ISSN: 2458-8989

Natural and Engineering Sciences

NESciences, 2026, 11 (3): 877-890 doi: 10.28978/nesciences.263068

Multimodal Deep Learning Framework for Agricultural Soil Classification and Environmental Risk Assessment Using RGB and Thermal Images

Ben Sujin B[1*], Arjun Thirunavukkarasu[2], M. Sangeetha[3]

[1]Deanship of Postgraduate Studies, Research and Innovation, University of Buraimi, Sultanate of Oman.

[2]Westmount Charter High School, Calgary, AB Canada.

[3]University of Technology and Applied Sciences, College of Engineering, Nizwa, Sultanate of Oman.

Abstract

Agricultural soil monitoring plays an important role in crop production, sustainable land management, and environmental protection. Integrating RGB and thermal imaging offers opportunities to analyse complementary visual and temperature-related characteristics through multimodal deep learning. This study presents a framework based on two parallel ResNet18 convolutional neural network branches, designed to extract features independently from RGB and thermal images. The extracted representations are combined through feature-level concatenation and processed by fully connected layers for classification. RGB images capture visible surface characteristics, including colour, texture, and patterns, whereas thermal images provide information about surface temperature variations. The experimental investigation uses paired road and sidewalk images to examine the effectiveness of multimodal feature fusion in surface-condition classification. The classification task considers dry, wet, snow, and ice surface conditions. Model performance is examined through classification metrics and comparative evaluation of different training configurations. The findings demonstrate the potential of combining visual and thermal information to improve automated surface-condition recognition. Although the framework offers a methodological foundation that may inform agricultural monitoring applications, the present experiment does not directly evaluate agricultural soil samples or establish soil-related environmental risks. Its application to agricultural soil classification would require representative paired RGB and thermal images, verified soil categories, and independent field or laboratory measurements. The framework therefore provides a basis for further investigation into multimodal approaches for agricultural soil monitoring and environmental assessment.

Keywords: Agricultural soil classification, Environmental risk assessment, Feature-level fusion, Multimodal deep learning, red, green, and blue imaging, Thermal imaging.

Available Online: 10/09/2026

1. Introduction

Side walk and pavement conditions plays a vital role for pedestrians' movements during the winter season in countries like Canada. Anyone can harm themselves if they do not care about the side walk surface condition. To prevent slippery and preventing from getting hurt, the pedestrians take much care before using the side walk. This creates a huge hurdle on office goers and other common people. Recent advancement in artificial intelligence, machine learning and deep learning techniques provides an automated alert to the

*Corresponding Author: Ben Sujin B, E-mail: bensujin.b@uob.edu.om

pedestrians about the side walk surface condition. The automated systems integrated with the AI, ML and deep learning techniques can provide valuable information of the side walk surface condition and prevents the pedestrians from slippery. However, the sidewalk surface condition can change frequently depending upon the environmental conditions, the pavement conditions and illuminations. Hence a normal deep learning technique may not be sufficient to predict the class of side walk surface conditions whether it is snow, icy, wet or dry. Most of the available systems or models uses the RGB images for sidewalk surface classification. The RGB images cannot extract all features of the sidewalk as the feature set may vary depends on the illumination. Hence relying only on the RGB image data set may prolong to wrong classification.

Thermal imaging on the other hand provides valuable information about the sidewalk surface condition. However, treating the thermal images alone will not be advisable in predicting the category of side walk surface condition. Looking after all developments and research gaps, the proposed research work uses the RGB and Thermal image pair as the input dataset to train the models and to classify the sidewalks based on its surface condition. The proposed model is a multimodal deep learning architecture which will combine the RGB image and Thermal image using feature level fusion. Separate RESNET models are used to train and extract the features from 512-dimensional RGB feature vector and 512-dimensional thermal image feature vector and later they are concatenated to a unified 1024-dimensional feature vector.

Multimodal image analysis has become more and more relevant to agricultural monitoring, especially when different sensing technologies supply complementary information on the environmental conditions (El Sakka et al., 2025; Wang et al., 2022). RGB images record surface characteristics that are visible to the human eye, while thermal images record variations in temperature (Vahidi et al., 2025; Zhou et al., 2021). The integration of these technologies using deep learning could pave the way for future advancements in agriculture, such as soil condition monitoring, moisture assessment, and precision land management (Miao et al., 2024; Vahidi et al., 2025). The present investigation is particularly concerned with the classification of road surfaces; however, the multimodal feature-fusion approach used is a methodological basis for further investigations in agricultural and environmental monitoring.

2. Literature Review

A systematic literature review was carried out from the recent publications under different themes coordinating the research objective. The review was done under the categories; Side Walk surface condition classification and risk assessment, deep learning applications on thermal and RGB images, multimodal deep learning and feature level fusion techniques. The following sections describes the review of literatures in a detailed manner.

2.1Review on Side Walk Surface Classification and Assessing the Risk

Accurate assessment of risk about the side walk surface condition is important to provide a safe movement for pedestrians especially during the winter season. The side walk surface characteristics changes frequently due to many environmental conditions. Environmental conditions such as snow, icy street, wetness can influence slippery walk surface for the people who use the pavements. Hence, the automatic detection of side walk surface condition has attracted the attention of people and researchers, particularly through latest Artificial Intelligence techniques and computer vision techniques.

Early such automated systems relied on conventional machine learning algorithms and hardware-based environment sensors. At present, it is observed that a high-quality imagery technique is available, a considerable improvement in Deep learning techniques and the efficiency of the such techniques are considerably improved, which leads to the development of sophisticated feature extraction and risk classification systems. A study which focus on real time road surface state condition based on deep convolution network. The RefNet incorporates a new residual block to collect the discriminative road surface features by significantly reducing the computing cost. The researchers proposed the five classification categories as snow, ice, wet and water and the same kind of classification is adopted in this research proposal. However, the application proposed in this research is pedestrian walk area or the side walk. The Nvidia Jetson AGX Xavier platform to perform the classification and turn out with an average precision of 88.93% and an inference throughput of 669 FPS. Whereas proposed an automated recognition of snow and ice road surfaces in the mountainous environment. They proposed a T-Net architecture that

uses a split transform merge architecture, asymmetric convolution and attention mechanism to enhance the feature extraction. However, the proposed model focused on side walk surface condition which can resemble the same conditions. The said model uses the RGB images, however the proposed research work uses the thermal images to improve the accuracy of risk assessment.

RSPNet, a light weight road surface perception network proposed by which overcomes the limitations of many previously proposed models in road surface classification. RSPNet uses a wavelet transform convolution to address the limitation and it incorporates the multi-scale depth wise convolution. An attractive catch from this research for the proposed model is the multimodal sensing which will improve the robustness when the acquired visual images get degraded. CNN based classification model using the Road Weather Information System (RWIS) imaging. In this research paper, the researcher proposed with automated examination of directly camera capture images unlike the analysis of cropped images. The paper presents that by increasing the image resolution to 224 X 224 pixels improved the performance significantly. Yumang et al. (2026) conveyed a surface detection mechanism using modified YOLOv8 architecture which utilizes the Einstein based activation function and alpha blending. The proposed configuration achieved a mean average precision of 65.51% and F1-score of 67.19%. The mechanism should not be directly equated with its accuracy but the modification applied on the YOLOv8 architecture. These review shows a big growth of using deep learning techniques in road and sidewalk surface detection and classification. Almost all the architectures which are reviewed here used the RGB images for classification and risk assessment. However, under different environmental conditions, the RGB image alone may not be sufficient to deduce the classification and may provide a reliable assessment. This motivates to use the thermal imaging along with the RGB imaging capability to finely predict and classify the side walk surface condition.

2.2A review on environmental variability and sidewalk surface datasets

It is observed from all literatures that the model's performance is heavily depending on the quality, quantity, variety and representativeness of the datasets. The dataset becomes imbalanced, when some dataset contains a damaged pavement surface and this is a great challenge in classifying the risk accurately.

ViSnow, a snow-covered road surface dataset to support the computer vision applications and algorithms. This dataset laid a basement for evaluating many CNN techniques and deep learning models. The also presented an automated dataset annotation while combining the CCTV imagery. The investigated the effect of illumination changes on the snow surface using the images acquired through the vehicle mounted cameras. The researchers concluded that sudden changes in the surface during dusk may reduce the accuracy and reliability of the conventional classification system. They proposed CycleGAN model to address the issue and reached to 78% of classification accuracy.

A detailed review of winter road surface condition modelling was presented which integrates the stationary sensing, remote sensing vehicle observations and machine learning techniques. The researchers showcased the issues pertains due to the heterogeneous data inculcation and uncertainty of environmental conditions. Considering all reviewed studies, it confirms that the motivation of using RGB combine Thermal image dataset will be value added in the accurate classification of side walk surface.

2.3Review on Deep Learning surface analysis

Based on the literature reviews, thermal imaging adds values to the image-based representation. Thermal images can be used to reveal patterns which are hard to be revealed by the RGB images. The thermal image classification model using a lightweight CNN algorithm and a low-dimensional image transformation technique. The different characteristics, Local Binary Patterns, Histogram of Oriented Gradients and Gabor filtering, of thermal images are investigated. The model claims an accuracy of 98.58%, 95.56% and 100% on three different thermal image datasets. One of the literatures was found interesting on thermal images even though the application was on different domain, detecting defects in cementitious composites. The thermal imaging can assess the structure abnormalities which can be extended to understand the pavement damages.

The pretrained CNN architectures for classifying the thermal images of the Volcanoes. They used the image data set acquired from the thermal cameras mounted at Mount Etna. VGG16 and AlexNet models were using to classify the thermal images and returned 90% accuracy. Another literature on the thermal images captured from electrical transfer rooms. They also presented the use of AlexNet, SqueezeNet, VGG19 and GoogleNet to classify the thermal images. The models provided the accuracies ranges from 86.98% to 100%.

The road slipperiness detection in arctic environments by sensor fusion techniques. and fuzzy logic systems. The system presented an accuracy of 87%, however the system did not use the pair of RGB and Thermal images. Another related system was detecting the pavement abnormalities through a light weight vehicle sensor. The study provided a technique of reduction in training time while maintaining the accuracy. A multimodal road surface recognition system was proposed for electric vehicles. The classification accuracy reported by the model was 98.49%. However, the study does not guarantee the classification on real world vehicle performance classification.

2.4A Review on Deep Learning for Feature Extraction and Feature Fusion

Recent days, the feature extractions done one fused feature are growing and adapted in various algorithms to gain a higher accuracy by retaining the information contained in each feature. A model of deep feature fusion on X-ray security image classification. They proposed a model using Random Weight Network and validated the stand-alone deep learning system with 83.55% accuracy. While applying the feature fusion technique, the test increased the accuracy to 97.44%. These results promises that the feature fusion can improve classification on images. A multimodal deep learning framework called as FusionNet for misinformation detection on social media. The researchers used the BERT model for the textual classifications and ResNet 18 for image classifications. The model proves and accuracy of 98.88% on social media content classification. Although the application domain differs from side walk surface classification, the study defines how encoders can be used in classification. A multi-feature fusion model for understanding irrelevant faces in short form videos. It used a multi head fusion network to integrate multiple feature representations. As the approach is directly related to facial recognition, it is treated in this research as a secondary study in terms of side walk surface detection using deep learning architectures.

2.5 Multimodal Deep Learning Applications in Agricultural and Environmental Monitoring

In recent years, the application of artificial intelligence in agricultural monitoring has been extended by the development of multimodal deep learning. Abu-Hammour et al. (2026) explored an environmentally enriched multimodal approach to olive disease classification and Berrocal et al. (2026) explored the combination of visual and soil-sensing information for maize disease severity assessment. Likewise, Deb and Rahman (2026) investigated the integration of UAV, satellite and IoT data in precision rice farming. The examples of these studies show the variety of agricultural applications where complementary data sources can be used to facilitate automated analysis. Similar methods have been explored in the context of rice crop health monitoring and agricultural sensing (Gupta and Chouhan, 2026; Li et al., 2026).

In the field of agricultural computer vision, RGB and thermal imaging have also been of interest. Begum et al. (2026) studied RGB and pseudo-thermal representations for plant disease classification, and Kulhari and Patwa (2026) studied the use of RGB and thermal information for leaf segmentation. El Sakka et al. (2025) conducted a review on the use of CNN in agricultural applications with multimodal data. Further research has considered environmental and crop-related information for agricultural monitoring (Vijayan and Chowdhary, 2026; Xiang et al., 2026; Xiao et al., 2026; Yang et al., 2025; Zhou et al., 2025).

Soil-related applications offer a specific link to environmental monitoring in agriculture. Guo et al. (2024) explored the use of multimodal feature fusion for the classification of heterogeneous soil data, and Miao et al. (2024) studied the identification of soil erosion using deep learning techniques from UAV images. Srivastava and Maity (2026) reviewed soil texture classification using AI and multimodal data fusion. Vahidi (2025) and Vahidi et al. (2025) discussed the issue of multimodal sensing for agricultural soil-moisture mapping. Zhou et al. (2021) summarized the application of infrared thermal imaging in crop water-stress assessment, while Zuo et al. (2023) studied the application of multimodal deep learning in tomato water-status detection.

Other studies show the use of multimodal sensing for purposes other than direct soil classification. Rana et al. (2026) discussed multimodal data analysis in precision farming, and Yuan et al. (2026) discussed machine-learning-based sensing for agricultural product quality and safety assessment. In the field of environmental monitoring, multimodal image analysis is explored in the segmentation of wildfires by Yue et al. (2025), who studied the fusion of RGB and thermal images. These studies provide related methodological and application-level views, but their research tasks are different from the agricultural soil classification.

2.6 Research Gaps

Multimodal deep learning, RGB imaging, thermal imaging and environmental sensing have been shown to have potential in agricultural monitoring in recent studies. But more research is needed to develop reliable agricultural soil classification systems based on paired RGB and thermal images. Representative agricultural datasets, validated soil categories, soil moisture and temperature variations, and the incorporation of independent field and laboratory measurements for environmental risk assessment are important challenges. The challenges offer future research opportunities for multimodal feature extraction and fusion methods for agricultural soil monitoring. The present experimental investigation is based on RGB and thermal feature fusion with road surface imagery, and needs to be tested separately with agricultural soil imagery.

3. Proposed Methodology

Based on the literatures, conventional systems employ RGB images for road surface classification. Many occasions those images fail to contribute all conditions of sidewalks under various environmental factors. Hence in the proposed model, we proposed a fusion model which tries to classify the sidewalk surface condition accurately in all type of weather conditions. The model fuses the RGB images and the thermal images which are captured by the vision system. This approach improves the robustness of sidewalk surface condition recognition under any type of environmental conditions.
The proposed model is carried out under four different stages as stated below.

a. Image data set preparation and preprocessing
b. Individual feature extraction
c. Fusing the feature vectors
d. Sidewalk surface classification and risk assessment.

The following figure depicts the workflow of the proposed model.

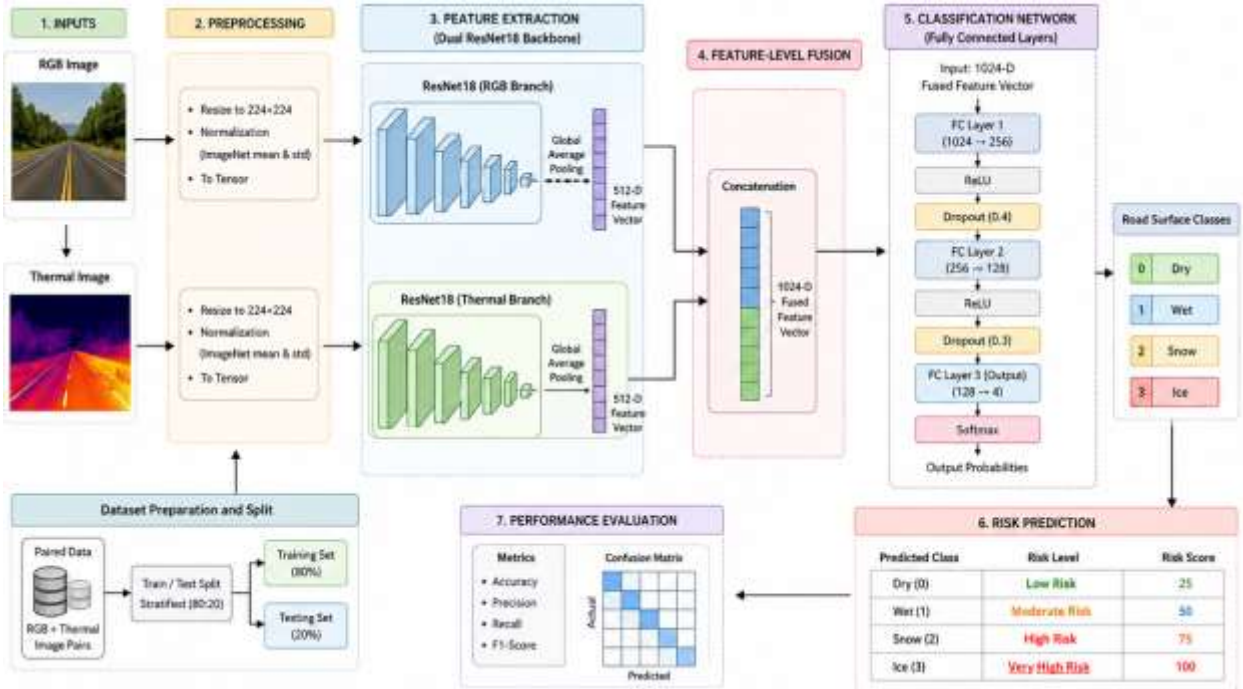


Figure 3.1: Work flow of the proposed methodology

3.1Image dataset preparation and Preprocessing

The vision system captured the paired image set which consists of the RGB image and the thermal image of the same location and the same skew shot. A total number of 1438 image sets were captured and used or training the proposed neural model. As illustrated in table 1, the captured images are grouped into different categories with different ID numbers for easy identification. This dataset preparation was handled manually in the Jetson Nividia system after the reception of images from the vision system.

Table 1: Image Categories

ID	Image Categories
01	Dry Pavement
02	Wet Pavement
03	Salted Pavement
04	Gravel Way
05	Puddle
06	Slush
07	Snow
08	Ice
09	Black Ice

An initial examination was conducted manually on all images under different categories. Through experience it was clearly evident that all these different nine categories of sidewalk surface to grouped or classified into a minimum number of classes in order to improve the performance of the system and to achieve accurate predictions. Hence the nine categories of sidewalk surface are later classified into four groups as Dry, Wet, Snow and Ice as depicted in Table number 2.

Table 2: Classification of Road Surface

ID	Image original categories	Target ID	Target Groups
01	Dry Pavement	T01	Dry
02	Wet Pavement	T02	Wet
03	Salted Pavement	T01	Dry
04	Gravel Way	T01	Dry
05	Puddle	T02	Wet
06	Slush	T03	Snow
07	Snow	T03	Snow
08	Ice	T04	Ice
09	Black Ice	T04	Ice

The final groups were assigned with numerical values in order to effectively classify the images during the model training and prediction.

Numerical Encoding: $Num(G) = \begin{pmatrix} Dry: 0; Wet: 1 \\ Snow: 2; Ice: 3 \end{pmatrix}$ …… (1)

The entire data set is sub divided into training sets and testing sets after encoding them into respective categories. The technically proven 80:20 stratified split is applied to the 1438 image pair. Totally 1150 data sets were used for training the model and 288 data sets were used for testing the model's performance.

The datasets are applied to the preprocessing stages as resizing and normalizing to maintain the uniformity in all acquired images. All images are resized to $\langle 224\ X\ 224 \rangle\ pixels$ to uniformly match the input dimension of ResNet18. During this stage it was ensured that all images are modified into uniform spatial dimension even though the captured image contains varied size and aspect ratio. As the ResNet18 accepts the fixed size input, resizing is mandatory to the specific size.

Once the resizing process is completed, each image set should be passed through the normalization. In order to normalize the image data, the entire image data set is encoded into the form of PyTorch tensors. During this transformation the standard image format (Height x Width x Channels) is been converted into the form of numerical tensor format (Channels x Height x Width). The transformed numerical tensor formatted image set will be given as the input for the PyTorch deep learning framework. Another important normalization is also adopted during this stage on the pixel values; the pixel values are scaled from the range 0 - 255 to 0–1.

The system then adopts the complete normalization through ImageNet dataset statistics which will consider the numerical tensor format data as the input. The normalization is focused towards stable gradient propagation and enables the network convergence during the model training.

The obtained mean and standard deviation values are depicted below.

$\mu = [0.485, 0.456, 0.406]$ …. (2)

$\sigma = [0.229, 0.224, 0.225]$ ….. (3)

Where,

μ : $mean\ pixel\ intensity\ and\ \sigma$: $SD\ for\ each\ color\ channel\ in\ the\ ImageNet\ dataset.$

As RestNet18 is a pretrained model and it is initially trained on the ImageNet dataset, the proposed model extracts the ImageNet statistics. This also enables the pre-trained feature representations to be used more efficiently during the training process. Experimentally it was determined that the gradients propagated more smoothly during the process of backpropagation by avoiding the vanishing gradient problem. The outcome enabled more stable gradient propagation, reduced time model convergence and improved training performance.

3.2Individual feature extraction (Dual Stream)

To improve the accuracy in training and prediction, the proposed model employs independent RestNet18 CNN. The features of RGB image and thermal images are extracted individually through these independent RestNet18 CNN models. The RestNet18 model architecture is given in figure 3.2.

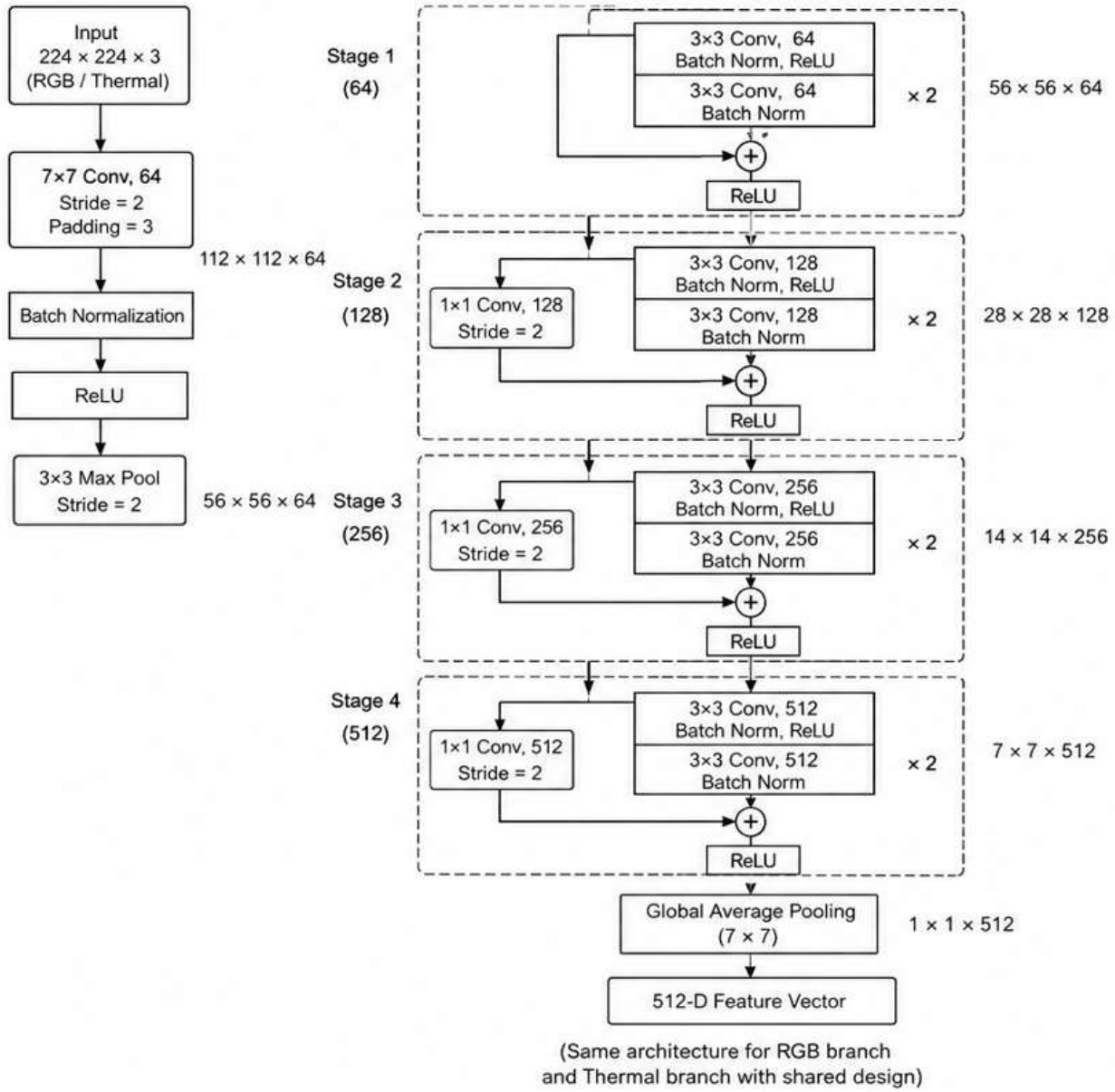


Figure 3.2: RestNet18 convolutional neural network model

As depicted in the above figure, the RestNet18 network accepts the images if size height as 224 pixels, width as 224 pixels and 3 RGB channels. The same input image size will be repeated for the thermal images. The initial convolution layer utilizes 7x7 convolution with 64 filters. The stride and padding are initialized

with 2, 3 respectively. This setup will extract the low-level features and capable of detecting color changes, textures, edges and simple patterns. This convolution stage will provide an output of image size as 112 x 112 x 64, which shows the feature maps are reduced to half and 64 feature maps were generated. After reducing the feature map, the images will be passed through batch normalization. The goal of this batch normalization stage is to normalize all feature values in order to attain a faster convergence while training and prediction. This also enables the networks to produce stable gradients and less internal covariate shift as given in equation 4.

$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \beta \text{ ….. (4)}$$

Where,

μ *is the batch mean,* μ *is the batch variance,* γ *is the scale parameter and* β *is the shift parameter.*

In order to introduces the non-linearity as like other neural network models, the proposed model utilizes the ReLU activation function. The RELU is chosen as it will provides faster computation and less expensive computation. ReLU removes the negative values efficiently. The ReLU activation function is illustrated in the following equation.

$$ReLU(x) = max(0, x) \text{ ….. (5)}$$

The next stage is to apply the image vectors to max pooling where the image size will be further reduced while preserving the important features in order to reduce the computation complexity. In this architecture, the 3x3 max pool with stride 2 is utilized and the output image size was 56 x 56 x 64.

3.2.1 Applying 64 filters in Stage 1

In stage 1 with 64 filters, each residual block consists of 3 x 3 convolutional neural network, batch normalization process and ReLU activation function. The entire block will run for two times and input will be copied directly as the dimensions of the input image remains unchanged. During the learning process, the residual learning is achieved as illustrated in equation 4.

$$O = F(x) + x \text{ ….. (6)}$$

Where O is for output, F(x) is the learned residual mapping and x is the input.

The major reason for applying the residual connection is instead of learning H(x), ResNet18 model learns $F(x) = H(x) - x$. Hence it changes to $H(x) = F(x) + x$. This allows the network model to train easier by adjusting the gradient flow to all deep hidden layers.

3.2.2 Applying 128 filters in Stage 2

During the stage 2, the feature map is reduced to half of its size as 28 x 28 x 128 where the filters are doubled to 128. As the dimensions are changed invariantly, the shortcuts cannot be applied as implemented in stage 1. Hence in stage 2, a projection shortcut method is utilized as 1 x 1 convolution and with stride values as 2. This is applied to the input image vector to match the dimensions of the channel. Also the input is down sampled and enabled with genuine residual addition.

The residual value is recalculated as

$$O = F(x) + W_s x \text{ ….. (7)}$$

Where W_s *is the projected convolution* $1 \; x \; 1$.

3.2.3 Stage 3 and Stage 4

During stage 3 and stage 4 the same process is repeated where the number of filters applied are 256 and 512 respectively. During stage 3, the network learns the higher-level semantic features like lane border markings, road boundary points and large texture patterns. Whereas in stage 4, the network extracts the abstract information about the road surface. The extracted information are cracks, snow, ice, wetness and thermal texture variations.

The output of the four stages of CNN is passed through the global average pooling process (GAP). For the GAP, the input vectors are accepted in the dimension of 7 x 7 x 512 and on each feature map the global average pooling is computed as given in equation 7.

$$G_{v = \frac{1}{H \; x \; W}} \sum_{i=1}^{H} \sum_{j=1}^{W} f_v(i,j) \text{ ….. (7)}$$

Where the values of H and W are 7 respectively and $f_v(i, j)$ is the activation at location (i,j) of v^{th} feature map.

Finally, the output feature vector will be of size 1 x 1 x 512 or simply 512- D vector. The same process will be repeated for the thermal images and a 512-d Vector feature map will be generated.

3.3Fusing the feature vectors

The major strength of the proposed model is the strong features which are obtained from the vision system. The side walk images are captured in RGB and Thermal format and independently those features are extracted. Carefully, those extracted features are to be fused to form the strong trained model. The extracted features from the RGB images such as sidewalk surface color, pattern, texture and its appearance, whereas the thermal images provided the contrast and surface temperature distributions. Available models depend on independent features to classify the side walk, whereas the proposed integrated modality allows the classification model to analyze the side walk carefully with more integrated feature dependent.

The extracted features are given as;

$Features_{RGB}^{i} \in \mathbb{R}^{512}$ ….. (8)

$Features_{Th}^{j} \in \mathbb{R}^{512}$ ….. (9)

After extracting the individual features, the features are concatenated carefully without loosing any feature representations. The concatenation technique is applied to safeguard all extracted features from the two different modalities and to fuse the features without adjusting any weighting coefficients.

For the i[th] feature sample the fused (concatenated) sample is represented as:

$\left[\left[Features_{fused}^{i} = \left[Features_{RGB}^{i} \oplus Features_{\mathrm{Th}}^{\mathrm{j}}\right]\right]\right]$ ….. (10)

The fused representation is derived as ,

$\left[\left[\mathrm{Features}_{\mathrm{fused}}^{\mathrm{i}} = [\mathrm{r}_1, \mathrm{r}_2, \ldots . \mathrm{r}_{512}, \mathrm{t}_1, \mathrm{t}_2, \ldots . \mathrm{t}_{512}]\right]\right]$ ….. (11)

The fused feature vector dimensionality will be changed to 1024 as;

$\mathrm{Dim}_{\mathrm{fused}} = \mathrm{Dim}_{\mathrm{RGB}} + \mathrm{Dim}_{\mathrm{TH}}$ ….. (12)

Which is $\mathrm{Dim}_{\mathrm{fused}} = 512 + 512$

Hence,

$\left[\left[\mathrm{Features}_{\mathrm{fused}}^{\mathrm{i}} = \mathbb{R}^{1024}\right]\right]$ ….. (13)

The final 1024-dimensional feature vector comprises of the features extracted independently from the two different network models. The feature concatenation provides 1024-dimension feature vectors which will increase the input features processed by the classification model. The proposed model uses a fully connected network layers which will progressively reduce the dimensionality as illustrated below.

1024→256→128→4

On the hidden space the features are reduced from 1024 dimensions to 254 using the dimensionality reduction technique illustrated below.

$\mathrm{H}_1 = \mathrm{RELU}\,(\mathrm{W}_1 * \mathrm{Features}_{\mathrm{Fusion}} + \mathrm{bias}_1)$ ….. (14)

Later, with the probability value of $\mathrm{P}_1 = 0.4$ is applied as given in the equation.

$\widetilde{\mathrm{H}_1} = \frac{\mathrm{M}_1 \odot \mathrm{H}_1}{1 - \mathrm{P}_1}$ ….. (15)

The later subsequent hidden layers will transform the dimensions to 128 and to 4 to process the features uniformly. The resulting vector will form the unnormalized feature set for the four categories of sidewalk classification.

3.4Side Walk Surface classification and risk assessment.

The final stage of the proposed methodology is to classify the side walk and to assess its risk for pedestrian walking. To carry out the classification process, the dimensionality reduced fused features class probabilities are computed using the softmax function. The softmax probability computation is given in equation.

$\mathrm{P}(\mathrm{y} = \mathrm{c} \mid \mathrm{I}_{\mathrm{RGB}}, \mathrm{I}_{\mathrm{TH}}) = \frac{\mathrm{e}^{\mathrm{z}_c}}{\sum_{\mathrm{j}=1}^{4} \mathrm{e}^{\mathrm{z}_j}}$ ….. (16)

The predicted side walk surface class is determined by:

$\hat{y} = \arg \max_{\mathrm{c} \in \{0,1,2,3\}} \mathrm{P}(\mathrm{y} = \mathrm{c} \mid \mathrm{I}_{\mathrm{RGB}}, \mathrm{I}_{\mathrm{TH}})$ ….. (17)

The proposed methodology utilizes the above regularization as it is relevant when a high dimensional fused representations with less training data.

The output classification set will be ,

$\hat{y} \in \{\mathrm{Snow}, \mathrm{Ice}, \mathrm{Dry}, \mathrm{Wet}\}$ ….. (18)

Mathematically, the risk assessment is represented as,

$R = \{25, \hat{y} = \text{Dry};\ 50, \hat{y} = \text{Wet};\ \hat{y} = 75, \text{Snow};\ \hat{y} = 100, \text{Ice} \ldots\ldots (19)$

The proposed methodology and framework lay a foundation for future sidewalk conditions for pedestrians and road conditions awareness for motorists when the system is incorporated with vehicle dynamics, environmental sensing and real-world validation.

4. Results and Discussion

The system was experimentally evaluated to understand the effectiveness in classifying the side walk surface condition with the RGB and thermal images pair. The experimentation system focused on four side walk surface conditions: Snow, Ice, Dry and Wet. The collected dataset is divided into training and testing dataset using the 80:20 stratified sampling strategy with the fixed random seed of 42. The collected RGB and Thermal image pairs was resized into 224 X 224 pixels and normalized using ImageNet mean and standard deviation methods.

As described in the previous section, the illustrated architecture employs two independent ResNet18 networks for RGB image set and Thermal image set. The network was trained with Adam Optimizer with the learning rate of 10^{-4} and the cross-entropy loss algorithm. The proposed model performance was evaluated using classification accuracy, precision, recall and F1-Score. The evaluation metrics were used to find the differences in the test data set. The implementation also supports the generation of class- wise classification reports and confusion matrices to examine the prediction behavior for the four side walk surface categories. Table 3 depicts the different parameters and its configurations used by the proposed model.

Table 3: Different parameters and its configurations

No	Parameter	Configuration
1	Input Data set	RGB and Thermal Image Pair
2	Target side walk classification	Snow, Ice, Dry and Wet
3	Input Image resolution	224 x 224
4	Feature Dimensions	RGB : 512, Thermal : 512
5	Feature Dimension after fusion	1024
6	Dropout Probabilities	0.4, 0.3
7	Training epochs	10, 30 (Extended)

4.1Comparative Analysis

To evaluate the sensitivity of the model, several experimentations carried out on the dataset. The evaluations include baseline configuration, transformation only processing, mild augmentation, strong augmentation, revised transformation without augmentation and extended training. Table 4 shows the evaluation of different metrics under different experimental configurations.

Table 4: Evaluation of metrics under different experimental configurations

Experimentation	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Baseline	85.38	86.03	85.38	82.06
Strong augmentation	66.15	60.8	66.15	62.53
Transformation only	89.23	82.83	89.23	83.48
Mild augmentation	77.69	71.92	77.69	73.21
Revised transformation, no augmentation	**89.23**	**91.47**	**89.23**	**85.11**
Extended training	81.54	79.49	81.54	79.74

As illustrated in the table, the revised transformation with no augmentation gives the highest accuracy and precision and that is visualized in the graph below. It is well understood that when the system is configured with augmentation, the accuracy is dipping down heavily and also the by increasing the number of epochs, the accuracy gain is less.

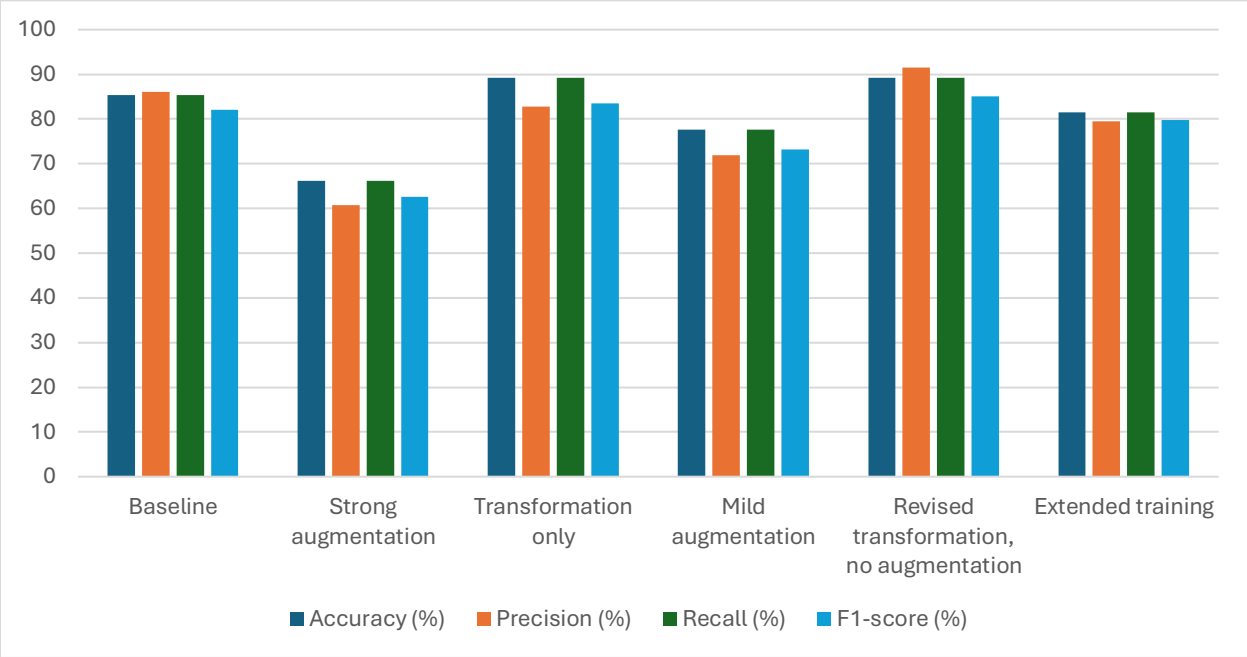


Figure 4.1: Illustration of evaluation metrics under different parameters

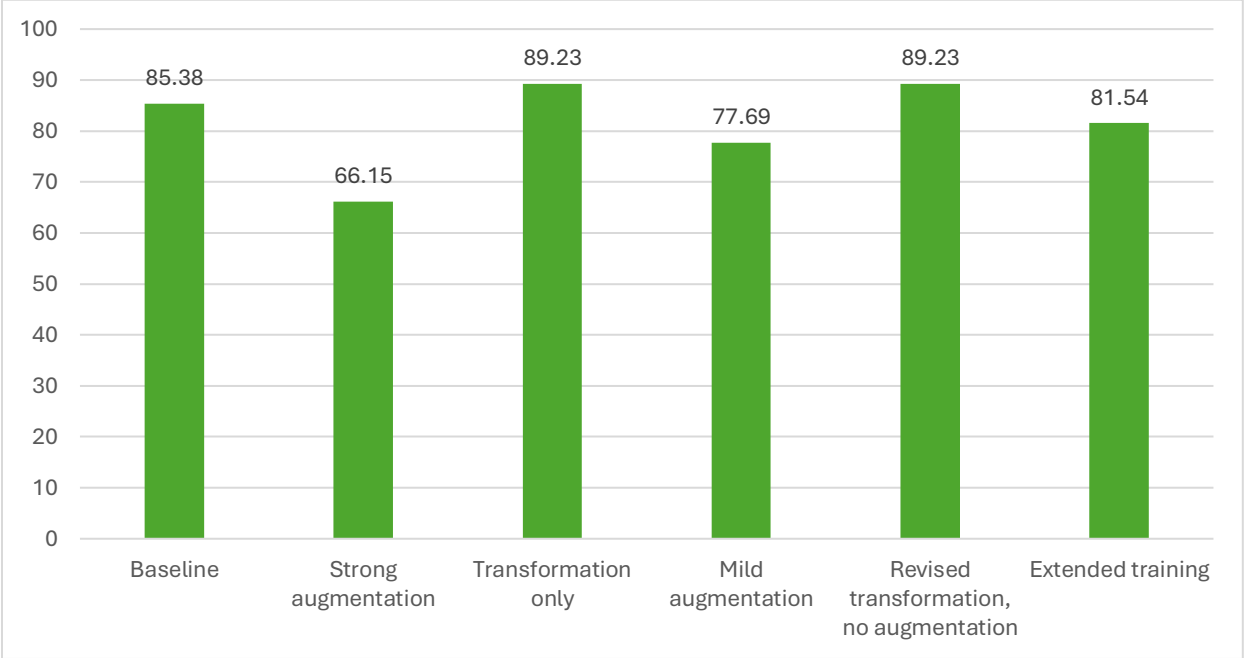


Figure 4.2: Illustration of accuracy measure under different parameters

4.2Training Loss and Convergence behavior loss

During the successive epochs, the convergence characteristics is recorded. Based on the recorded characteristics, it is observed that the various preprocessing and different augmentation strategies considerably triggered the optimization performance of the two channel ResNet architecture. Figure 4.3 shows the training loss under various configuration for ten epochs.

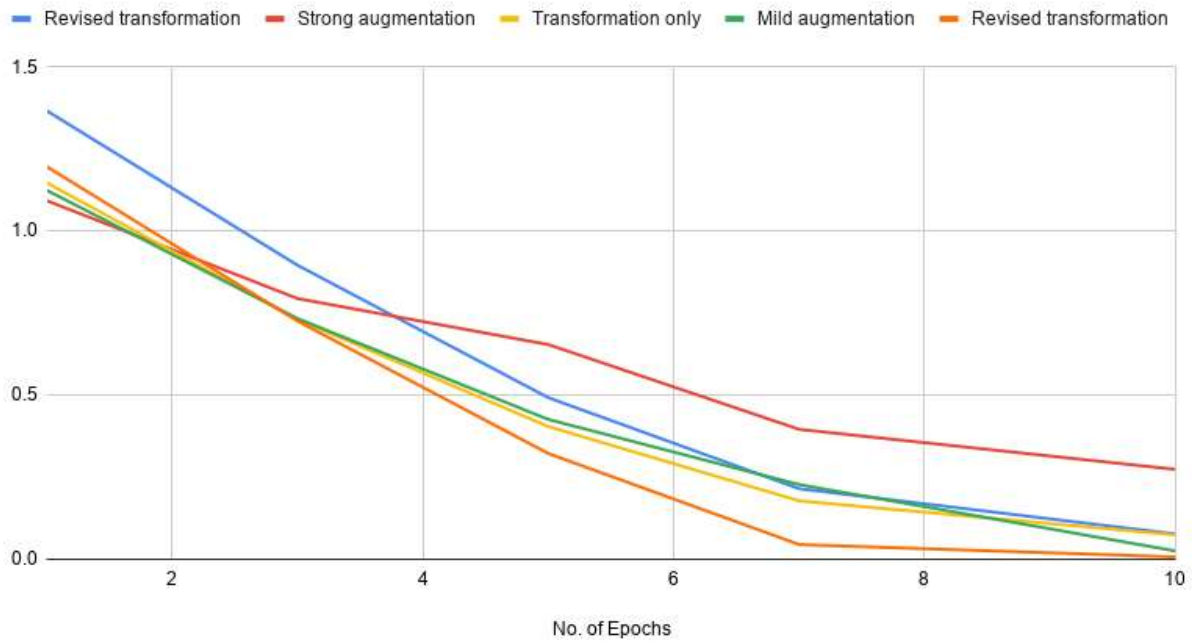


Figure 4.3 Training loss vs No. of Epochs

The above figure clearly indicates the overall reduction of training loss in various configurations which leads to the conclusion that the optimization algorithm successfully reduced the training loss which is the training objective. It is also noted that the convergence rate is differed in various configurations.
The complementary RGB and thermal data also has potential applications outside of road-surface analysis. In agricultural settings, the visible and thermal properties can be used to monitor soil-surface conditions and moisture and temperature variations (Vahidi et al., 2025; Zhou et al., 2021). But, for agricultural soil classification and environmental risk assessment, there is a need for dedicated agricultural datasets and independent field or lab measurements (Guo et al., 2024; Miao et al., 2024; Srivastava and Maity, 2026). The framework proposed in this research is thus not an experimentally proven soil classification system, but rather a possible methodological basis for further research in agriculture.

5. Conclusion

The multimodal deep learning framework proposed for sidewalk surface classification provides a risk alert to all users and provide active advises of the side walk conditions. The systems showcase a reliable accuracy in the risk assessment and hence the proposed architecture can be embedded with any autonomous system or handheld devices. Even though the proposed model exemplifies the feasibility of integrating the RGB and thermal images successfully for side walk surface classification, there are few limitations that to be acknowledged. The obtained results show sensitivity to augmentation intensity and training duration. Another limitation is the lack of environmental metadata integration. Future studies can be focused on larger datasets, attention-based feature fusion mechanisms which can incorporate the road surface classification too. Future research may extend the multimodal framework to agricultural soil monitoring by incorporating paired RGB and thermal images, verified soil categories, and field-based environmental measurements. Such investigations could examine the applicability of multimodal deep learning to soil-condition classification and environmental risk assessment under different agricultural conditions.

References

1. Abu-Hammour, Z. S., Mashagbeh, M. F. A., AlSmadi, N. M., Altalla, E. N., Ayasrah, A. B., Alnasra, H. A., & Almanasir, I. H. (2026). Toward Smart Agriculture: A Novel Environmentally Enriched Multimodal Deep Learning Framework for Olive Peacock Spot Disease Stage Classification and Severity Estimation. Applied Sciences, 16(15), 7669.

2. Begum, S., Naresh, E., & Srinidhi, N. N. (2026). Hybrid deep learning-based multimodal framework for plant leaf disease classification using RGB, Excess Green (ExG), and pseudo-thermal representations with MobileNetV2. Scientific Reports.
3. Berrocal, O., Sarmiento, C., & Figueroa, J. (2026). Multimodal Deep Learning Model for Real-Time Severity Assessment of Maize Stalk Rot Integrating Visual and Soil Sensing Data. Smart Agricultural Technology, 102038.
4. Deb, N., & Rahman, T. (2026). Multi-modal Deep Learning Framework Using UAV, Satellite, and IoT Data for Precision Rice Farming and Sustainable Agricultural Production. Sensing and Imaging, 27(1), 190.
5. El Sakka, M., Ivanovici, M., Chaari, L., & Mothe, J. (2025). A review of CNN applications in smart agriculture using multimodal data. Sensors, 25(2), 472.
6. Guo, Q. M., Zhan, L. T., Yin, Z. Y., Feng, H., Yang, G. Q., & Chen, Y. M. (2024). Multi-modal fusion deep learning model for excavated soil heterogeneous data with efficient classification. computers and Geotechnics, 175, 106697.
7. Gupta, S., & Chouhan, S. A. (2026). A multimodal deep learning for rice crop health monitoring using vegetation index feature selection. Scientific Reports.
8. Kulhari, A., & Patwa, S. (2026). An adaptive attention-based multimodal deep learning framework integrating RGB and thermal imaging for robust leaf segmentation in real-world environments. International Journal of System Assurance Engineering and Management, 1-13.
9. Li, L., Song, Y., Zhang, Y., Guo, P., Wang, X., Ma, Z., & Yan, S. (2026). An AI-Driven Multimodal Sensing Framework Integrating UAV Imagery and Environmental Sensors for Intelligent Farmland Monitoring. Sensors, 26(11), 3456.
10.Miao, S., Liu, Y., Liu, Z., Shen, X., Liu, C., & Gao, W. (2024). A novel attention-based early fusion multi-modal cnn approach to identify soil erosion based on unmanned aerial vehicle. IEEE Access, 12, 95152-95164.
11.Rana, N., Dahiya, P., Mehta, S., & Ladohia, S. (2026). Introduction to Multimodal Data Analysis in Precision Farming: Challenges, Techniques and Applications. In Multimodal Artificial Intelligence in Precision Agriculture (pp. 3-24). CRC Press.
12.Srivastava, A., & Maity, R. (2026). AI-Driven Soil Texture Classification: Integrating Machine Learning, Deep Learning, and Multimodal Data Fusion for Precision Soil Conservation. In Advancements in Soil Conservation: From Traditional Practices to AI-Driven Solutions (pp. 401-433). Cham: Springer Nature Switzerland.
13.Vahidi, M. (2025). Advancing Soil Moisture Mapping with Multi-Modal Drone Sensing and AI: Integrating GPR, Hyperspectral, and Thermal Imaging for Precision Agriculture.
14.Vahidi, M., Shafian, S., & Frame, W. H. (2025). Multi-Modal sensing for soil moisture mapping: Integrating drone-based ground penetrating radar and RGB-thermal imaging with deep learning. Computers and Electronics in Agriculture, 236, 110423.
15.Vijayan, S., & Chowdhary, C. L. (2026). RiceMultiModalNet: a proof-of-concept biologically constrained multi-modal deep learning for rice disease detection integrating phenological stage and environmental risk. Scientific Reports.
16.Wang, D., Cao, W., Zhang, F., Li, Z., Xu, S., & Wu, X. (2022). A review of deep learning in multiscale agricultural sensing. Remote Sensing, 14(3), 559.
17.Xiang, J., Pan, J., Zhang, H., Liu, X., Liu, B., Tian, J., & Yan, S. (2026). Multimodal Deep Learning for Pest and Disease Recognition and Crop Growth Assessment in Open-Field Agricultural Environments. Agriculture, 16(13), 1414.
18.Xiao, S., Chen, Y., Lu, M., Wang, J., Guo, R., Xu, X., & Song, Y. (2026). Multimodal Method for Pest Recognition Using Field Images and Environmental Data in Smart Agriculture. Agriculture, 16(12), 1268.
19.Yang, Z. X., Li, Y., Wang, R. F., Hu, P., & Su, W. H. (2025). Deep learning in multimodal fusion for sustainable plant care: A comprehensive review. Sustainability, 17(12), 5255.
20.Yuan, Z., Wang, X., Wu, X., & Zhang, X. (2026). Multimodal Sensing and Machine Learning for Nondestructive Quality and Safety Assessment of Agricultural Products. Chemosensors, 14(10), 233.
21.Yue, T., Huang, H., Wang, Q., Song, B., & Chen, Y. (2025). A multimodal deep learning framework for accurate wildfire segmentation using RGB and thermal imagery. Applied Sciences, 15(18), 10268.
22.Yumang, A. N., Paglinawan, A. C., & Dingel, B. D. (2026). EIN+Y: An Einstein activation function based YOLOv8 with alpha blending for road surface condition classification. Franklin Open, 16, 100708.

23.Zhou, C., Cao, Y., Ming, B., Luo, J., Xu, F., Zhang, J., & Dong, M. (2025). A multimodal deep learning framework for intelligent pest and disease monitoring in smart horticultural production systems. Horticulturae, 12(1), 8.
24.Zhou, Z., Majeed, Y., Naranjo, G. D., & Gambacorta, E. M. (2021). Assessment for crop water stress with infrared thermal imagery in precision agriculture: A review and future prospects for deep learning applications. Computers and Electronics in Agriculture, 182, 106019.
25.Zuo, Z., Mu, J., Li, W., Bu, Q., Mao, H., Zhang, X., ... & Ni, J. (2023). Study on the detection of water status of tomato (Solanum lycopersicum L.) by multimodal deep learning. Frontiers in Plant Science, 14, 1094142.