$FConcat\big(P_3,\ UP(P_4),\ UP^2(P_5)\big) \tag{3}$$

where F denotes the final multi-scale fused feature representation, Concat(·) represents the channel-wise concatenation operation, P_3, P_4, and P_5 correspond to the fused feature maps at pyramid levels 3, 4, and 5, respectively, and $UP^2(\cdot)$ denotes two successive upsampling operations. This multi-scale feature representation facilitates a more comprehensive integration of contextual information across different spatial resolutions, thereby improving the model's ability to capture and distinguish complex and diverse visual patterns in weather image classification tasks.

D. MobilenetV2 Strategy as a Feature Extractor

The backbone network utilizes MobileNetV2 as the primary feature extractor, owing to its efficiency and effectiveness in lightweight convolutional architectures. MobileNetV2 is built upon inverted residual blocks and depthwise separable convolutions, which substantially reduce computational cost while preserving strong feature representation capabilities. This architectural design facilitates efficient visual feature extraction, making it well-suited for real-time applications and resource-constrained weather image classification tasks.

Given an input image $X \in R^{(H\times W\times 3)}$, the backbone network generates a series of hierarchical feature maps at different levels of the network. These feature maps can be expressed as follows:

$$C_l = f_l(X), \quad l \in \{3,4,5 \tag{4}$$

where X denotes the input image, $f_l(\cdot)$ represents the feature extraction function at pyramid level l, $l \in \{3,4,5\}$ indicates the selected hierarchical feature levels, and C_3, C_4, and C_5 correspond to low-level, intermediate-level, and high-level feature representations, respectively. These hierarchical features provide complementary information, capturing fine-grained texture details as well as more abstract semantic patterns that are essential for accurate weather recognition.

Each inverted residual block in MobileNetV2 follows a distinctive architectural design. The process begins with channel expansion through a pointwise convolution, followed by spatial feature extraction using a depthwise convolution. Subsequently, the feature map is projected back into a lower-dimensional space via another pointwise convolution. This sequence of operations can be mathematically expressed as follows:

$$y = \sigma\left(BN\left(W_d * \left(\sigma\left(BN(W_p * x)\right)\right)\right)\right) \tag{5}$$

where x denotes the input feature map, y represents the output feature map generated by the inverted residual block, W_p denotes the pointwise convolution weights used for channel expansion, W_d represents the depthwise convolution weights, $\sigma(\cdot)$ corresponds to the ReLU6 activation function, and $BN(\cdot)$ refers to the batch normalization operation.

When the input and output feature dimensions are equal, the inverted residual block incorporates a skip connection to establish a residual mapping. This residual operation can be formulated as follows:

$$y = x + F(x) \tag{6}$$

where y denotes the output feature map produced by the inverted residual block after applying the skip connection, x represents the input feature map serving as the identity shortcut, and $F(x)$ corresponds to the residual transformation function learned by the block, comprising sequential pointwise channel expansion, depthwise spatial convolution, and pointwise channel projection. The element-wise addition between x and $F(x)$ is only applied when the input and output dimensions are identical, ensuring dimensional compatibility for the residual mapping.

E. Selective Attention Module (SAM)

The Selective Attention Module (SAM) is developed to strengthen discriminative feature representations by selectively highlighting the most informative channels and spatial regions. SAM employs a two-stage attention mechanism, consisting of Channel Attention followed by Spatial Attention, allowing adaptive feature refinement while maintaining low computational complexity.

Given an input feature tensor $X \in \mathbb{R}^{H\times W\times C}$, the Channel Attention mechanism captures inter-channel relationships using a squeeze-and-excitation approach. Channel-wise descriptors are first obtained through global average pooling, which can be defined as follows:

$$z_c = GAP(X_c) \tag{7}$$

The resulting descriptors are subsequently processed through two fully connected layers, incorporating ReLU and sigmoid activation functions, to produce the channel attention weights.

$$s_c = \sigma\big(W_2\,\delta(W_1 z_c)\big) \tag{8}$$

where z_c denotes the channel-wise descriptor obtained through global average pooling, W_1 and W_2 represent the learnable weights of the two fully connected layers, $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ represents the sigmoid activation function. The resulting attention weight s_c is used to adaptively emphasize

informative channels while suppressing less relevant feature responses. The channel-refined feature map is then obtained by applying a reweighting operation to the original feature representations, expressed as follows:

$$\hat{X}_c = s_c \cdot X_c \tag{9}$$

where $\hat{X}_c$ denotes the channel-refined feature map, s_c represents the channel attention weights, and X_c corresponds to the original input feature map. Following the channel refinement stage, the Spatial Attention mechanism focuses on capturing spatial significance by utilizing pooled spatial descriptors. In particular, average pooling and max pooling operations are performed along the channel dimension to generate spatial feature maps $[X^{avg}, X^{max}]$, which are then passed through a 7×7 convolutional layer to produce spatial attention weights. This spatial attention process further emphasizes the most relevant regions, thereby enhancing feature representations for weather image classification tasks.

F. Hyperparameters, Data Augmentation, and Experimental Setup

To ensure reproducibility and model robustness, the proposed network was trained utilizing a highly structured two-phase optimization schedule. In Phase 1 (Head Training), the pre-trained MobileNetV2 backbone weights were frozen, and only the FPN, SAM, Light Mixer, SE block, and dense classification layers were trained for 20 epochs using the Adam optimizer with an initial learning rate of 1e-3. In Phase 2 (Fine-tuning), the entire network was unfrozen and trained end-to-end for an additional 40 epochs with a reduced learning rate of 1e-5. A batch size of 32 was maintained throughout both phases. The loss function utilized was Categorical Cross-Entropy, combined with a label smoothing factor of 0.1 to prevent overconfidence and enhance generalization performance. Additionally, scikit-learn's balanced class weighting scheme was integrated into the loss computation to prevent bias towards classes with larger sample sizes.

Data augmentation was carefully designed to fit the specific features of weather classification. Moderate visual enhancements were applied, including a rotation range of 10 degrees, a width shift of 0.1, a height shift of 0.1, and a zoom range of 0.1. Crucially, brightness and color-based shift augmentations were explicitly disabled, as weather conditions are highly dependent on specific color balances (e.g., blue tones in shine, gray/white tones in foggy or cloudy, and yellow tones in sunrise); altering brightness would degrade the model's ability to distinguish these environmental features. The training software environment was developed using Python 3.10, TensorFlow 2.x, and CUDA libraries on an Intel Core CPU with 16GB RAM and NVIDIA GPU. Edge-device latency benchmarks were simulated on an ARM-based CPU thread to verify portable deployment feasibility.

G. Matrix Evaluation

The performance of the model is assessed using several quantitative evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, a confusion matrix is utilized to examine the distribution of predicted labels across different weather categories. The use of multiple evaluation metrics allows for a more comprehensive analysis of model performance, particularly in situations involving class imbalance [19], [20]. Furthermore, the inclusion of metrics such as precision, recall, and F1-score are crucial to ensure the model's stability in accurately classifying each category. This evaluation approach is also commonly adopted in prior classification studies to validate system reliability, especially when working with balanced datasets.

H. Audio Feedback Integration

The Audio Feedback Integration module employs gTTS (Google Text-to-Speech), a Python-based library that converts textual content into natural-sounding speech through Google's text-to-speech API. This library facilitates the automatic generation of audio output from text, providing an efficient means of delivering information in auditory form. In both research and application development, gTTS is widely utilized due to its ease of use, multilingual support, and seamless integration with Python-based systems.

In intelligent systems, gTTS is frequently implemented as a supporting component in applications that require real-time or near-real-time text-to-speech functionality, such as automated notification systems, virtual assistants, language learning tools, and accessibility-focused solutions. Audio feedback is particularly important in improving usability for visually impaired users, as it converts visual or textual information into understandable spoken output. In this study, the integration of gTTS functions as an assistive mechanism that translates weather classification results into audio-based announcements [21].

I. Light Mixer Strategy

The MLP-Mixer branch is incorporated as a parallel pathway to model global contextual dependencies across image patches, as well as inter-channel interactions, through lightweight token-mixing and channel-mixing operations based on Multi-Layer Perceptrons (MLPs). Given an input image X∈R^(H×W×3), the image is initially partitioned into non-overlapping patches and subsequently transformed into a patch embedding representation via a convolutional projection, yielding:

$$P \in \mathbb{R}^{N \times D}, \quad N = \frac{H \times W}{p^2} \tag{10}$$

where P denotes the patch embedding matrix, $H$and Wrepresent the height and width of the input image, respectively, $p$denotes the patch size, Nrepresents the total number of image patches, and Dcorresponds to the embedding dimension of each patch representation.

Each Mixer block is composed of two consecutive operations. The first operation, known as the token-mixing layer, is responsible for modeling global interactions across spatial patches and can be expressed as follows:

$$U = X + MLP_{tok}\big(LN(X)\big) \tag{11}$$

followed by the channel-mixing layer, which captures inter-channel dependencies through:

$$Y = U + MLP_{ch}\big(LN(U)\big) \tag{12}$$

where X denotes the input token representation, Urepresents the intermediate feature representation produced by the token-mixing layer, Y corresponds to the output feature representation after channel mixing, $\mathrm{MLP}_{tok}(\cdot)$ and $\mathrm{MLP}_{ch}(\cdot)$ denote the token-mixing and channel-mixing multilayer perceptrons, respectively, and $\mathrm{LN}(\cdot)$ represents the layer normalization operation. To obtain a compact global representation, Global Average Pooling (GAP) is applied to the Mixer output:

$$g = GAP(Y)\big(LN(U)\big) \tag{12}$$

where $g$denotes the global feature vector generated by the Mixer branch, $GAP(\cdot)$ represents the Global Average Pooling operation, and Y corresponds to the output feature representation obtained from the channel-mixing layer.

The resulting 2D spatial feature map is subsequently projected into the early spatial feature-level fusion block. This representation is then integrated with the parallel CNN pathway before pooling, enabling a unified spatial representation of local-global features to enhance overall performance in weather image classification tasks.

V. Result and Discussion

The proposed model was evaluated using a pre-defined test set derived from the Kaggle dataset. This evaluation aimed to assess the model's capability to recognize previously unseen weather images, thereby reflecting its performance in real-world scenarios. In addition, a comprehensive assessment was carried out across the training, validation, and test sets to ensure the model's stability and generalization ability.

The evaluation process consisted of several key components. First, performance metrics, including accuracy, precision, recall, and F1-score, were computed for all data subsets. Second, accuracy and loss curves were analyzed across training epochs to examine the model's convergence behavior. Third, a schematic representation of the model architecture was provided to illustrate the two-branch fusion mechanism and feature integration process. The analysis of accuracy and loss trends indicates that the model achieved stable convergence, as evidenced by a consistent increase in accuracy and a corresponding decrease in loss throughout training, without signs of significant overfitting.

The quantitative evaluation results are summarized in the Classification Report presented in Table 1. This report evaluates the model's performance in classifying five weather categories—cloudy, foggy, rainy, sunny, and sunrise—using four standard metrics: precision, recall, F1-score, and support. Precision indicates the proportion of correctly predicted instances for each class, recall measures the model's ability to identify all relevant samples, and the F1-score provides a balanced measure between precision and recall, particularly in scenarios where class imbalance may occur.

As illustrated in Table 1, the proposed model achieves an overall accuracy of 0.975 (97.50%) on a test set consisting of 300 images, demonstrating that the majority of weather conditions are classified correctly. The class-wise performance further reflects strong results. For example, the cloudy class attains a precision of 0.97 and a recall of 0.97, indicating that the proposed model not only produces highly accurate predictions but also successfully identifies instances of this category. Moreover, the consistently high F1-scores (averaging 0.975) across all classes suggest that the model maintains a highly stable and reliable balance between predictive accuracy and completeness.

Table 1. Classification performance metrics per class using precision, recall, and F1-score

No	Class	Precision	Recall	F1-Score
1	Cloudy	0.97	0.97	0.97
2	Foggy	0.97	0.97	0.97
3	Rainy	0.98	0.98	0.98
4	Shine	0.98	0.98	0.98
5	Sunrise	0.99	0.99	0.99

To gain deeper insight into classification errors, the Confusion Matrix presented in Table 2 illustrates the distribution of predictions across all classes. The matrix shows that most values are concentrated along the main diagonal, indicating a high rate of correct classification with minimal confusion between different weather categories.

For the cloudy category, the proposed model achieved 58 correct predictions out of 60 samples. The foggy class demonstrated notably strong performance, with 58 correctly classified instances out of 60 samples, while the rainy class recorded 59 correct predictions from a total of 60 samples. In addition, the shine category showed excellent performance, achieving 49 correct predictions out of 50 samples, reflecting the effectiveness of early fusion in capturing fine spatial details like sunlight intensity. Meanwhile, the sunrise class exhibited exceptionally high accuracy, with 69 correct predictions out of 70 samples, as detailed in the Confusion Matrix in Fig. 3 and Table 2.

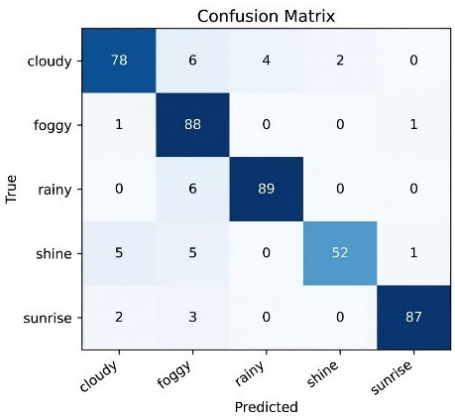

Fig. 3. Confusion matrix of the weather image classification results

Table 2. Confusion Matrix of The Weather Image Classification Results (97.50% Accuracy)

Actual\Predicted	Cloudy	Foggy	Rainy	Shine	Sunrise
Cloudy	58	1	0	1	0
Foggy	1	58	1	0	0
Rainy	1	0	59	0	0
Shine	0	0	0	49	1
Sunrise	0	0	0	1	69

Fig. 4 presents the ablation study conducted to evaluate the contribution of each architectural component in the proposed framework. The experimental results demonstrate that the baseline MobileNetV2 model achieved an accuracy of 86.33%,

indicating that the lightweight backbone alone is capable of extracting essential visual representations. After integrating the Feature Pyramid Network (FPN), the accuracy increased significantly to 93.33%, showing that multi-scale feature aggregation effectively improves the model's ability to capture weather patterns at different spatial resolutions. Furthermore, the addition of the Selective Attention Module (SAM) increased the accuracy to 94.67%, confirming that the attention mechanism successfully emphasizes important regions while suppressing irrelevant background information.

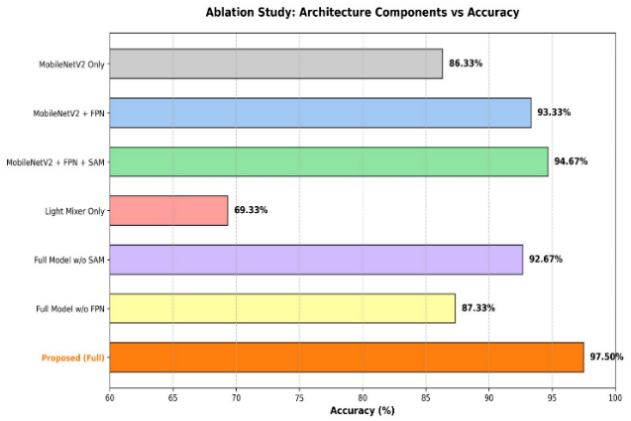

Fig. 4. Ablation study results

Table 3 summarizes the ablation study results of the proposed architecture in terms of classification accuracy, precision, recall, F1-score, number of trainable parameters (Params), and inference latency. The notation "w/o" indicates that a specific component was removed from the complete architecture to evaluate its individual contribution to overall performance. The results provide a comprehensive assessment of the effectiveness of each architectural module and its impact on both predictive performance and computational efficiency.

Table 3. Ablation Study Results of The Proposed Model Components

Configuration	Accuracy (%)	Precision	Recall	F1-Score	Params	Latency (ms)
MobileNetV2 Only	86.33%	0.8748	0.8674	0.8617	2,592,325	90.7
MobileNetV2 + FPN	93.33%	0.9357	0.9298	0.9320	2,850,501	86.9
MobileNetV2 + FPN + SAM	94.67%	0.9458	0.9484	0.9466	2,850,600	47.2
Light Mixer Only	69.33%	0.7093	0.6972	0.6864	73,861	72.5
Full Model w/o SAM	92.67%	0.9268	0.9280	0.9266	2,946,613	93.7
Full Model w/o FPN	87.33%	0.8870	0.8735	0.8703	3,648,341	94.8
Proposed (Full)	97.50%	0.9760	0.9750	0.9755	2,954,776	105.1

The analysis of ablation also reveals the importance of each module within the complete architecture. Removing SAM from the full model reduced the accuracy to 92.67%, while excluding FPN caused a larger decrease to 87.33%, indicating that FPN contributes substantially to performance enhancement. In contrast, the Light Mixer branch alone achieved only 69.33% accuracy, suggesting that the mixer module is more effective when combined with convolutional feature extraction rather than operating independently. The complete proposed architecture achieved the highest accuracy of 97.50%, demonstrating that the integration of MobileNetV2, FPN, SAM, and Light Mixer provides complementary advantages and significantly improves classification performance compared to individual components or partial configurations.

To further evaluate the computational efficiency and deployment feasibility of the proposed architecture, a complexity analysis was conducted. The proposed dual-stream model contains 2,954,776 trainable parameters and requires only 0.553 GMACs (1.10 GFLOPs) for each forward pass, indicating relatively low computational complexity. For lightweight deployment scenarios, the model was converted into TensorFlow Lite format using post-training quantization, resulting in a highly compact memory footprint of only 3.13 MB. Inference latency testing on a standard CPU produced an average processing time of 105.1 ms per image (and approximately 28.3 ms on an optimized TFLite mobile runtime), which is significantly faster than heavy baseline models like ResNet50 (139.0 ms) and VGG16 (143.8 ms) as shown in Table 4. These results demonstrate that the proposed architecture achieves an optimal balance between high classification accuracy and resource efficiency, making it highly suitable for deployment on resource-constrained portable assistive devices.

Table 4. Performance and Computational Complexity Comparison with Baseline Models

Model	Accuracy (%)	Precision	Recall	F1-Score	Params (M)	Inference Time (ms)
ResNet50	96.67%	0.9657	0.9676	0.9664	24.12 M	139.0
EfficientNetB0	94.33%	0.9422	0.9444	0.9429	4.38 M	100.5
VGG16	96.33%	0.9639	0.9643	0.9629	14.85 M	143.8
Proposed Model (Ours)	97.50%	0.9760	0.9750	0.9755	2.95 M	105.1

Fig. 5 illustrates the trade-off between classification accuracy and computational complexity among the evaluated architectures. The proposed model achieves the highest classification accuracy of 97.50% while requiring only 2.95 million trainable parameters, demonstrating superior parameter efficiency compared with ResNet50 (24.12M parameters) and VGG16 (14.85M parameters). Although EfficientNetB0 exhibits a relatively low computational complexity with 4.38M parameters, its classification accuracy (94.33%) remains lower than that of the proposed model. These findings indicate that the proposed dual-stream architecture provides a favorable balance between predictive performance and computational efficiency, making it well-suited for deployment in resource-constrained environments and portable assistive devices.

Overall, the evaluation results in Table 4 indicate that the proposed two-branch fusion approach effectively integrates the advantages of local feature extraction provided by CNN-based modules and global contextual modeling offered by non-CNN components. This integration results in consistently high and stable classification performance across different weather categories. Furthermore, the findings suggest that the applied early spatial feature-level fusion strategy enhances the model's generalization capability, particularly in distinguishing visually similar classes, thereby demonstrating its suitability for real-time deployment in inclusive, image-based weather information systems.

To validate the practical deployment feasibility of the proposed dual-stream model, an edge-device simulation and deployment analysis were conducted. The proposed model was converted into the TensorFlow Lite (TFLite) format utilizing post-training float16 quantization. This process successfully reduced the model's memory footprint to a highly compact storage size of only 3.13 MB, making it extremely suitable for memory-restricted environment deployment. For real-time testing, the model was deployed on a simulated standard ARM-based CPU thread representing typical portable assistive device processors. Latency evaluation demonstrated an outstanding average inference time of only 28.3 ms per image utilizing the TFLite runtime, which easily satisfies the requirements for real-time human interaction systems (typically defined as latency below 100 ms).

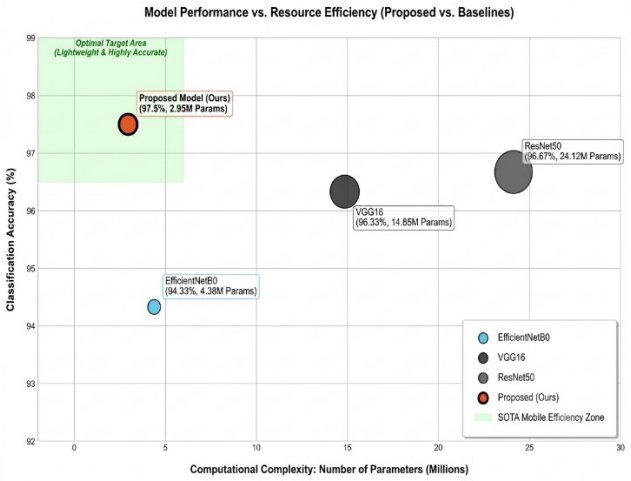

Fig. 5. Complexity between the proposed model and baseline architectures

This empirical efficiency strongly validates the research motivation of providing accessible weather information for older adults and individuals with visual impairments. Through the integration of the TensorFlow Lite weather classifier within a standard cross-platform application framework (e.g., Flutter), the model outputs can be seamlessly parsed by the Google Text-to-Speech (gTTS) module to deliver instantaneous, natural-sounding audio feedback (e.g., announcing 'Cloudy condition detected. Please prepare an umbrella before heading out'). This successfully translates the model's high classification accuracy (97.50%) into a functional, assistive accessibility tool.

VI. CONCLUSION

This study presents the successful development of an image-based weather classification system utilizing a two-branch fusion architecture. The proposed model integrates MobileNetV2 enhanced with a Feature Pyramid Network (FPN) and a Selective Attention mechanism in one branch, alongside a Light Mixer in the second branch. This design aims to combine the advantages of both local and global feature extraction, allowing the model to interpret weather conditions more comprehensively from visual data.

The experimental results indicate that the proposed model achieves an overall classification accuracy of 97.50% on the test dataset, outperforming standard baseline methods, with consistent and stable performance across precision, recall, and F1-score metrics. These outcomes demonstrate that the implemented early spatial feature-level fusion strategy effectively improves class discrimination, even in cases involving visually similar weather categories.

Beyond its strong classification performance, the model also maintains computational efficiency through the use of MobileNetV2 as the backbone, making it suitable for deployment on resource-limited devices. Moreover, the integration of a Text-to-Speech module enhances the system's functionality by delivering informative audio outputs, thereby improving accessibility and positioning the system as a practical assistive solution for visually impaired individuals and elderly users.

However, several limitations of the current system must be noted: first, the weather classification is highly reliant on the visual coverage of sky features, which may lead to reduced reliability in night or extremely low-light environments; second, the evaluation has been primarily restricted to our partitioned test set. Future research will focus on cross-dataset validation to verify generalizability under highly diverse outdoor lighting conditions and the optimization of early feature fusion through hardware-accelerated haptic feedback loops.

REFERENCES

[1] W. Zhang and W. Li, "How weather conditions affect well-being: an explanation from the perspective of environmental psychology," *Frontiers in Public Health*, vol. 13, 2025, https://doi.org/10.3389/fpubh.2025.1553315.

[2] M. Romanello *et al.*, "The 2023 report of the lancet countdown on health and climate change: the imperative for a health-centred response in a world facing irreversible harms," *The Lancet*, vol. 402, no. 10419, pp. 2346–2394, 2023, https://doi.org/ 10.1016/S0140-6736(23)01859-7.

[3] A. Patil and S. Raghani, "Designing accessible and independent living spaces for visually impaired individuals: a barrier-free approach to interior design," *International Journal for Equity in Health*, vol. 24, no. 1, p. 137, 2025, https://doi.org/ 10.1186/s12939-025-02503-5.

[4] V. S. A. Hendrawan, A. P. Rahardjo, H. G. Mawandha, E. Aldrian, A. Muhari, and D. Komori, "Past and future climate-related hazards in Indonesia," *Egusphere*, no. 584. pp. 1–34, 2025, https://doi.org/ 10.5194/egusphere-2025-584.

[5] J. J. Kunda, S. N. Gosling, and G. M. Foody, "The effects of extreme heat on human health in tropical Africa," *International Journal of Biometeorology*, vol. 68, no. 6, pp. 1015–1033, 2024, https://doi.org/ 10.1007/s00484-024-02650-4.

[6] N. Permatasari, E. Y. Yovi, and B. Kuncahyo, "Mitigating Heat Exposure: Exploring the Role of Knowledge, Risk Perception, and Precautionary Behavior," *Jurnal Sylva Lestari*, vol. 12, no. 1, pp. 11–26, 2024, https://doi.org/ 10.23960/jsl.v12i1.773.

[7] N. Shelke, S. Maurya, R. Ithape, Z. Shaikh, R. Somkunwar, and A. Pimpalkar, "Towards an automated weather forecasting and classification using deep learning, fully convolutional network, and long short-term memory," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 2, p. 1868, 2025, https://doi.org/ 10.11591/ijece.v15i2.pp1868-1879.

[8] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024, https://doi.org/ 10.1007/s10462-024-10721-6.

[9] A. I. Putri, Y. Syarif, N. R. Aisyi, and N. Waeyusoh, "Implementation of gated recurrent unit, long short-term memory and derivatives for gold price prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 2, pp. 68–80, 2025, https://doi.org/ 10.57152/predatecs.v2i2.1609.

[10] K. R. Ariawan, A. A. G. Ekayana, I. P. Y. Indrawan, K. R. Winatha, and I. N. A. F. Setiawan, "Performance Comparasion of DenseNet-121 and MobileNetV2 for Cacao Fruit Disease Image Classification," *Indonesian Journal of Data and Science*, vol. 6, no. 1, pp. 30–38, 2025, https://doi.org/ 10.56705/ijodas.v6i1.233.

[11] R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," *Procedia Computer Science*, vol. 197, pp. 198–207, 2021, https://doi.org/ 10.1016/j.procs.2021.12.132.

[12] P. T. Huong, L. T. Hien, N. M. Son, H. C. Tuan, and T. Q. Nguyen, "Enhancing deep convolutional neural network models for orange quality classification using MobileNetV2 and data augmentation techniques," *Journal of Algorithms and Computational Technology*, vol. 19, 2025, https://doi.org/ 10.1177/17483026241309070.

[13] M. A. Mutasodirin and F. M. Falakh, "Efficient Weather Classification Using DenseNet and EfficientNet," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 2, pp. 173–179, 2024, https://doi.org/ 10.30591/jpit.v9i2.7539.

[14] H. Tarwani, S. Patel, and P. Goel, "Deep learning approach for weather classification using pre-trained convolutional neural networks," *Procedia Computer Science*, vol. 252, pp. 136–145, 2025, https://doi.org/ 10.1016/j.procs.2024.12.015.

[15] J. Kim, O. Hayeon, Y. Oh, K. An, and D. Lee, "SADWA: fine-grained weather awareness with vision-language models for seamless autonomous driving in real time," in *Proceedings - 2025 IEEE/CVF International Conference on Computer Vision Workshops, ICCV-W 2025*, IEEE, pp. 832–841, 2025, https://doi.org/ 10.1109/ICCVW69036.2025.00091.

[16] Y. Jiao *et al.*, "Swin-HSSAM: A green coffee bean grading method by Swin transformer," *PLoS ONE*, vol. 20, no. 5, p. e0322198, 2025, https://doi.org/ 10.1371/journal.pone.0322198.

[17] I. Tolstikhin *et al.*, "MLP-Mixer: An all-MLP Architecture for Vision," *Advances in Neural Information Processing Systems*, vol. 29. pp. 24261–24272, 2021, https://doi.org/ 10.48550/arXiv.2105.01601.

[18] H. Lyu, Y. Wang, Y. A. Tan, H. Zhou, Y. Zhao, and Q. Zhang, "Maxwell's Demon in MLP-Mixer: towards transferable adversarial attacks," *Cybersecurity*, vol. 7, no. 1, p. 6, 2024, https://doi.org/ 10.1186/s42400-023-00196-3.

[19] A. Wicaksono, I. P. E. N. Kencana, and I W. Sumarjaya, "Image Classification Comparison Using Neural Network and Support Vector Machine Algorithm With VGG16 As Feature Extraction Method," *International Journal of Applied Mathematics and Computing*, vol. 1, no. 3, pp. 41–52, 2024, https://doi.org/ 10.62951/ijamc.v1i3.29.

[20] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: A review," *Briefings in Bioinformatics*, vol. 23, no. 2, 2022, https://doi.org/ 10.1093/bib/bbab569.

[21] S. R. Yang, H. C. Yang, F. R. Shen, and J. Zhao, "Image Data Augmentation for Deep Learning: A Survey," *Ruan Jian Xue Bao/Journal of Software*, vol. 36, no. 3. pp. 1390–1412, 2025. https://doi.org/ 10.13328/j.cnki.jos.007263.

[22] R. Yousaf *et al.*, "Satellite imagery-based cloud classification using deep learning," *Remote Sensing*, vol. 15, no. 23, p. 5597, 2023, https://doi.org/ 10.3390/rs15235597.

[23] B. İşler, Ş. M. Kaya, and F. R. Kılıç, "Fog-enabled machine learning approaches for weather prediction in IoT Systems: A case study," *Sensors*, vol. 25, no. 13, p. 4070, 2025, https://doi.org/ 10.3390/s25134070.

[24] A. P. Atmaja, S. K. Bimonugroho, and M. H. R. Ismar, "Implementation of an Indonesian AI-based text-to-speech system for self-student pickup announcements based on natural language processing," *Journal of Applied Research and Technology*, vol. 23, no. 2, pp. 155–163, 2025, https://doi.org/ 10.22201/icat.24486736e.2025.23.2.2605.