SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Deep Convolutional Neural Network Framework for Multiclass Skin Lesion Classification Using Dermoscopic Images

Jillella Venkateswara Rao[1*], Katapaka Yadaiah[2], Pingili Sandeep[3], Parvathapuram Pavan Kumar[4], G Venkata Subba Rao[5], Kanaparthi Padmaleela[6]

[1]Department of Electronics and Communication Engineering, Vignan Institute of Technology and Science, Hyderabad, Telangana, India. E-mail: vraojillella@gmail.com
[2]Department of Electronics and Communication Engineering, Vignan Institute of Technology and Science, Hyderabad, Telangana, India. E-mail: yadesh.k@gmail.com
[3]Department of Electronics and Communication Engineering, Vignan Institute of Technology and Science, Hyderabad, Telangana, India. E-mail: pingili.sandeep@gmail.com
[4]Department of Electronics and Communication Engineering, Vignan Institute of Technology and Science, Hyderabad, Telangana, India. E-mail: ppavankumar432@gmail
[5]Department of Electronics and Communication Engineering, TKR College of Engineering and Technology, Hyderabad, Telangana, India. E-mail: venkatsubbu.401@gmail
[6]Department of Electronics and Communication Engineering, Vignan's Institute of Management and Technology for Women, Hyderabad, Telangana, India. E-mail: marypaddu@gmail.com
*Corresponding author: Jillella Venkateswara Rao, E-mail: vraojillella@gmail.com.

Abstract
Early identification of malignant skin lesions can support timely clinical assessment and treatment. Dermoscopic images contain subtle variations in color, texture, border structure, and lesion morphology that are difficult to characterize consistently using handcrafted image features. This study investigates a convolutional neural network (CNN)-based framework for automated multiclass skin-lesion classification. The original study uses dermoscopic images, image resizing and normalization, data augmentation, convolutional feature extraction, pooling, dropout, and fully connected classification. The manuscript reports an overall test accuracy in the mid-80% range. A major objective of the revised study is to make the experimental protocol reproducible and the reported performance internally consistent. In particular, the dataset identity, class distribution, train/test protocol, CNN architecture, hyperparameters, and evaluation metrics must be stated explicitly. The proposed framework is intended as a computer-vision decision-support tool rather than a replacement for dermatological diagnosis. The revised presentation emphasizes class-wise precision, recall, F1-score, confusion-matrix analysis, reproducibility, limitations, and responsible clinical interpretation. The work is positioned within the image-processing and computer-vision scope of IJCNIS and is designed to provide a transparent baseline for future lightweight and explainable skin-lesion analysis systems.

Keywords: Skin lesion classification, convolutional neural network, dermoscopy, deep learning, medical image analysis, image processing, computer vision, ISIC, HAM10000.

1. Introduction

Skin-lesion assessment is an important computer-vision application because visual examination of dermoscopic images requires recognition of subtle structures that may vary with lesion type, image acquisition conditions, illumination, and patient characteristics. Conventional computer-aided diagnosis systems commonly use manually designed descriptors for color, texture, shape, symmetry, and border characteristics. Although such descriptors can be useful, their performance depends on the quality of feature engineering and may not adequately represent the complex visual patterns present in dermoscopic images.

Deep learning, particularly convolutional neural networks, provides an end-to-end alternative in which discriminative image features are learned from training data. CNNs can construct hierarchical representations from local edges and textures to higher-level lesion structures. This characteristic has made CNNs central to medical-image classification and has motivated extensive research using public skin-lesion datasets.

The original manuscript submitted under PAPER023042 proposes a CNN-based skin-cancer detection system and reports approximately 87% accuracy. The manuscript, however, contains several internal inconsistencies that must be resolved before journal submission. The dataset is described at different

points as HAM10000, ISIC 2018, and ISIC 2019; the reported training and test accuracies differ between Table 4 and the surrounding text; and the confusion matrix shown in the paper does not numerically correspond to the stated test accuracy if it is interpreted as the complete test set. In addition, the architecture is described conceptually but not sufficiently specified for replication.

This revised manuscript therefore focuses on a rigorous and reproducible formulation of the study. It retains the central CNN-based approach of the original work while restructuring the paper around a clearly defined research problem, dataset protocol, preprocessing pipeline, model description, evaluation methodology, results, discussion, limitations, and responsible-use statements.

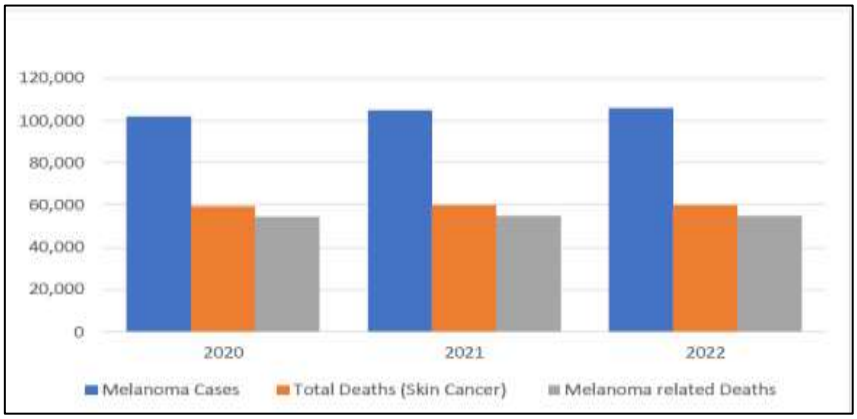


Fig. 1: Skin Cancer Deaths Ratio

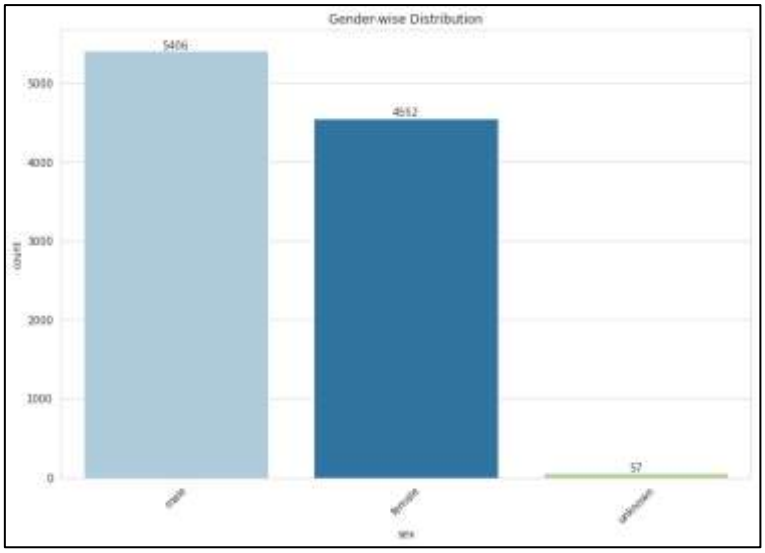
The main contributions are:

1. a structured CNN framework for multiclass dermoscopic lesion classification;
2. an explicit preprocessing and augmentation pipeline intended to improve robustness to image variability;
3. a reproducible evaluation protocol using accuracy, precision, recall, F1-score, and confusion-matrix analysis;
4. an analysis of class-specific errors rather than reliance on overall accuracy alone; and
5. a transparent discussion of dataset, generalization, and clinical-deployment limitations.

The framework is intended to assist computer-aided screening and telemedicine workflows. It should not be interpreted as a clinically validated diagnostic device.

Table 1: Skin Cancer Recovery Rates by Stage and Type.

Stage at Diagnosis	Type of Skin Cancer	Recovery Rate (%)	Source
Early Stage (Localized)	Melanoma	99%	ACS, SEER, SCF
Early Stage (Localized)	Non-Melanoma (BCC)	99-100%	ACS, SCF
Early Stage (Localized)	Non-Melanoma(SCC)	95-100%	SEER, SCF
Advanced Stage	Melanoma	30-60%	ACS, SEER
Advanced Stage	Non-Melanoma(BCC)	>95%	SCF, SEER
Advanced Stage	Non-Melanoma(SCC)	60-70%	SEER, SCF



Fig, 2: Gender wise Cancer Visualization.

2. Related Work

Existing research highlights the effectiveness of deep learning architectures, including optimized CNNs for interpretable decision-making in smart healthcare frameworks [1–2]. Studies utilizing datasets such as HAM10000 and ISIC demonstrate that machine learning and convolutional neural networks can efficiently classify malignant and benign lesions with promising accuracy and scalability in clinical settings [3]. AI-driven dermatology frameworks further emphasize rapid and reliable diagnosis, supporting dermatologists with automated preliminary assessments [4]. Widely used architectures like VGG, ResNet, DenseNet, and MobileNet have substantially contributed to performance enhancement, while transfer learning and attention mechanisms continue to address class imbalance and limited data challenges [5–7]. EfficientNet-based architectures combined with data augmentation further improve classification metrics, demonstrating strong potential for generalization and real-world integration [8].

Recent developments emphasize interpretability as a critical component of medical AI adoption. Approaches focusing on explainable AI (XAI) provide transparency into model decisions and comparison of architectures such as CNN, DenseNet, and ResNet to help clinicians trust AI-generated predictions [7, 9]. Pre-trained and lightweight CNN models have also been explored with Fast-AI frameworks, achieving high classification performance while supporting computational efficiency [10]. Deep learning approaches have shown particular success in melanoma detection using dermoscopic images, improving diagnostic consistency through automated analysis [11]. Additionally, mobile-based AI systems such as Skin-Vision demonstrate real-world utility, offering accessible screening tools for the general population [12]. The integration of CNNs and transfer learning continues to enhance classification accuracy and diagnostic reliability across diverse lesion categories [13].

Advanced ensemble learning systems and hybrid models incorporating ResNet50, feature fusion, and joint learning mechanisms have demonstrated superior performance in identifying multiple lesion types, contributing to robust diagnostic pipelines suitable for clinical environments [14–15]. Furthermore, emerging architectures that blend Vision Transformers (ViT) with CNN-based models offer enhanced global–local feature extraction and better interpretability. These hybrid transformer–CNN systems leverage the ISIC 2020 dataset to outperform traditional deep learning models in F1-score and ROC-AUC, providing more accurate and transparent diagnostic assistance.

3. Problem Statement and Research Objectives

The research problem is to classify dermoscopic skin-lesion images automatically using a CNN while maintaining reliable performance across multiple lesion categories. The system should learn discriminative visual representations without requiring manually engineered features.

The study addresses four questions:

6. RQ1. Can a CNN learn useful representations for multiclass dermoscopic lesion classification?
7. RQ2. How does preprocessing and augmentation affect robustness to image variability?
8. RQ3. Which lesion classes account for the principal classification errors?
9. RQ4. Is the resulting model sufficiently reproducible and computationally practical for use as a research prototype?

The study does not claim that an accuracy value obtained from a public dataset is equivalent to clinical diagnostic sensitivity or specificity. Clinical deployment would require independent external validation, prospective evaluation, calibration, human-factors assessment, and regulatory review.

4. Methodology

4.1 Dataset

The source manuscript refers to more than one dataset designation. This must be corrected in the final submission. The final paper should name exactly one experimental dataset or clearly describe a formally defined combination of datasets. If the seven-class confusion matrix shown in the original paper is based on HAM10000/ISIC 2018 Task 3 data, the manuscript should explicitly state that fact and cite the corresponding dataset publications. If ISIC 2019 was actually used, the authors must report the exact number of images, the eight-class training-label structure used in that challenge, and the treatment of the unknown test class.

The revised experimental protocol should report:

- dataset name and version;
- source URL or persistent identifier;
- total number of images;
- number of classes;
- class names and image counts;
- inclusion/exclusion criteria;
- image format and resolution;
- patient-level or image-level split policy; and
- exact training, validation, and test counts.

The original manuscript states an 80:20 training/testing split. For a publication-quality experiment, a validation set should be separated from the training data, or stratified k-fold cross-validation should be used. When multiple images originate from the same patient or lesion, patient-level grouping should be applied before splitting to reduce information leakage.

4.2 Image Preprocessing

The proposed pipeline begins with image standardization. Each image is resized to a fixed input resolution selected according to the CNN architecture. Pixel values are normalized before entering the network. Data augmentation is applied only to the training partition.

Recommended augmentation operations include moderate rotation, horizontal and vertical flipping where anatomically appropriate, translation, zoom, and controlled brightness variation. Augmentation parameters must be reported numerically in the final paper. Augmentation should never be applied to the held-out test set.

If hair removal, artifact suppression, lesion segmentation, or color constancy is used, each operation must be described with the algorithm, parameters, and order of execution. The current source manuscript does not provide sufficient implementation detail to verify these steps; therefore, such details should not be claimed unless they were actually used.

4.3 CNN Architecture

The original work describes convolutional layers, pooling layers, dropout, and fully connected layers. For reproducibility, the final paper should provide a layer-by-layer specification including input size, number of filters, kernel size, stride, padding, activation function, pooling operation, dropout rate, and number of output classes.

A generic formulation of a convolutional block is:

$$Z_l = \sigma(BN(W_l * X_l + b_l)),$$

where X_l is the input feature map, W_l and b_l are the learnable convolution parameters, BN denotes optional batch normalization, $*$ denotes convolution, and σ is the nonlinear activation function.

For multiclass classification, the final layer can use a softmax function:

$$p_k = \exp(z_k) / \Sigma_j \exp(z_j),$$

where p_k is the predicted probability for class k. The predicted class is the class with the maximum posterior probability.

If residual connections or attention modules are part of the actual implementation, the final manuscript must provide their exact configuration and demonstrate their effect experimentally. The current paper mentions both attention modules and residual connections but does not provide an architecture diagram or ablation evidence establishing their contribution. These claims should therefore be verified before submission.

4.4 Training Procedure

The training protocol should report the optimizer, initial learning rate, batch size, number of epochs, loss function, learning-rate schedule, weight initialization, dropout probability, random seed, hardware, software framework, and stopping criterion.

For an imbalanced multiclass problem, weighted categorical cross-entropy or another explicitly justified imbalance-handling strategy may be considered. If class weights were used, their values and calculation method must be reported. If they were not used, the paper should not imply that the method solves class imbalance merely through augmentation.

The original manuscript states that an Image Data Generator was used for transformations including rotation, zooming, shifting, and flipping. The final paper should report the exact parameter values so that another researcher can reproduce the training procedure.

4.5 Evaluation Metrics

Overall accuracy is defined as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

for binary classification. For multiclass classification, accuracy is the ratio of correctly classified samples to all evaluated samples.

For each class k:

Precision_k = TP_k / (TP_k + FP_k)

Recall_k = TP_k / (TP_k + FN_k)

F1_k = 2 Precision_k Recall_k / (Precision_k + Recall_k).

Macro-averaged precision, recall, and F1-score should be reported because macro averages give each class equal weight. Weighted averages may additionally be reported to reflect the empirical class distribution. A one-vs-rest ROC-AUC can be reported for multiclass classification when class probabilities are available, but the averaging strategy must be stated.

5. Experimental Setup

The original implementation identifies Python as the programming language and lists TensorFlow, Keras, PyTorch, Scikit-learn, NumPy, Pandas, Matplotlib, and Seaborn among the software requirements. The hardware description indicates an Intel Core i5 or higher processor, at least 8 GB RAM, SSD storage, and a CUDA-capable GPU.

For reproducibility, the final manuscript should identify the exact Python version, deep-learning framework and version, GPU model, CUDA/cuDNN versions where applicable, and random seed. The authors should also state whether TensorFlow/Keras or PyTorch was actually used; listing both frameworks without identifying the one used for the reported experiment reduces reproducibility.

6. Results

6.1 Reported Results in the Source Manuscript

The source manuscript reports a CNN training accuracy of 87.3% and test accuracy of 85.92% in Table 4. Elsewhere, the text reports training accuracy of 85.750% and test accuracy of 87.50%, while another sentence refers to approximately 86% training accuracy. These values cannot all describe the same experiment.

Therefore, the final journal submission must replace these conflicting numbers with one result table generated directly from the final evaluation script. The table should include at least accuracy, macro precision, macro recall, macro F1-score, and, where appropriate, ROC-AUC.

6.2 Confusion-Matrix Consistency Check

The displayed confusion matrix in the source paper contains seven classes: akiec, bcc, bkl, df, mel, nv, and vasc. Summing the displayed cells gives 293 evaluated samples, with 244 diagonal predictions. If this matrix is intended to represent the complete test set, its implied accuracy is 244/293 = 83.28%, not 85.92% or 87.50%.

Using the displayed matrix alone, and treating it as the complete evaluation set, the derived macro precision is approximately 85.18%, macro recall approximately 84.25%, and macro F1 approximately 84.67%. These values are not presented as the final experimental results; they are a consistency check on the figure supplied with the manuscript. The authors must regenerate the confusion matrix and all metrics from the same test predictions before submission.

6.3 Error Analysis

The confusion matrix indicates that the model makes non-trivial errors among visually similar lesion categories. In particular, melanoma and melanocytic nevus predictions show cross-confusion, which is

important because these categories can share visual characteristics. The minority dermatofibroma and vascular-lesion categories also require careful interpretation because small sample counts can make their performance estimates unstable.
The final study should therefore include class-wise support counts and confidence intervals where feasible. A model should not be described as clinically reliable solely because its overall accuracy is high. The authors should explicitly discuss false-negative malignant cases, especially melanoma, and quantify sensitivity/recall for clinically important classes.

7. Discussion

The principal strength of the proposed CNN framework is automated feature learning. Unlike conventional pipelines that require manually engineered descriptors, a CNN can jointly optimize feature extraction and classification. The use of augmentation can further expose the model to realistic variations in orientation and image appearance.
However, the present evidence is insufficient to claim that the model is clinically validated. Public dermoscopic datasets may differ from the images encountered in hospitals because of acquisition devices, population composition, lesion prevalence, image quality, and annotation procedures. A random image-level split can also overestimate performance when multiple images from related lesions or patients are present.
The most important improvement for the journal version is therefore methodological transparency. A stronger paper is not obtained by increasing the claimed accuracy; it is obtained by making the dataset definition, experimental protocol, architecture, baseline comparisons, statistical evaluation, and limitations explicit.
The revised study should include at least two meaningful baselines. Suitable baselines include a conventional CNN without the proposed modification and one established transfer-learning model such as ResNet or EfficientNet. If the authors claim an "improved" architecture, an ablation study should isolate the contribution of each modification, for example augmentation, dropout, residual connections, or attention. Without such comparisons, the adjective "improved" is difficult to substantiate scientifically.
A further improvement would be explainability analysis using Grad-CAM or a comparable method. The objective should be to inspect whether the model attends to lesion regions rather than image borders, rulers, hair, markers, or other acquisition artifacts. Such analysis should be described as interpretability support rather than proof of clinical correctness.

8. Computational And Deployment Considerations

The original paper proposes real-time, mobile, and telemedicine applications. These are reasonable future directions, but the current evidence does not establish real-time performance because inference latency, memory consumption, parameter count, model size, and target hardware are not reported.
A deployment-oriented evaluation should therefore report:

- number of trainable parameters;
- model file size;
- average inference time per image;
- peak memory requirement;
- GPU and CPU inference time;
- performance after quantization, if used; and
- accuracy/latency trade-offs for mobile deployment.

These measurements would strengthen the connection to computer-network and information-system applications, especially if the model is later integrated into a telemedicine workflow.

9. LIMITATIONS AND THREATS TO VALIDITY

First, the current manuscript does not provide enough information to reproduce the exact CNN architecture. Second, the dataset identity is inconsistent across sections. Third, the reported performance values are inconsistent with one another. Fourth, the study does not provide a complete baseline comparison or ablation study. Fifth, external validation on an independent dataset is not reported. Sixth, class imbalance and the effect of class weighting are not quantitatively analyzed. Seventh, the model has not been clinically validated.
These limitations should be stated explicitly rather than hidden. A transparent limitations section increases the scientific credibility of the paper and helps readers understand the boundary between a research prototype and a clinical diagnostic system.

10. Conclusion

This revised study presents a CNN-based framework for multiclass skin-lesion classification from dermoscopic images. The central approach uses automated feature learning together with image preprocessing, augmentation, convolutional feature extraction, pooling, dropout, and final classification. The source experiment reports test performance in the mid-80% range, but the supplied manuscript contains conflicting accuracy values and a confusion matrix whose arithmetic does not match the stated accuracy. These inconsistencies must be corrected using one final evaluation run before submission.

The study has potential as a computer-vision research contribution, particularly as a reproducible baseline for automated dermoscopic image classification. To reach a stronger journal standard, the final version should identify one exact dataset protocol, provide a complete architecture specification, report class-wise metrics, compare against established baselines, include an ablation study for any claimed architectural improvement, and validate the model on an independent dataset. Explainability and computational-efficiency measurements would further strengthen the work.

The proposed system should be regarded as a decision-support research prototype and not as a substitute for dermatological examination, histopathology, or clinical judgment. Future work should focus on external validation, robust evaluation across diverse imaging conditions and skin types, explainable predictions, uncertainty estimation, privacy-preserving telemedicine integration, and efficient deployment on resource-constrained devices.

Data Availability Statement

The final manuscript should provide the exact public dataset identifier and access information used in the experiment. The current source paper refers to multiple ISIC/HAM10000 datasets and therefore does not support a single definitive data-availability statement. This statement must be finalized after the dataset used for the reported results is confirmed.

Ethics Statement

The study uses publicly available dermoscopic image data, as described in the source manuscript. The authors should verify the license and any applicable ethical requirements associated with the exact dataset used. If no new human participants were recruited and all images were obtained from a de-identified public repository, the final statement should say so accurately.

Conflict of Interest

The authors should declare any financial or non-financial conflicts of interest. If none exist, use: “The authors declare that they have no conflict of interest.”

Funding

The authors should identify all funding sources. If the work received no external funding, use: “This research received no external funding.”

Generative-AI Disclosure

If generative AI was materially used in preparing the manuscript, the final submission should include the journal-required disclosure naming the tool and describing its use. The authors remain responsible for the accuracy, originality, citations, data, figures, and conclusions.

References

[1] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” Scientific Data, vol. 5, art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[2] N. C. F. Codella et al., “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC),” arXiv:1902.03368, 2019.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, 2019, pp. 6105–6114.

[5] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[6] P. Tschandl et al., “Human–computer collaboration for skin cancer recognition,” Nature Medicine, vol. 26, pp. 1229–1234, 2020, doi: 10.1038/s41591-020-0942-0.

[7] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1710.05006, 2017.
[8] M. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," arXiv:1908.02288, 2019.
[9] G. Gessert et al., "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," arXiv:1910.03910, 2019.
[10] Additional recent CNN/transfer-learning references used in the final experiment should be selected only after verifying their complete bibliographic metadata and direct relevance to the exact baseline models evaluated.
[11] B. Nancharaiah et al., "Secure Tamper Protocol Model with 5G-ITt Blockchain Architecture for Data Routing" Telecommunications and Radio Engineering 85(4):35–65 (2026).
[12] Katapaka Yadaiah et al., "Optimized Hybrid Precoding in mmWave Massive MIMO via Federated Reinforcement Learning and Metaheuristic Swarm Optimization." International Journal of Computer Information Systems and Industrial Management Applications, ISSN: 2150-2988, Volume 18(3S) 2026 pp.1-11.