DOI: 10.52756/ijeim.2025.v01.i02.005 Int. J. Eng. Infor. Manag. 1(2): 54-62 (2025)

International Journal of Engineering and Information Management
Journal homepage: www.ijeim.in

Robust SCM Packaging Authentication through Aggregation-Based CSP Darknet53 with Integrated Spatial Pyramid Pooling

Brijit Bhattacharjee[1*], **Subhajeet Brojabasi**[2], **Aparna Das**[1]

[1]Department of Computer Science and Engineering, Swami Vivekananda Institute of Science & Technology, Kolkata-700145, West Bengal, India
[2]Department of Computer Science & Engineering Cyber Security & Data Science, Brainware University, Kolkata - 700125, West Bengal, India

***Corresponding Author:** brijit002@gmail.com

Article Information

Type of Article: Original
Received: Dec 13, 2024
Accepted: Apr 6, 2025
Published: Apr 15, 2025

Keywords:
Darknet
DNN
Object Detection
Packaging Authentication
Spatial Pyramid Pooling.

Cite this article:
Brijit Bhattacharjee, Subhajeet Brojabasi, & Aparna Das (2025). Robust SCM Packaging Authentication through Aggregation-Based CSP Darknet53 with Integrated Spatial Pyramid Pooling. *International Journal of Engineering and Information Management*, 1(2), 54-62.
DOI: 10.52756/ijeim.2025.v01.i02.005

Abstract

The Supply chain business model is being reshaped and transformed by artificial intelligence (AI) using contemporary, environmentally friendly AI-based techniques. Customers frequently receive the incorrect items in parcel boxes when they order online. Supply Chain (SC) companies uphold a simple return policy for clients, which incurs additional expenses for a specific order, to preserve their reputation. A certain amount of time is spent on the return procedure, which can result in losses for significant returns. By incorporating deep learning algorithms into the SC packaging process, the suggested method attempts to address this issue. Before the things are delivered, the darknet architecture in this document detects incorrect product delivery. Semantic properties are extracted from the video footage using the following framework. The suggested method is made with CSPDarknet53 to enhance learning capacity. The spatial pyramid pooling candidate window preserves 6×6 features as a connection parameter while providing special information processing to the various CNN layers. The DPM model is used for the encoding process. In the suggested information construction, a selective, optimistic search pattern path augmentation localized by RELU activation is used to choose features.

1 Introduction

Automation, along with digitization, became an important aspect of Supply Chain Activity (SCA) influenced by Deep learning algorithms (Dolgui & Ivanov, 2021; Mahrishi et al., 2021). A fuzzy-based inference system designed by Özkan & İnal (2014), which is referred to as ANFIS, is capable of solving decision-making problems related to supplier selection (refer to Figure 1). The SCOR model proposed by Lima-Junior & Carpinetti (2019) predicts SC performance using a perceptron neural network. Carrera et al. (2020) proposed an automatic model for supplier selection based on performance, characteristics, and management capabilities. Thompson (1990) proposed an intuition-based vendor profile analysis for multi-attribute qualitative analysis. AHP/ANP are AI-based commercial tools to resolve SCA's inventory-based problems. Formulation of Operational and production data in automated digitalization improves the overall performance of SC activities (Dubey et al., 2019; Simonyan & Zisserman, 2014; He et al., 2015b; Ioffe & Szegedy, 2015; Howard et al., 2017; Jia et al., 2014; Tieleman & Hinton, 2012).

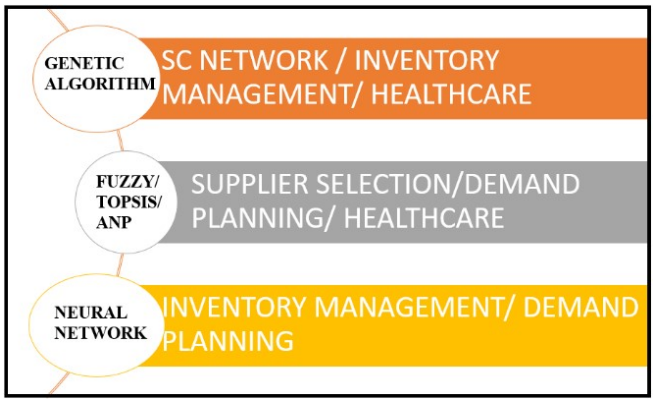

Figure 1: AI-based research tools and Algorithms to resolve Supply Chain problems.

The multi-criterion decision-making problem (MCDM) explores various learning capabilities of neural networks. The criteria for selection and evaluation distribution concerning input and output data automatically using Adaptive Neuro-Fuzzy Inference System (ANFIS) enabling improved accuracy and efficiency in data-driven modeling. ANFIS is a neuro-fuzzy system for function approximation in clustering and pattern recognition. ANFIS integrates least squares with gradient descent using a hybrid learning rule. ANFIS applies least squares to reduce the impact of measurement error while updating premise parameters using gradient descent. The ANFIS maintains a layered approach where each layer passes through two pie layers and repeatedly updates its weights. Multiple layers with multiple weights finally produce a sum for all the evaluated weights from the pile and N nodes. The ANFIS model was tested on supplier data for quality, delivery, and technology evaluation. Decision-makers determine the performance criteria of the suppliers. Based on the given input, a rank parameter is proposed according to the level of the suppliers. The ultimate performance is achieved with such input applied to the training phase. The experiment was carried out with 1000 epochs, 0.7733 residual with 0.002 for training and 0.01 MSE score for testing. During Gol-Mohammadi's experiment, the ANFIS model scored 2.67 × 10-17 MSE for training. MSE for overall stage operation is 0.0134. The final observation for this experiment concludes the decrease in the supplier score based on quality, delivery, and price. Cox (1999) proposed a machine learning-based contamination detection approach for food recognition and quality, calorie prediction. Wang et al. (2019) implements evolutionary computation techniques with transfer learning for seafood and wine manufacturing business organizations. Integration of fault detection with autonomous computing proposed a unique approach Wang et al. (2018) for smart manufacturing. Hydropower pinch analysis (HYPOPA) proposed an optimization approach for multipurpose reservoirs, focusing on the effect of climate change on the supply chain management system (Goodfellow et al., 2013). Applications are focusing on demand forecasting, where the BULLWHIP approach solves demands according to customer requirements. This approach efficiently optimizes inventory with limited cost and increased customer loyalty (Huang et al., 2017) and proposes a performance forecasting approach for the high-cost logistics industry. Zhang et al. (2010) integrated IoT and blockchain to create a supply chain provenance system to forecast demand for food. He et al. (2015b) proposed a multimodal classification to estimate the maturity of food through feature concatenation of hyperspectral images. This approach reduces food wastage along with food production costs. A similar type of problem is also solved by the Inertial Measurement Unit (IMU). The approach proposed by Yue et al. (2016) focuses on information security from data manipulation to debit/credit card fraud identification.

2 Proposed Approach

In Figure 2, the solution is explained with diagrams during the packing process, and the items are checked with the camera. The camera, applying our proposed approach, detects the object inside the box and sends information to the system. The system also maintains a specific record of the contents inside the box with a specified ID. If both items are matched, then packing continues; otherwise, it will tag the

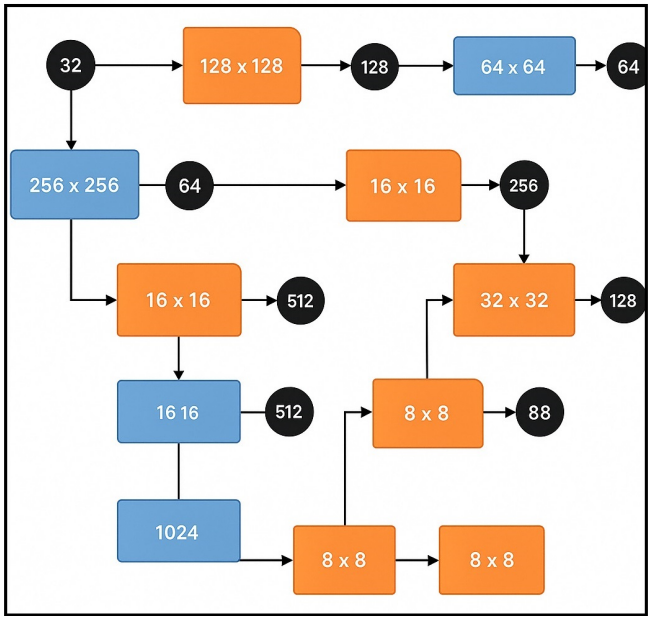

Figure 2: Framework of the Proposed Approach.

Table 1: Average Precision Evaluation Observations

COCO AP_{50}	FPS	Hardware
50%	109	GTX 1080 Ti
40%	52	Intel Core i9-9900K
42%	49	Nvidia Jetson TX2

box as status unmatched. Then, those unmatched ID boxes are packed separately. The deep learning approach is in this proposed approach via backbone architecture, pooling layer, and aggregation network.

2.1 Backbone Architecture

The proposed architecture consists of CSPDarknet53 (Wang et al., 2019). CSPDarknet53 efficiently increases the learning ability of CNN by reducing the computation effort by 10 to 20 percent. CSPNet distributes the computation to several layers of CNN to reduce the energy waste caused by unnecessary computation. The following approach reduces the impediment of computation for PeleeNet (Wang et al., 2018) by up to 50% (Wang et al., 2019). Concerning memory cost, Dynamic Random Access Memory (DRAM), immediately, large space CSPNet replaced water fabrication with cross-channel pooling (Goodfellow et al., 2013) to cut down memory usage by up to 75% (Wang et al., 2019). The average precision on a few tested observations is mentioned in Table 1.

The CSPDarknet53 (Wang et al., 2019) maintains a dense block and transition layer for each stage (Huang et al., 2017). The dense input and output are shown via Equation (1) given below.

$$m_1 = s_1 \times m_0 \tag{1}$$

$$m_2 = s_2 \times [m_0, m_1]..., m_j \tag{2}$$

$$s_j \times [m_0, m_{j-1}] \forall m \in \text{input} S \in \text{weight} \tag{3}$$

The weight updated with gradient propagates P represented as,

$$m_1^{'} = \phi(s_1, P_0), m_1^{'} = \phi(s_2, P_0, P_1), ..., m_n^{'} = \phi(s_j, P_0, P_1, ..., P_{j-1}) \tag{4}$$

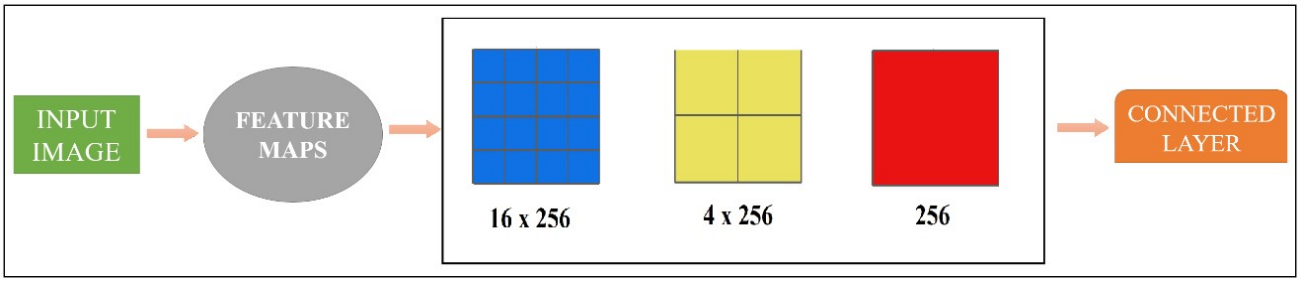

Figure 3: Spatial Pyramid Pooling Layer Network Structure.

The CSPDarknet53 (Wang et al., 2019) composed with partial dense block reevaluated as Equations (5-8):

$$m_J = s_j * [m_0, m_1, ..., m_{j-1}], m_T = s_T * [m_0, m_1, ..., m_j], m_U = s_U * [m_0, m_T] \tag{5}$$

$$s_j^{'} = \phi(s_j, P_0, P_1, ..., P_{j-1}) \tag{6}$$

$$s_T^{'} = \phi(s_T, P_0, P_1, ..., P_j) \tag{7}$$

$$s_U^{'} = \phi(s_U, P_0, P_T) \exists m_T, s_T^{'} \tag{8}$$

2.2 The Spatial Pyramid Pooling Layer

The arbitrary input fed to the convolution layer generates a variable-sized output. Fixed-length vectors are used to perform pooling operations applying the Bag-of-words (BoW) approach (Zhang et al., 2010). Through spatial information in local spatial bins, spatial pyramid pooling (He et al., 2015a; Yue et al., 2016) improves the performance of the Bow approach. For an image with an arbitrary size, a deep network is adopted with a spatial pyramid layer for pooling operation. In Figure 3, SPP uses maxpooling for each spatial bin. The generated output is a 256-dimensional vector (Rota Bulò et al., 2017). The global average pooling improves accuracy by reducing overfitting through the words method. The following network structure was trained by back-propagation with a fixed number of inputs. SPP architecture also influences ZF-5 (Zeiler et al., 2011) with five CNN layers. Convnet 5 (Howard, 2013) was modified with the two pooling layers applied in feature maps. Overfeat 5/7 modified with SPP by a larger feature map and a larger filter number, 512. In SPP-NET, a Candidate window is extracted as a selective, optimistic search formulation. The CNN applies a large number of windows for each image, which will consume much greater time for feature extraction. The Binary support vector machine also maintains a similar approach with the least amount of time. A deformable part model (DPM) extracts features with encoded feature maps and requires a predefined window size. Deep Pyramid DPM is popular for complementary approaches like sliding windows and region-based detection. DPM evaluates the global maximum score function by a dynamic programming algorithm. All DPM parts generalize multi-resolution models via a sub-sampling layer. DPM reveals an object geometry filter by implementing an implicit convolution layer. DPM – CNN vectorized filter confined within a subarray of the route. To maintain maximum nonlinearity, vector elements are added with scalar bias. Sampling layers and fixed depth approximation measure loopy structures within a detection window. End-to-end training is carried out within the HOG-based DPM during end-to-end optimization. Fixed-depth network explodes with an unrolling Inference algorithm converging to a particular fixed point. DPM also helps construct a feature pyramid by eliminating the max pooling layer. Specific visual structures of Max pooling are needed to prevent subsampling. Stand and hard negative mining can overlap negative samples with a 30-42% intersection over the union ratio. The candidate window uses spatial pyramid information of 1×1, 2×2, 3×3, 6×6 pool features.

2.3 Path Aggregation Network

The following framework localizes instances on the edges of the low-level pattern to enhance localization capability. Bottom-up path augmentation is carried out by producing similar spatial measure feature

maps. Applying the basic structure of ResNet (He et al., 2015a), PAN generates a new feature map from a lateral connection using the Coarser map. The (3×3) convolution layer processes the fused feature map followed by ReLu (Krizhevsky et al., 2012). In dark net 53, the prediction takes place by Coordinate C_p, C_q, C_r, and C_s. The corners are represented via (mp, mq) with a foundation for width and height as f_w, f_n. The prediction is carried out with the sigmoid functions corresponding to

$$N_p = S(CP) + mp, Nq = 5(C_q) + mq, Nwfwecw\&N_h = fnecn \tag{9}$$

The sigmoid function is capable of mapping number lines into a limited range by converting a scalar value into a probabilistic measure. In deep learning approaches, such activation is inspired by biological neural networks. In general, the sigmoid function is Capable of converting mood y output into a probability score. Sig Sigmund function converts a scalar value into zero and one for a range of positive and negative sigmoid functions, which requires hyperbolic tangent formation in metric. Application of the sigmoid referred to as the arctangent function formula. In logistic regression, the sigmoid function helps to diagnose medical problems via stochastic measurements, such as the activation function, which introduces nonlinearity to neural networks. The system ignores predictions by assigning ground truth for each object. The $N_p, N_q, N_w,$, and N_h are used to predict the classes using multi-label classification. The S function formulates a complex domain by overlapping labels to achieve better accuracy. Scaling and feature extraction were carried out with the Pyramid networks.

3 Experiments and Results

Figure 4 illustrates an evaluation of multiple object detection using precision and recall values. A sample of 10 images for each topic was tested on this model. Higher mean and lower standard deviation represent better accuracy for detecting that particular object. VGG (Simonyan & Zisserman, 2014) framework formed with a stack of CNN layers, has 4096 channels for 2 and 2000 in the last. The ConvNet configuration is distributed among weight layers like A with 11, B A-LRN with 11, B with 13, C with 16, D with 16, and E with 16 layers. Network parameter ranges from 133 to 144.

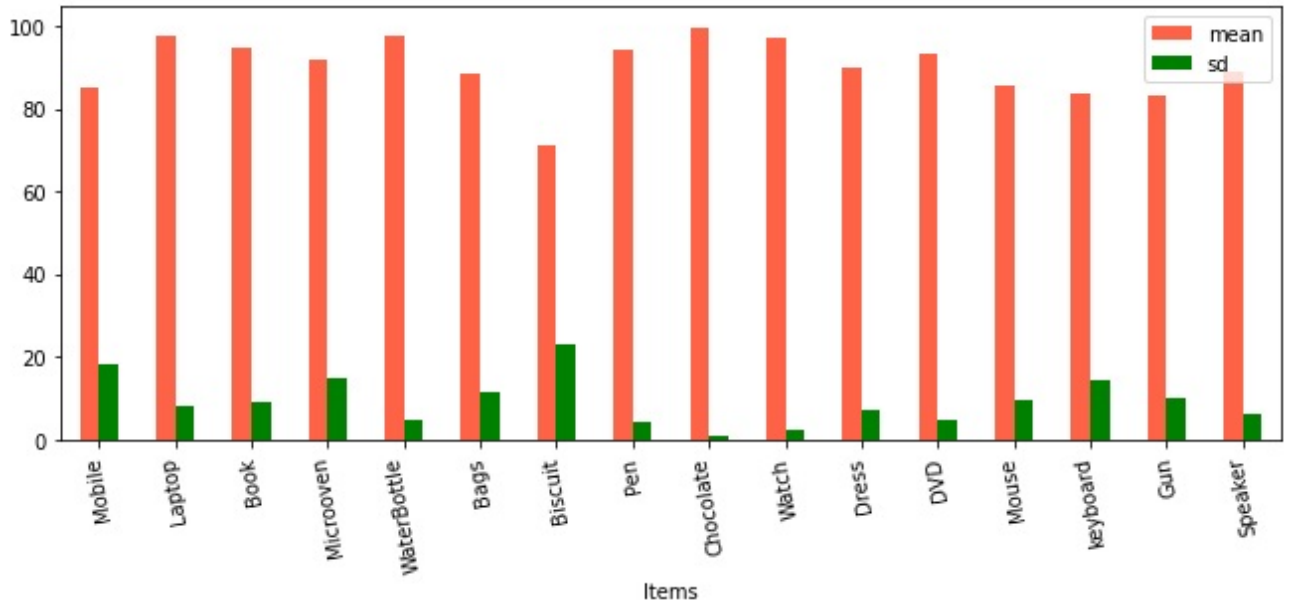

Figure 4: Precision and Recall evaluation based on multiple objects with mean and standard deviation.

He et al. (2015a) inherited the concept of ResNet which was introduced by Simonyan & Zisserman (2014) to prepare a 3×3 filter with an equal feature map size. Applying scale Augmentation (Simonyan & Zisserman, 2014), the image is sampled into a range between 256 and 480 pixels. Batch normalization (Ioffe & Szegedy, 2015) is carried out before activation. Thirty-four layer ResNet exhibits low training

Table 2: Comparison of the proposed model with other related Deep learning models

Model	Method	Activation	Pooling	Training
VGG (Simonyan & Zisserman, 2014)	Train the given input using ConvNet with 224*224 RGB form. The following network is equipped with Local Response Normalisation.	RELU	Maxpooling over 2*2 Pixel	Training Procedure carried out with batch size 256 and momentum to 0.9.
ResNet (He et al., 2015a)	Efficient and popular.	RELU	Average and Softmax	Equivalent to VGG (Simonyan & Zisserman, 2014)
Mobilenet (Howard et al., 2017)	Stride with conv/s2 with filter shape 3*3*3*32. Continued with FC/SI at 1024*1000	RELU	Softmax	The training Process was carried out using RMSPROP (Tieleman & Hinton, 2012) by asynchronous gradient descent.
Proposed Approach	Applying CSPDarknet53 backbone architecture with spatial pyramid pooling layer and path aggregation network as the neck.	RELU 6	Softmax	It can be updated using advanced backbone architecture.

error and generalized validation. MobileNet (Howard et al., 2017) factorizes a standard CNN into a depthwise CNN reformed as a pointwise convolution. With twenty-eight layers, spatial resolution is reduced by average pooling. In Howard et al. (2017), dense convolution is optimized by the GFMM function. Re-ordering of memory is implemented via a cafe package (Jia et al., 2014). The MobileNet started with conv/s2 with filter shape 3×3×3×32 and input size 224×224×3. Then 3×3×32 dw with conv dw/s1 stride with input size 112×112×32. Then, a series of conv/s1 and conv dw/s1, the filter shape becomes 1×1×256×512 with input size 14×14×256.The conv dw/s2 with filter shape 3×3×512 dw and input size 14×14×512. Finally, conv/s1 is implemented with filter shape 1×1×1024×1024 and input size 7×7×1024. The avgPool/s1 is implemented with a filter shape pool of 7×7 and an input size of 7×7×1024. The classification was carried out with SoftMax with a final size of 1×1×1000.

4 Comparative Studies

Retina Net is used to compute computational feature maps and perform object classification using bounding box regression. The goal of this approach is to localize the coordinate for object identification. The CNN performs a bounding box upon the labelled dataset with corresponding coordinates from the region of interest. The training process is carried out by calculating the difference between predicted and ground truth-bounding boxes. Retina-Net resolves the training data imbalance using focal loss. The architecture of Retina-Net is influenced by the feature Pyramid Network (FPN) capture feature with different resolutions at a specific scale. The modulation factor provides a robust solution for object recognition. The experiments with robust training range from 0.5 to 5 with focal loss 2. Each sample image with 100k anchors uses heuristic sampling for each mini-batch. During the experiment Retina net sample, the predicted probability for 107 negative and 105 positive windows. Nonmaximum suppression (nms) applied to online hard example mining (OHEM) enforces 1:3 positive negative variants during each minibatch with a large class in balance. Two-stage detectors classify boxes using region pooling operation. The baseline trained owing OHEM achieves 36.0AP with a gap of 3.2AP for dense detector training. Another backbone architecture is Deconvolutional Single Shot Detection (DSSD), which uses a progressively smaller SSD convolution layer. The prediction was carried out with a 3×3-Dimensional filter with non-maximum suppression NMS to eliminate redundant bounding box predictions to improve the efficiency of object detection.

The DSSD training process is carried out by matching the default box with each ground truth box with a threshold of 0.5. The Confidence loss maintains a positive-negative ratio of 3:7. Data augmentation is carried out with random flipping and photometric distortion. To detect small objects, DSSD uses a random expansion augmentation trick. The experiment was carried out with SSD321, which achieved a map score of 76.4. There are three variations found where SSD321 combined with $Pm(b), Pm(c)$, and $Pm(d)$ secure map between 76.9 to 77.0. The experiment achieved a 78.6 map by integrating Dm (Eltw - Prod) with $Pm(c)$ and SSD321. In comparison with residual–101, the improvement is shown in detecting large objects. The DSSD519 – Residual 101 secures a 3.3% better map than R-FCN (Özkan & İnal, 2014). The Following approach incorporates Top-down Modulation (TOM) to achieve a lower-level feature map from top-down contextual features. The lateral connection approach feeds the bottom–up feature as input to each layer of TDM. TDM network is distributed into the lateral module (L) and the top-down module (T). Lateral features are produced via L, while T is used to combine these lateral features with top-down features. The training problem is influenced by L while updating gradients from the object

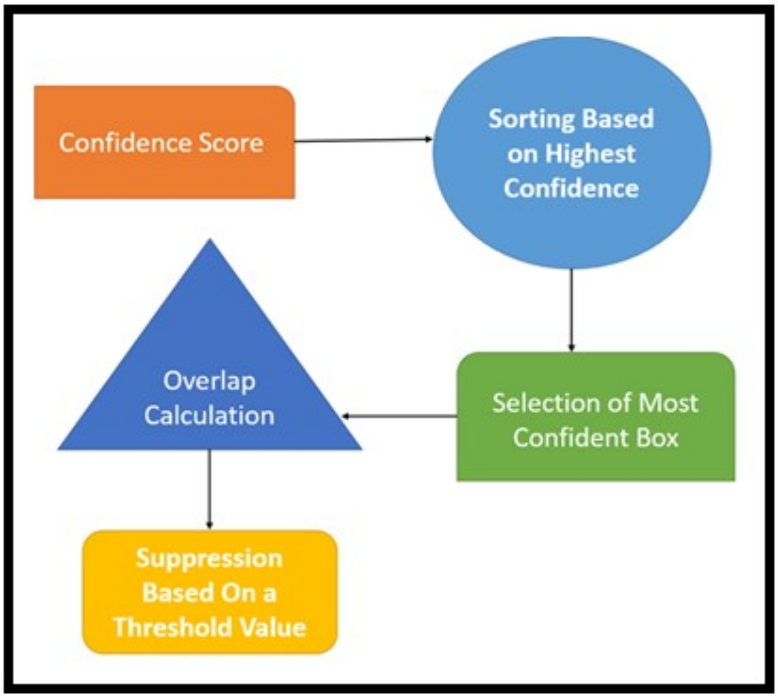

Figure 5: Bounding box and Confidence box formation.

detection. To preserve P-Down feature transmission, L learns to transform low–level features and T learns to transform semantic information. L is a ReLU nonlinear 3×3 convolution layer. The resolution of each L must be higher than its previous one. TDM conducts extensive experimental evaluations to ensure substantial performance. TDL added each layer to the TDL network while performing training on the decision task. During each iterative experiment, TDM combined with 4 TDM top-down modules scored AP 26.2, AP 50, 45.7, and AP75 27.2. TDM with VGG16 secured 29.9 AP and 50.9 AP50 with two top-down modules. The baseline with five features has AP32.1, AP50 59.2, and AP75 39.8. TDM with ResNet To1 has AP94.4, AP 54.4, and AP75 39.8 for three top-down modules. TDM with IRNV2 gets AP 38.1, AP50, 58.6, and AP 75, 40.7. TDM networks are efficient at transmitting contextual features with finer detail. The scale-up is carried out with the R01-Pool feature from L2 – normalize layers. Yue et al. (2016) proposes a unified implementation of faster R-CNN, RFCN and SSD. The box proposal was generated using the combined approach of R-CNN and F-R-CNN. In this approach, images with different spatial locations are divided into variable aspect ratios named anchors. The loss was evaluated with the image and model parameters. At different scales, tiling a collection of boxes generates a regular grid of anchors referred to as predictors. Region-based Fully Convolution Network (R-F(N) Applies a positive-sensitive cropping mechanism that is more efficient in comparison to ROT Pooling operation. The feature extraction is carried out using VGG-16, ResNet-101, and Inception v2. The Number of Proposals is set to 300 to reduce the risk of recall. The approach uses stride to improve mAP by a factor of 5%, automatically increasing the running time by 69%. To achieve regression, target ground truth instances must match with each anchor. The argmax matching algorithm is used to design positive and negative anchors, and the matching anchor maintains a box encoding function with 510g weight and height Proposition. Input size configuration Carried out with $M \times M$ fix shape. Downscaled formation. The post-process detection maintains a 0.6 Iou threshold of the metric $[0.5 : 0.05 : 0.95]$. The non-max suppression runs with 40 ms fastest models. The vector for average Precision evaluates AP vectors using horizontal flipping and box refinement visual attractiveness using rigorous criteria on large objects. (Howard et al., 2013) uses a Coarser-resolution feature map to accurately localize lower-level semantics in spatial resolution. The single-scale baseline approaches are unable to decide a large number of anchors to improve accuracy for Scale variance. The approach maintains a mini-batch to benefit lighter-weight head formulation. Applying this segmentation proposal. The output size increased from 14×14 to 28×28. It secures 3.4 AP during small-scale object detection. The feature pyramids act Like

a Deep mask that can be masked with 34, 64 and forwarded to 128, 512 pixels with 7×7 larger MLP. Canonical objects maintain a Padding of 25% with $\sqrt{2}$ larger corresponding sizes.

5 Conclusion

The research develops an effective SCM packaging authentication system through its integration of CSPDarknet53 architecture and Spatial Pyramid Pooling while implementing an aggregation-based approach. The proposed method resolves the frequent problem of inaccurate deliveries in e-commerce logistics through complete packing box content recognition prior to shipment. The predictive system delivered reliable performance according to experimental results that showed mean precision reaching between 80 and 93 and standard deviations varying between 14 and 4. The supply chain obtains advanced capabilities from the fusion of CSPDarknet53 with selective semantic feature extraction and ReLU-based localized path augmentation, which are deep learning methods. Spatial pyramid pooling protects fundamental spatial frameworks within different layers, so the framework provides detailed object verification in complex packaging scenes. The solution achieves enhanced sustainability through its combination of CSPDarknet531 with selective semantic feature extraction, both improved by ReLU-based localized path augmentation. These intelligent systems in SCM operations lead to an industrial shift through operational improvements and better environmental accountability.

Conflict of Interest

The author declares that there is no conflict of interest in this work.

Data Availability

Data may be available on request.

References

Carrera, D. A., Mayorga, R. V., & Peng, W. (2020). A soft computing approach for group decision making: A supply chain management application. *Applied Soft Computing*, *91*, 106201. https://doi.org/10.1016/j.asoc.2020.106201.

Cox, A. (1999). Power, value and supply chain management. *Supply Chain Management an International Journal*, *4*(4), 167–175. https://doi.org/10.1108/13598549910284480.

Dolgui, A. & Ivanov, D. (2021). 5g in digital supply chain and operations management: fostering flexibility, end-to-end connectivity and real-time visibility through internet-of-everything. *International Journal of Production Research*, *60*(2), 442–451. https://doi.org/10.1080/00207543.2021.2002969.

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., & Hazen, B. T. (2019). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, *226*, 107599. https://doi.org/10.1016/j.ijpe.2019.107599.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. arXiv preprint arXiv:1302.4389. https://doi.org/10.48550/arxiv.1302.4389.

He, K., Zhang, X., Ren, S., & Sun, J. (2015a). Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385. https://doi.org/10.48550/arxiv.1512.03385.

He, K., Zhang, X., Ren, S., & Sun, J. (2015b). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916. https://doi.org/10.1109/tpami.2015.2389824.

Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. arXiv preprint arXiv:1312.5402. https://doi.org/10.48550/arxiv.1312.5402.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. https://doi.org/10.48550/arxiv.1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2261–2269). https://doi.org/10.1109/CVPR.2017.243.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167. https://doi.org/10.48550/arxiv.1502.03167.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093. https://doi.org/10.48550/arxiv.1408.5093.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume *25* (pp. 1097–1105). https://doi.org/10.1145/3065386.

Lima-Junior, F. R. & Carpinetti, L. C. R. (2019). Predicting supply chain performance based on scor® metrics and multilayer perceptron neural networks. *International Journal of Production Economics*, *212*, 19–38. https://doi.org/10.1016/j.ijpe.2019.02.001.

Mahrishi, M., Morwal, S., Muzaffar, A. W., Bhatia, S., Dadheech, P., & Rahmani, M. K. I. (2021). Video index point detection and extraction framework using custom yolov4 darknet object detection model. *IEEE Access*, *9*, 143378–143391. https://doi.org/10.1109/access.2021.3118048.

Rota Bulò, S., Neuhold, G., & Kontschieder, P. (2017). In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2126–2135). https://doi.org/10.48550/arXiv.1712.02616.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arxiv.1409.1556.

Thompson, K. N. (1990). Vendor profile analysis. *Journal of Purchasing and Materials Management*, *26*(1), 11–18. https://doi.org/10.1111/j.1745-493x.1990.tb00494.x.

Tieleman, T. & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning. https://cir.nii.ac.jp/crid/1370017282431050757.

Wang, C., Liao, H. M., Yeh, I., Wu, Y., Chen, P., & Hsieh, J. (2019). Cspnet: A new backbone that can enhance learning capability of cnn. arXiv preprint arXiv:1911.11929. https://doi.org/10.48550/arxiv.1911.11929.

Wang, R. J., Li, X., & Ling, C. X. (2018). Pelee: a real-time object detection system on mobile devices. arXiv preprint arXiv:1804.06882. https://doi.org/10.48550/arxiv.1804.06882.

Yue, J., Mao, S., & Li, M. (2016). A deep learning framework for hyperspectral image classification using spatial pyramid pooling. *Remote Sensing Letters*, *7*(9), 875–884. https://doi.org/10.1080/2150704x.2016.1193793.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *International Conference on Computer Vision* (pp. 2018–2025). https://doi.org/10.1109/iccv.2011.6126474.

Zhang, Y., Jin, R., & Zhou, Z. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0.

Özkan, G. & İnal, M. (2014). Comparison of neural network application for fuzzy and anfis approaches for multi-criteria decision making problems. *Applied Soft Computing*, *24*, 232–238. https://doi.org/10.1016/j.asoc.2014.06.032.