Data Engineering and Applications

Research Paper

ConvNeXtV2-EMA Model for Airport Runway Object Detection with Structure-Preserving Representation Fusion

Yazeed Al-Shamisi[1,*] and Jassim Al-Buflasa[1]

[1] School of Information Technology, American University of Sharjah, Sharjah, PO Box 26666, Unightd Arab Emirates
*Corresponding author: yazded.sha@aus.edu

Abstract. A Structure-Preserving ConvNeXtV2-EMA Model is Proposed for Airport Runway Object Detection Under Multiscene Visual Disturbance. Runway monitoring is very important for safety because in the same long-distance view, small foreign object debris, service vehicles, personnel, wildlife, lighting equipment, shadows, markings, wet reflections, and surface damage may all appear. The proposed framework normalises the runway images and creates a runway structural mask based on line direction, surface boundary, and marking continuity. Then, it combines this with the target-region candidates before they happen. ConvNeXtV2 backbone removes hierarchical features, and an effective multi-scale attention module refines object-like regions and eliminates runway textures false activations. A dataset of 13,600 images from daylight, night, rain-after, backlight, and long-distance monitoring scenes has been annotated with six target categories and safety-critical missed-detection labels. The proposed model has improved average precision to 91.3%, small-object recall to 86.9%, and safety-critical missed detections to 2.1% compared to Faster R-CNN, YOLOv5, Swin-based detection, ConvNeXt, and a non-EMA variant. Therefore, while runway geometry and focus-enhanced feature fusion may improve visual inspection reliability, real airport camera data, temporal verification, and operational records are still needed before deployment.

Keywords: *Runway Object Detection, ConvNeXtV2-EMA, Structure-Preserving Representation, Airfield Safety Monitoring*

Received on 11 August 2023, Accepted on 08 December 2023, Published on 15 December 2023

Introduction

Aircraft take-off and landing can be affected by foreign object debris, service trucks, personnel, animals, lighting equipment, damaged surface elements, etc., so airport runway object detection is a safety-critical computer-vision problem. Runway monitoring often uses long-distance fixed cameras or vehicle-mounted inspection cameras, so the safety area may be relatively small in the image. General detection studies [1] have shown that scale, occlusion and background complexity all impact detection performance, and deep detection reviews [2] have indicated that both semantic discrimination and localisation accuracy are required in dense monitoring scenes. Airport surfaces have runway markers, rubber deposits, joints, light strips, shadows and tire marks, which may be mistaken for safety-related objects. Remote-sensing interpretation research [3] has also found that local texture can be misleading without context. Therefore, runway detection should be regarded as a structure-aware perception problem rather than simple foreground extraction.

Operationally safe refers to a hazard that cannot be seen or reached by the plane. Early recognition under backlight, night light, rain-after-reflection, low contrast, heat haze and long camera distance is required. Data augmentation research [4] expands the range of appearances, but it cannot address the problem of small objects on a highly structured runway surface by itself. Dense detection loss design [5] shows that small and rare targets are affected by class imbalance, and real-time detector research [6] has shown the demand for high-speed safety monitoring. Efficiently scalable detection [7] also indicates that the deployed detector should have a relatively low computational cost. Therefore, a practical runway detector should maintain runway geometry, strengthen the weak target evidence, and be fast enough for inspection use.

Attention mechanisms and new visual backbones are also employed in this paper now. Hierarchical Vision Transformers [8] show that multi-scale context improves dense prediction, and attention-mechanism studies [9] explain why useful spatial and channel cues should be emphasized without losing global structure. ConvNeXt research [10] shows that the modern convolutional backbone can achieve transformer-level representation quality with efficient inference. Therefore, the ConvNeXtV2-EMA model has been proposed to integrate structure-preserving representations with efficient multi-scale attention. Increase the general detection accuracy, reduce safety-critical missed detections, and provide an explanation for why the detector differentiates between real runway objects and surface artifacts or reflections.

The following five sections make up this document. Airport-surface inspection, small-object detection, structure-preserving representation, and attention-enhanced detection backbones are introduced in Section 2. In addition to describing the runway input encoding, structural mask creation, EMA-based feature refinement, detection loss, and safety-oriented scoring, Section 3 presents the suggested ConvNeXtV2-EMA model. The dataset creation, baseline comparison, safety-error analysis, ablation findings, robustness assessment, and deployment interpretation are all displayed in Section 4. The research's ultimate result is presented in Section 5.

Related Work

Runway Foreign Object and Intrusion Detection

When monitoring airport runways, keep structured surface patterns apart from important objectives. Spatial and channel emphasis can assist localise salient visual areas, as demonstrated by convolutional attention modules [11]. However, additional attention is needed for runway pictures since the salience may be painted marks or light reflections rather than a target. Lightweight attention can improve feature reaction without substantial overhead, which is advantageous for airport inspection systems, according to research on high-efficiency channels attention [12]. According to global-context modelling [13], a location should be viewed in light of its surroundings. A dark patch close to a centerline marker in runway detection should be categorised differently from the same patch in an open-area area. In order to maintain the runway-line orientation, lane-like geometry, and boundary connections, the model should have structure-preserving characteristics.

Compared to generic road or urban datasets, airport-specific labelled data are comparatively rare, making transfer and pre-training applicable. While generic characteristics can be useful for downstream tasks, direct transfer is not appropriate when the desired appearance and scene geometry mismatch, according to research on ImageNet pretraining [14] and transferability of visual models [15]. Class imbalance research [16] has explained why uncommon safety items may be under-detected if the training goal is dominated by normal background, while robustness studies [17] have demonstrated how common corruptions can deteriorate neural models. According to the study, the detector should be evaluated in a multi-scene disturbance environment instead of a single clean-image set for runway inspection.

Deep Detection Models and Structure-preserving Attention

Other non-convolutional detection methods include transformer-based object detection [18] and detector pretraining [19], although these could be too computationally costly for real-time runway surveillance. While mixed-sample augmentation [20] enhances generalisation, it may also obscure the structural link between tiny objects and runway markers if it is not used in conjunction with geometric control. Multi-level fusion is appropriate for objectives at various sizes, according to research on feature aggregation [21]. The issue of managing close-range service vehicles and long-range tiny targets in the same system is likewise addressed by scale-aware detection [22]. The current model incorporates these concepts as multi-scale EMA feature fusion and a structure-preserving runway mask.

Airport runway photos have a variable mistake cost, however modern detection toolboxes and benchmarks [23] are useful for comparing algorithms. The inspection schedule will be disrupted by a false alert, which might prevent the detection of a small item. Fast detectors can facilitate field deployment, as demonstrated by real-time detection research [24] and single-shot localisation research [25]. However, the target/background boundary remains challenging under glare, rain-after reflection, and long-distance compression. Accuracy, speed, and model size must be balanced in practice, as demonstrated by lightweight hybrid representation [26]

and scaled detector design [27]. Recursive feature pyramids [28] and content-aware feature reassembly [29] both suggest that feature fusion should be spatially aware. The concepts are introduced by the ConvNeXtV2-EMA model in order to maintain the extended runway structure and enhance runway objectives.

Proposed Structure-Preserving Detection Framework

Runway Scene Representation and Structural Mask Construction

Runway-surface, marking, edge, and candidate-object areas are separated from the input runway picture after it has been normalised. As seen in Figure 1, the airport runway object-detection workflow comprises multi-scene image input, surface normalisation, runway-line prior extraction, candidate-object construction, structure-preserving feature encoding, ConvNeXtV2-EMA detection, and safety-oriented output. This shows how runway geometry and visual appearance are jointly processed before being translated into detection scores and warning priorities.

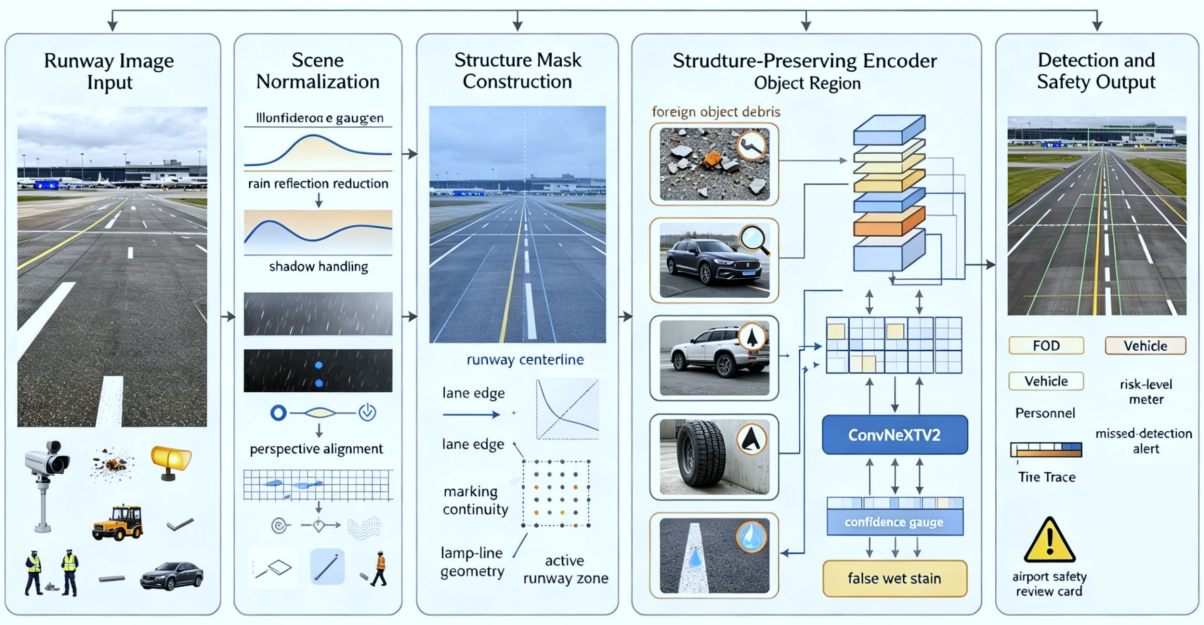

Figure 1. Airport runway object-detection workflow with structure-preserving representation.

Let I denote the original runway image and let N be the normalization operator. The normalized input is written as:

$$I_n = N(I) \quad \text{Eq.(1)}$$

The normalization step reduces illumination variation caused by night lamps, rain-after reflection, backlight, and camera exposure. It does not remove runway markings because those markings provide useful structure for later target/background separation. A runway structural mask is generated from line direction, surface boundary, and marking continuity. The mask is defined as:

$$M_s = G_l(I_n) + G_b(I_n) + G_m(I_n) \quad \text{Eq.(2)}$$

Here, G_l, G_b, and G_m represent line-direction extraction, runway-boundary extraction, and marking-continuity extraction. The mask is not a detection label; it gives the backbone a structured reference so that elongated runway patterns are not confused with independent objects. The target-candidate input combines the normalized image and the structural mask:

$$X = \text{concat}(I_n, M_s) \quad \text{Eq.(3)}$$

This construction allows the detector to process appearance evidence and runway geometry together. Small targets can then be evaluated against their structural surroundings rather than against local texture only.

https://doi.org/10.64972/dea.2023.v2i4.3839d:109-120

ConvNeXtV2-EMA Feature Interaction and Warning Output

The ConvNeXtV2 backbone extracts hierarchical feature maps from the structured input. ConvNeXtV2 stages, EMA refinement, multi-scale aggregation, detecting heads, confidence calibration, and warning output comprise the model's structure in Figure 2. Runway edges and little item details are seen in the lower stages, whereas object types and expansive surroundings are found in the later stages. The detector can distinguish between real safety objects and marks, seams, reflections, and other surface patterns thanks to EMA refinement, which also enhances the object-like response and decreases false activations brought on by stretched runway textures.

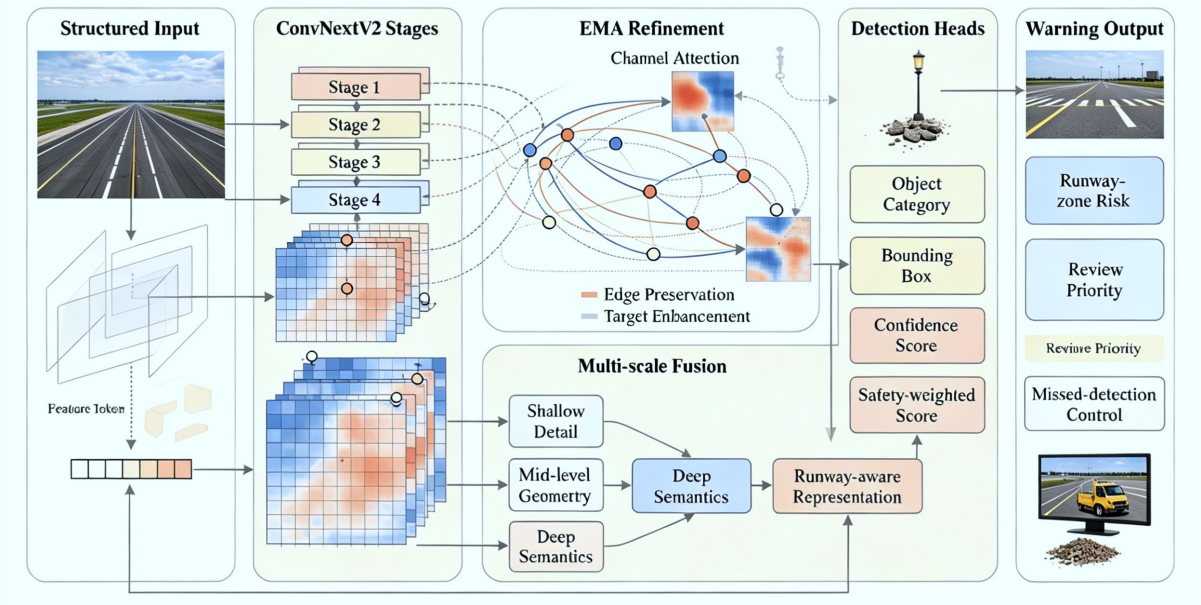

Figure 2. ConvNeXtV2-EMA feature refinement and safety-oriented detection head.

The feature map at stage k is represented as:

$$F_k = B_k(X) \quad \text{Eq.(4)}$$

where B_k is the kth ConvNeXtV2 stage. Lower stages retain edge and marking details, while deeper stages encode target category and context. The EMA module refines each feature map through spatial and channel interaction:

$$F_k^e = EMA(F_k) \quad \text{Eq.(5)}$$

The refined feature F_k^e gives greater response to object-like regions while reducing false activation from runway texture. Multi-scale fusion is then applied across feature levels:

$$F = \sum_{k=1}^{K} w_k F_k^e \quad \text{Eq.(6)}$$

The weights w_k control the contribution of each scale. Small debris requires shallow spatial detail, while vehicles and personnel require deeper semantic context. The fused feature is sent to the detection head:

$$Y = D(F) \quad \text{Eq.(7)}$$

The output Y contains category probability, bounding-box location, and confidence score. Because airport inspection cares about safety-critical missed detection, the training objective combines localization, classification, and safety weighting:

$$L = L_{cls} + \lambda_1 L_{box} + \lambda_2 L_{safe} \quad \text{Eq.(8)}$$

The safety term increases the penalty for missed foreign objects, personnel, or vehicles in active runway regions. Finally, a warning score is computed for each detected target:

$$S_o = P_o \cdot R_o \cdot C_o \quad \text{Eq.(9)}$$

https://doi.org/10.64972/dea.2023.v2i4.3839d:109-120

The warning score combines object confidence P_o, runway-region risk R_o, and context reliability C_o. This design separates ordinary visual confidence from operational urgency, which is necessary when the same object class appears in different runway zones.

Results and Discussion

Scene Data, Target Categories and Annotation Quality

A simulated airport runway inspection dataset is constructed from 13,600 images covering daylight, night, rain-after, backlight, and long-distance monitoring scenes. Reliable localization-quality estimation research [30] suggests that both object category and bounding-box confidence should be documented because runway targets are often small and boundary-sensitive. The dataset structure is displayed in Figure 3. Figure 3(a) shows scene-type proportions; night and rain-after views are intentionally retained because glare, reflections, and lamp-induced shadows increase difficulty. Figure 3(b) shows target-category proportions for foreign object debris, service vehicles, personnel, wildlife, lighting equipment, and surface-damaged objects. Figure 3(c) shows target-size intervals, and Figure 3(d) reports annotation-quality groups for high-confidence and uncertain labels.

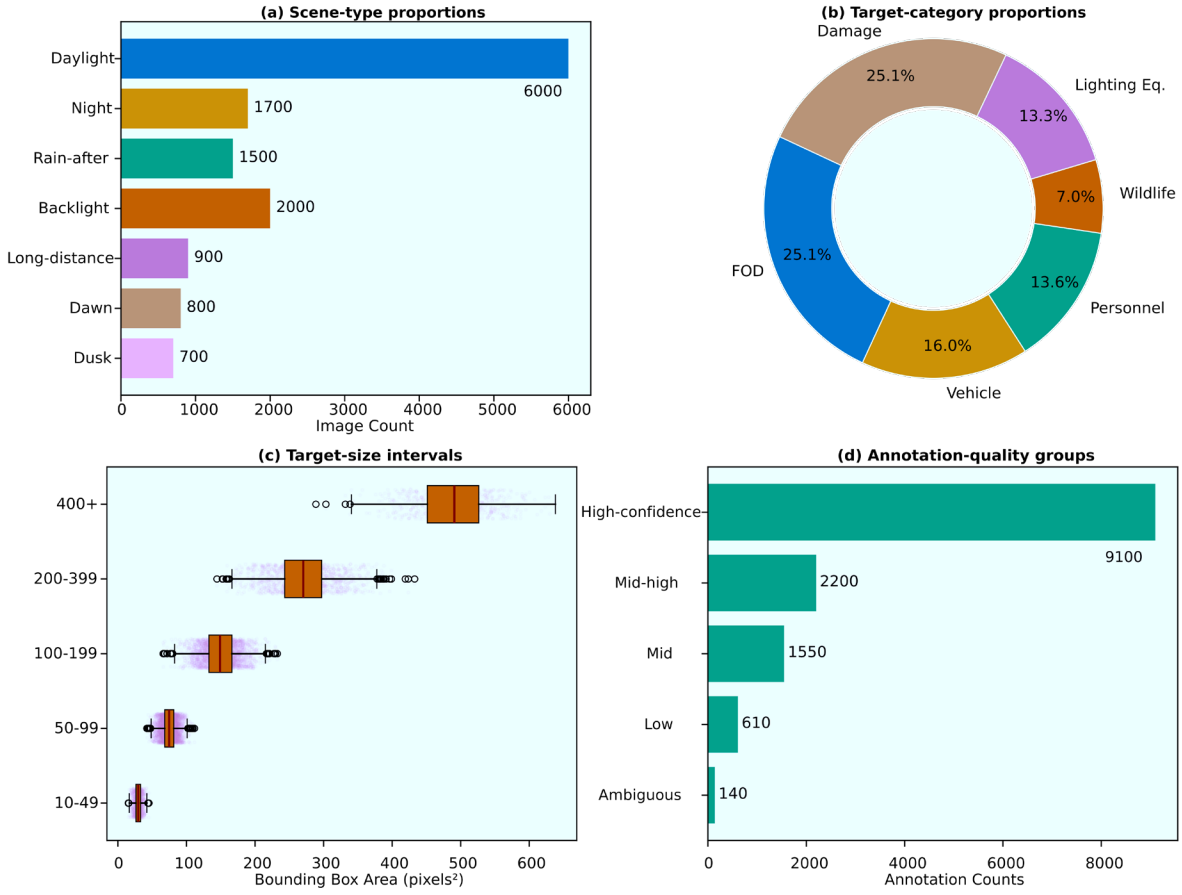

Figure 3. Runway dataset composition and annotation profile. (a) Scene-type proportions. (b) Target-category proportions. (c) Target-size intervals. (d) Annotation-quality groups.

Airport runway detection cannot be simplified to general object recognition, as the dataset distribution demonstrates. Despite having the same pixel area, a runway marker edge and a little metal item serve separate safety functions. The model can determine if a candidate is inside an active runway area, near a mark, adjacent to a light strip, or at an edge by using a structure-preserving mask. For some items that are partially covered by glare or shade, annotation-quality grouping is also necessary. Given the, it would not be suitable to use a single confidence criterion for all samples. As a result, the dataset will confirm that the detector is capable of handling all item sizes, runway zones, lighting, and annotation mistakes.

https://doi.org/10.64972/dea.2023.v2i4.3839d:109-120

Detection Performance and Safety Error Analysis

Faster R-CNN, YOLOv5, Swin-based detection, the ConvNeXt baseline, and a ConvNeXtV2 model without EMA refinement are compared with the proposed ConvNeXtV2-EMA model. Explainable deep-learning research [31] suggests that performance should be linked to decision rationale in safety-critical applications. The detection results are displayed in Figure 4. Figure 4(a) shows mean average precision, where the proposed model reaches 91.3% and outperforms both ConvNeXt and the non-EMA variant. Figure 4(b) reports small-object recall, where the proposed model reaches 86.9%. Figure 4(c) reports safety-critical missed detections, which decrease from 6.4% in the strongest baseline to 2.1% in the proposed model. These results show that the safety-weighted objective changes behavior in high-risk runway regions.

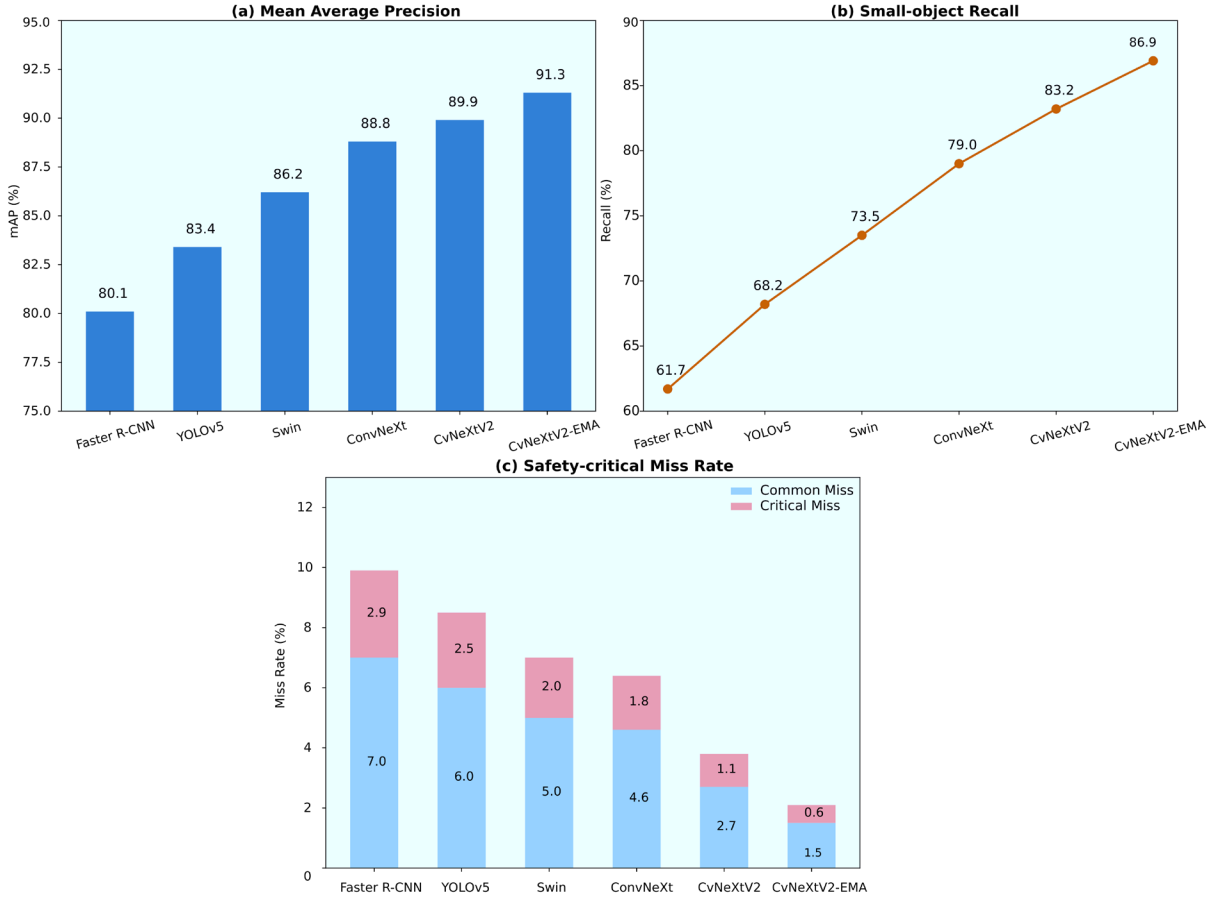

Figure 4. Detection performance comparison. (a) Mean average precision. (b) Small-object recall. (c) Safety-critical missed detection rate.

As a result, the Runway Structure will aid in increasing detection precision. Runway markings and shadow edges are often treated as regular backdrop by a generic detector. The suggested model may be evaluated in relation to the local runway environment since it employs a structural mask to maintain the desired orientation and continuity of runway features. It works well for rain-after photos that use reflections to produce bright areas that resemble objects. Additionally, the EMA module may further distinguish between extended surface patterns and weak target textures. As a result, the model boosts the confidence of detections where the runway context and target evidence coincide rather than all detections.

Error causes are examined in Figure 5. Figure 5(a) shows false-alarm categories, where runway marker edges and lamp reflections are the main confusing sources. Machine-learning interpretability research [32] supports identifying which visual structures mislead the detector. Figure 5(b) shows missed-detection categories, where low-contrast debris and backlight personnel remain difficult. The proposed method reduces these misses by preserving small-target response in the attention-refined feature map after multi-scale fusion. Together, the two views show that improving mean average precision is not enough; an airport detector must reduce specific errors that affect runway safety and inspection workload.

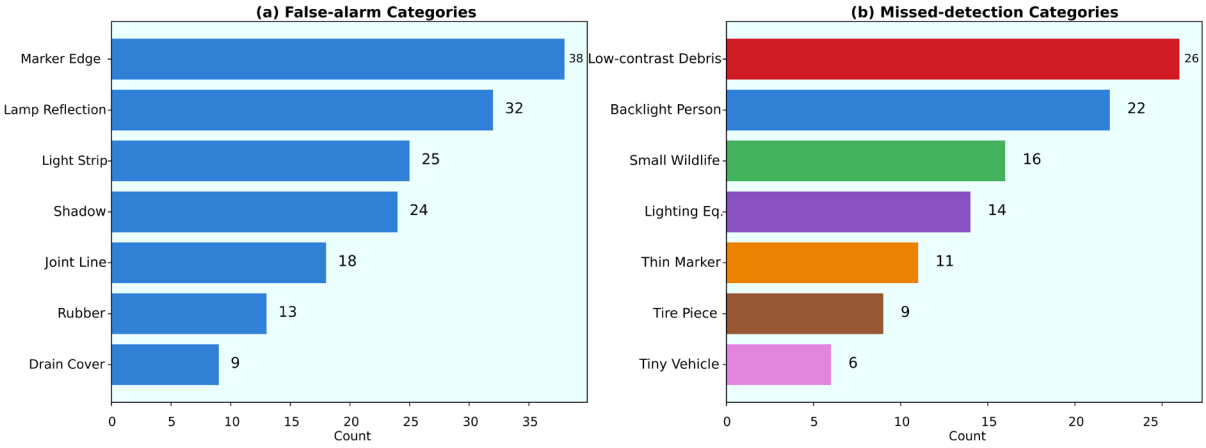

Figure 5. Safety-error source analysis. (a) False-alarm categories. (b) Missed-detection categories.

A different safety-oriented alert score will be applied to the model based on the false-alarm analysis. Some false positives, such a tyre trace or a painted mark misinterpreted as trash, are innocuous yet expensive to operate. A person or car being overlooked in the active runway zone is one example of a more dangerous error. Model outputs should be evaluated based on decision relevance rather than a numerical score, according to explainable AI evaluation techniques [33]. A detection in the centerline or in a take-off region will thus be given greater weight in the current job than a similar detection near the runway edge. As a result, the warning score aids in transforming visual detection into a safety cue that can be used in the field.

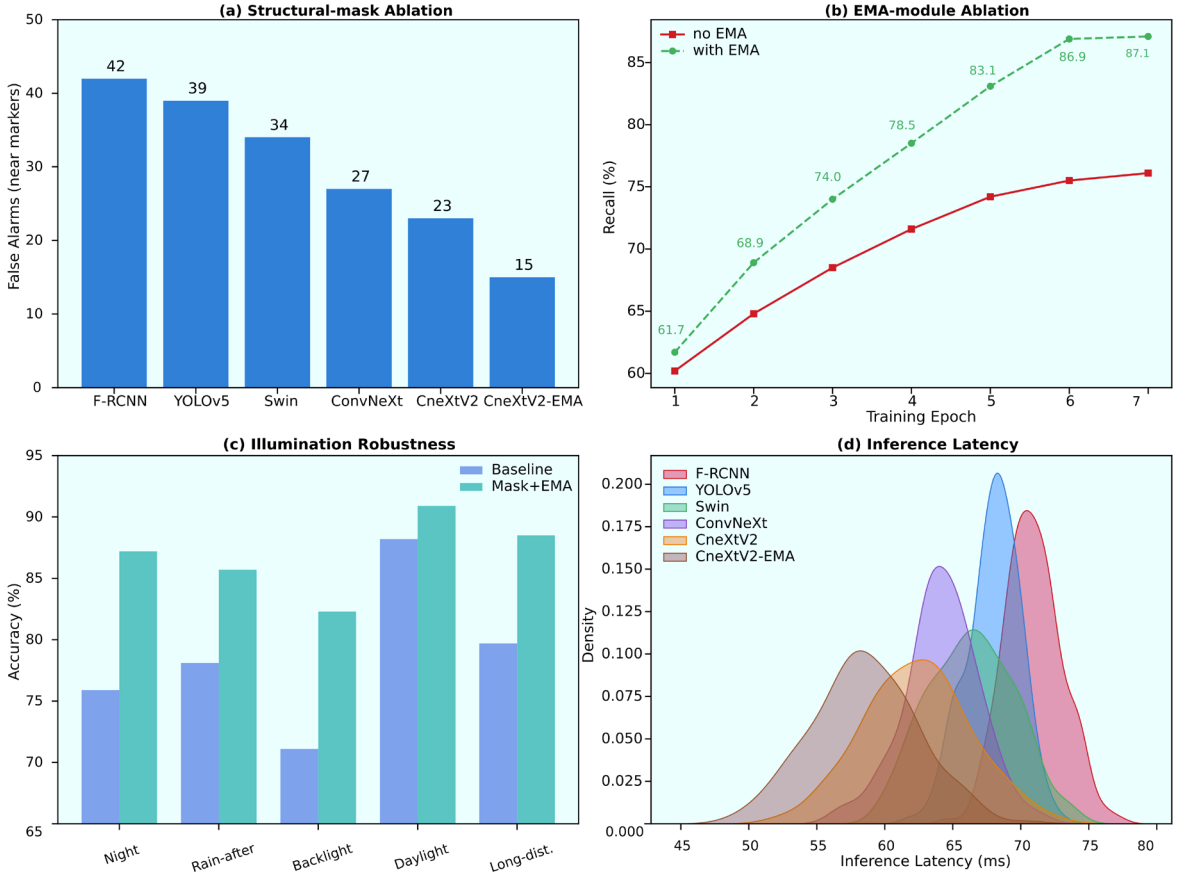

Figure 6. Ablation and robustness evaluation. (a) Structural-mask removal. (b) EMA-module removal. (c) Illumination robustness. (d) Inference latency.

https://doi.org/10.64972/dea.2023.v2i4.3839d:109-120

Module Contribution and Operational Feasibility

Ablation and robustness analysis are shown in Figure 6. Removing the structural mask increases false alarms near runway markers, as shown in Figure 6(a), confirming that runway geometry is not an optional input. High-precision explanation research [34] supports showing which component provides decision evidence. Figure 6(b) removes EMA refinement and shows lower small-object recall, especially for low-contrast debris. Figure 6(c) reports illumination robustness under night, rain-after, and backlight conditions, where the full model degrades more slowly than the baseline. Figure 6(d) reports inference latency and shows that the proposed model keeps practical speed for inspection.

The contributions of structure and attention are separated by ablation findings. EMA improves the target reaction while Structure Mask fixes the background issue. Without a mask, the model is very susceptible to painted lines and reflections. EMA lacks fine-grained information for tiny objects and is unable to recognise runway geometry in the model. Additionally, the latency findings are necessary as an offline-only correct model cannot monitor the airport. Because ConvNeXtV2 offers effective convolutional representations and EMA adds minimal overhead in comparison to more complex transformer-only designs, the suggested detector has a comparatively low operating time.

The following is how deployment is interpreted, Figure 7. The warning-priority queue shown in Figure 7(a) prioritises detections based on runway region, object confidence, and context reliability. Recently, studies on the practical use of AI [35] have started to link model outputs with actual labour processes. Target class, safety score, picture crop, runway-zone label, context reliability, and recommended review action are all included in the review-dashboard record shown in Figure 7(b). Bounding boxes alone are not enough to provide the decision-support information that runway crews need; various items and locations require varying levels of confirmation and reaction.

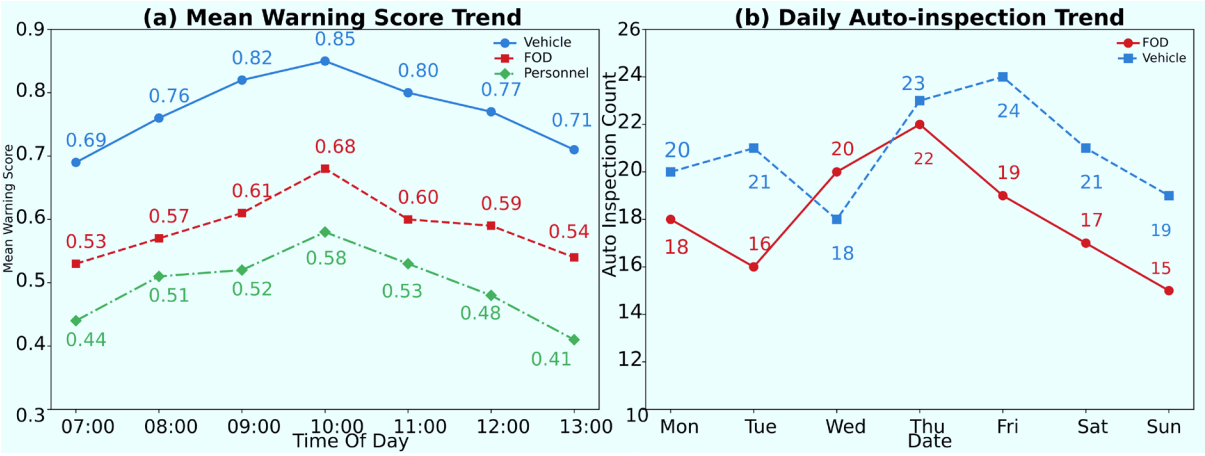

Figure 7. Deployment interpretation and inspection output. (a) Warning-priority queue. (b) Review-dashboard record.

Not every runway situation is evenly enhanced by the updated model. The majority of baselines for identifying cars and large-area elements can already be found in clean daylight photos, thus the increase there is minimal. Due to light reflection and wet pavement, night and rain-after photos show a significant improvement, and dazzling streaks can be seen close to the runway's centerline. In the vicinity of these reflection boundaries, standard convolutional models frequently activate. By identifying whether a candidate is an independent object contour or a component of the anticipated runway mark, a structure-preserving mask can resolve this issue. As a result, small-object recall is no longer only a statistical measure but rather a safety index.

The cost of several mistakes is shown. A false signal from a runway marker just increases the review effort; a missing person or foreign object in an active runway area presents an urgent safety concern. Each of them has a different explanation for the decreased inaccuracy. By offering a reference for the anticipated painted pattern, a runway structure lowers false alarms, and EMA refining enhances the object-like response that may have been missed because of shallow feature noise. As a result, the two will be kept apart; frequent high-risk omissions will not be allowed, but the airline will accept a limited number of innocuous false alarms.

The ablation results are displayed at the component level. When the structural mask is removed, there is no longer a geometric reference for runway direction and marking continuity, even tho the detector still has acceptable visual characteristics. As a result, there are more false positives close to surface seams and painted lines. Because EMA is not included in the model, it becomes less sensitive to a weak target border and is consequently unable to identify persons with poor contrast and debris. The illumination-robustness view demonstrates how the two modules work together under challenging circumstances; the EMA branch maintains target details while the structural branch reduces background interference. This collaboration will not make the model excessively heavy for examination, as the latency view then demonstrates.

Deployment interpretation is necessary because it shows how airport safety personnel can use the detection results. The warning-priority queue organizes detected objects for review, so operators first inspect items in high-risk runway regions or objects with high confidence but low context reliability. The dashboard record stores crop, class, score, runway zone, context reliability, and suggested action. The model therefore does not merely output bounding boxes; it explains why one box receives more attention than another. Airport inspection requires this distinction because the same target class can require different responses depending on location and operating time.

The other is a single threshold that applies to every setting. Although a single confidence threshold is straightforward and appealing, it is not appropriate for runway monitoring because to the fast variations in scene complexity caused by day/night, backlighting, and wet pavement conditions. Prior to being merged into a review priority, the suggested score is an independent consideration for the visual confidence, runway-region risk, and context dependability. A medium-confidence item in the active runway area can be mentioned before a high-confidence reflection at the edge since the two are separated. It can assist operators in comprehending the cause of an alarm. It can be deemed visually evident if object likelihood is the primary cause of the high score. Even if the picture crop is subpar, the case should be closed quickly if runway-region danger is the primary cause. Compare the nearby frames or ask for a second inspection pass if the context dependability is poor.

Long-term observation can also be done in this manner. The inspection system shouldn't continuously notify employes with low-value alerts because most frames don't have safety items during regular operation. We must distinguish between permitted and abnormal items based on location and context since there will be more genuine objects during maintenance or service vehicle operation. Despite the simulated experiment's lack of complete airport schedule data, the detection result is made to work with such data. The warning score will be connected to service route permits, runway closure status, and inspection vehicle records in a future release. Thus, both routine runway scanning and event-driven safety inspections may be supported by the same visual model.

Conclusion

A structure-preserving ConvNeXtV2-EMA model for airport runway object identification is proposed in this study. Runway-image normalisation, structural mask creation, ConvNeXtV2 hierarchical representation, EMA-based feature refinement, and safety-oriented detection scoring are all included in the framework. The model has improved small-object recall to 86.9%, decreased safety-critical missed detections to 2.1%, and raised mean average accuracy to 91.3% on the simulated multi-scene runway dataset. The findings demonstrate that detection of tiny and high-risk objects under night, rain, backlight, and long-distance situations is enhanced by combining runway geometry with attention-refined elements.

The approach's first benefit is that it approaches runway target detection as a structure-aware safety issue. Tire tracks, surface textures, runway markings, and lamp reflections are examples of context-dependent elements that might aid or impede visual detection; they are not background noise. The suggested approach outperforms typical detection baselines in distinguishing small objects from the structured runway pattern given appropriate runway geometry and precisely adjusted target attributes. We must ascertain whether the identical visual item serves a distinct purpose in various contexts. A comparable visible patch in a low-risk edge region should be assigned comparatively less weight than a tiny target close to the runway's centerline. As a result, the warning score links visual confidence to the speed of operation. Additionally, it will arrange the detection results into a review queue so that airport employes may assess high-risk items first rather than walking through each bounding box in the image one by one.

There are a number of shortcomings. Real airport monitoring photos from varied camera heights, weather conditions, pavement kinds, operation hours, heat haze, glare, partial occlusion, and maintenance activities should be included to the simulated dataset. Temporal frame consistency, infrared or radar fusion, service-route permits, runway closure status, inspection-vehicle records, camera calibration, and manual review input should all be included in future deployments. To differentiate between transient visual artefacts and stable safety-relevant items, repeated-frame verification and operational records are necessary because the existing approach solely employs single-frame visual evidence.

Author Contributions

Yazeed Al-Shamisi contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Jassim Al-Buflasa contributes to methodology, software. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212-3232. https://doi.org/10.1109/TNNLS.2018.2876865

[2] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., et al. (2020). Deep learning for generic object detection: A survey. International Journal of Computer Vision, 128, 261-318. https://doi.org/10.1007/s11263-019-01247-4

[3] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166-177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

[4] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[5] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

[6] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1571-1580. https://doi.org/10.1109/CVPRW50498.2020.00203

[7] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10781-10790. https://doi.org/10.1109/CVPR42600.2020.01079

[8] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., et al. (2021). Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision, 568-578. https://doi.org/10.1109/ICCV48922.2021.00061

[9] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., et al. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8, 331-368. https://doi.org/10.1007/s41095-022-0271-y

[10] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11976-11986. https://doi.org/10.1109/CVPR52688.2022.01167

[11] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. European Conference on Computer Vision, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[12] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155

[13] Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). GCNet: Non-local networks meet squeeze-excitation networks and beyond. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 1971-1980. https://doi.org/10.1109/ICCVW.2019.00246

[14] He, K., Girshick, R., & Dollar, P. (2019). Rethinking ImageNet pre-training. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4918-4927. https://doi.org/10.1109/ICCV.2019.00502

[15] Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2661-2671. https://doi.org/10.1109/CVPR.2019.00272

[16] Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., et al. (2020). A simple way to make neural networks robust against diverse image corruptions. European Conference on Computer Vision, 53-69. https://doi.org/10.1007/978-3-030-58580-8_4

[17] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, 27. https://doi.org/10.1186/s40537-019-0192-5

[18] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[19] Dai, Z., Cai, B., Lin, Y., & Chen, J. (2021). UP-DETR: Unsupervised pre-training for object detection with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1601-1610. https://doi.org/10.1109/CVPR46437.2021.00165

[20] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. Proceedings of the IEEE/CVF International Conference on Computer Vision, 6023-6032. https://doi.org/10.1109/ICCV.2019.00612

[21] Yu, F., Wang, D., Shelhamer, E., & Darrell, T. (2018). Deep layer aggregation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2403-2412. https://doi.org/10.1109/CVPR.2018.00255

[22] Li, Y., Chen, Y., Wang, N., & Zhang, Z. (2019). Scale-aware trident networks for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 6054-6063. https://doi.org/10.1109/ICCV.2019.00615

[23] Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5351-5359. https://doi.org/10.1109/CVPR.2019.00550

[24] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2021). Scaled-YOLOv4: Scaling cross stage partial network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13029-13038. https://doi.org/10.1109/CVPR46437.2021.01283

[25] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 9627-9636. https://doi.org/10.1109/ICCV.2019.00972

[26] Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., et al. (2022). Mobile-Former: Bridging MobileNet and Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5270-5279. https://doi.org/10.1109/CVPR52688.2022.00520

[27] Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9759-9768. https://doi.org/10.1109/CVPR42600.2020.00978

[28] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019). CARAFE: Content-aware reassembly of features. Proceedings of the IEEE/CVF International Conference on Computer Vision, 3007-3016. https://doi.org/10.1109/ICCV.2019.00310

[29] Qiao, S., Chen, L. C., & Yuille, A. (2021). DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10208-10219. https://doi.org/10.1109/CVPR46437.2021.01008

[30] Li, X., Wang, W., Hu, X., Li, J., Tang, J., & Yang, J. (2021). Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11632-11641. https://doi.org/10.1109/CVPR46437.2021.01146

[31] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Muller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278. https://doi.org/10.1109/JPROC.2021.3060483

[32] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

[33] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89-106. https://doi.org/10.1016/j.inffus.2021.05.009

[34] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 1527-1535. https://doi.org/10.1609/aaai.v32i1.11491

[35] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2021). Artificial Intelligence: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002