PimpleNet: Breaking Out the CNN

PurelyBiome

Abstract— **User-submitted facial images contain valuable information about visible acne severity, but this signal can be difficult to interpret consistently due to variation in lighting, camera angle, skin tone, makeup, and image quality. This project develops a CNN-based acne severity classification model to support PurelyBiome's Kit Activation workflow and strengthen personalization beyond microbiome data alone. Using public facial image datasets and representative PurelyBiome customer images, the final transfer-learning model achieved 95.56% validation accuracy, a 95.30% weighted F1-score, and a 90.81% macro F1-score across three acne severity classes: Clear Skin, Mild/Moderate, and Severe. The model also achieved 100% accuracy on 50 held-out PurelyBiome customer images, demonstrating strong generalization to representative user-submitted photos. By estimating visible acne severity from user images, the model provides a structured visual signal that can be paired with microbiome profiles to support more personalized skincare recommendations. Over time, these image-derived acne severity targets may help PurelyBiome identify microbiome patterns associated with acne-prone states and use microbiome data as an early signal that a user's skin system is trending toward acne. Integrated with microbiome insights, this visual layer can make PurelyBiome's recommendations more targeted, objective, and responsive to each user's changing skin needs.**

I. Introduction

Personalized skincare is often limited by incomplete information. Most users choose products based on broad skin-type categories, visible symptoms, or trial and error, even though skin concerns such as acne can vary significantly across individuals and across different areas of the same face. For PurelyBiome, which combines at-home skin microbiome testing with personalized skincare recommendations, user-submitted facial images provide an opportunity to add a second layer of personalization: the visible presentation of the skin.

During PurelyBiome's Kit Activation workflow, users submit facial images alongside their microbiome sample. These images contain valuable information about acne severity, distribution, and progression. A model that can automatically classify acne severity from these images could help PurelyBiome better understand the user's current skin state and connect that visual signal with microbiome results. This is especially useful because acne is not always evenly distributed; users may experience more visible breakouts on specific areas such as the forehead, cheeks, chin, or jawline. By identifying visible acne patterns, PurelyBiome can move toward more targeted skincare routines, including product suggestions tailored to specific areas of the face.

This project develops a convolutional neural network-based acne severity classification model for user-submitted facial images. The model is designed to classify images into three acne severity categories: Clear Skin, Mild/Moderate, and Severe. Using public facial image datasets and representative PurelyBiome customer images, the final transfer-learning model achieved 95.56% validation accuracy, a 95.30% weighted F1-score, and a 90.81% macro F1-score across the three classes. The model also achieved 100% accuracy on 50 held-out PurelyBiome customer images. These results demonstrate the potential of image-based acne grading as a practical tool for improving skincare personalization.

Beyond providing a visual severity estimate, the model also creates structured acne severity targets that can be paired with microbiome data. Over time, this may allow PurelyBiome to identify microbiome patterns associated with acne-prone states and use microbiome measurements as an early signal that a user's skin system is trending toward acne. In this way, the image model supports both immediate personalization based on visible skin presentation and longer-term learning from the relationship between microbiome profiles and acne severity.

The goal of this work is to build a working prototype that can support PurelyBiome's product workflow. By combining visual acne severity with microbiome insights, PurelyBiome can make recommendations that are more objective, localized, and responsive to each user's changing skin needs. This approach creates a foundation for future features such as progress tracking, region-specific product recommendations, acne segmentation, and more personalized routines informed by both biological and visual skin data.

II. Related Work

Computer vision has become an increasingly important tool for analyzing acne severity from facial images. Traditional acne assessment methods often rely on manual grading, lesion counting, or rule-based image processing, which can be difficult to scale across real-world user-submitted images with variation in lighting, pose, skin tone, makeup, and image quality. Recent work has shown that convolutional neural networks can learn visual features useful for acne detection, lesion localization, and severity grading directly from facial images.

Shen et al. proposed an automatic facial acne diagnosis method based on convolutional neural networks, demonstrating that CNNs can classify acne-related visual categories from facial images [1]. Their work helped establish deep learning as a practical approach for automated acne image analysis. Huynh et al. later developed an automated acne object detection and severity grading approach, showing that acne analysis can be framed not only as whole-image classification, but also as lesion detection and grading [2]. Wen et al. explored interpretable CNN models for acne lesion localization and severity evaluation, highlighting the importance of both prediction accuracy and interpretability in acne-related image analysis [3].

Public datasets have also played an important role in making acne grading research more reproducible. Wu et al. introduced the ACNE04 dataset as part of their work on joint acne image grading and lesion counting using label distribution learning [4]. ACNE04 includes acne severity annotations and lesion count information, making it useful for training and evaluating acne grading models. More recent work has continued to build on acne image datasets for related tasks such as lesion segmentation, lesion scoring, and overall severity estimation. These approaches show a broader trend toward automated systems that can estimate acne severity from facial images in a scalable and repeatable way.

Transfer learning is commonly used in skin image classification because task-specific dermatology datasets are often smaller than general computer vision datasets. Pretrained CNN architectures provide visual feature extractors that can be fine-tuned for specific downstream tasks. EfficientNet is especially relevant because it was designed to improve accuracy and efficiency through compound scaling of network depth, width, and image resolution [**?**]. This makes EfficientNet well suited for practical workflows where model performance, training efficiency, and future deployment feasibility are all important.

Most existing acne classification work focuses on acne assessment as a standalone computer vision task. In contrast, this project applies acne severity classification within PurelyBiome's personalized skincare workflow. The goal is not only to classify visible acne severity, but also to create structured image-derived acne severity targets that can be paired with microbiome data. This connection may allow PurelyBiome to identify microbiome patterns associated with acne-prone states and eventually use microbiome measurements as an early signal that a user's skin system is trending toward acne. By integrating image-based acne grading into Kit Activation, this work connects computer vision with microbiome-informed personalization and supports future extensions such as region-specific recommendations and acne segmentation.

III. Dataset

This project used a combination of public facial image datasets and representative PurelyBiome customer images. Acne-positive images were sourced from ACNE04, a facial acne dataset introduced by Wu et al. for joint acne grading and lesion counting. ACNE04 includes facial images annotated by acne severity, making it suitable for image-level acne severity classification.

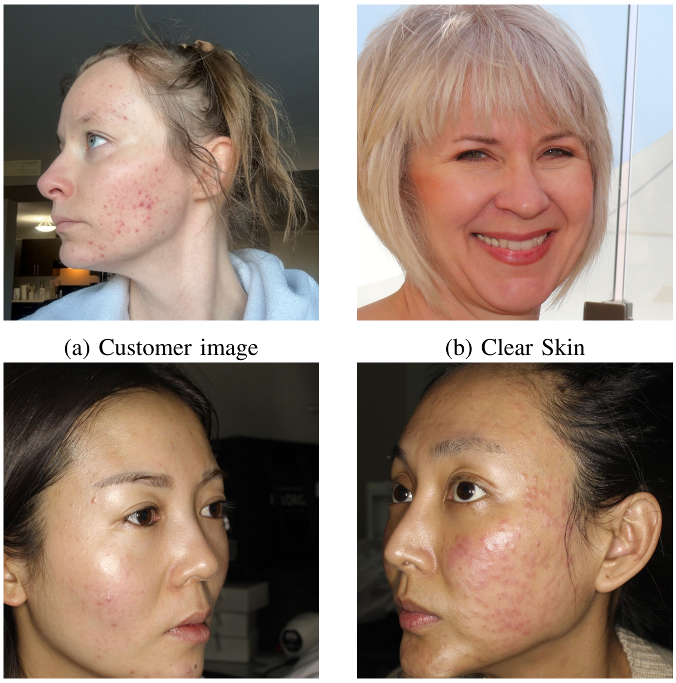
The original ACNE04 severity labels were consolidated into two acne-positive categories. Classes 0 and 1 were combined to form the Mild/Moderate class, while classes 3 and 4 were combined to form the Severe class. This grouping created broader and more practical severity categories for product-facing use.

To represent the Clear Skin class, images were selected from the FFHQ dataset, a high-quality facial image dataset containing diverse human faces. These images provided examples of faces without visible acne severity, allowing the model to distinguish clear skin from acne-present cases.

The final dataset was organized into three categories: Clear Skin, Mild/Moderate, and Severe. The final class distribution included 1,106 Clear Skin images, 1,106 Mild/Moderate images, and 271 Severe images, for a total of 2,483 images. Of these, 1,987 images were used for training and 496 images were used for validation. An additional 50 representative PurelyBiome customer images were reserved as held-out testing data.

Because the Severe class contained fewer images than the Clear Skin and Mild/Moderate classes, the dataset was imbalanced. This imbalance was addressed during model training using class weighting, where the underrepresented Severe class was assigned a higher training weight. This helped reduce bias toward the more frequent classes and encouraged the model to learn acne-related features from the smaller Severe class.

These categories were selected to align with PurelyBiome's Kit Activation workflow, where model outputs should be simple, interpretable, and useful for skincare personalization. The held-out customer images reflect the type of user-submitted photos PurelyBiome receives during Kit Activation and help connect the dataset to the company's real-world product workflow. Together, ACNE04, FFHQ, and held-out customer images provided the foundation for developing an acne severity grading model to support personalized skincare recommendations.



(a) Customer image

(b) Clear Skin

(c) Mild/Moderate

(d) Severe

Fig. 1: Representative images used in the dataset, including a held-out PurelyBiome customer image, a clear-skin FFHQ image, and acne-positive examples from the Mild/Moderate and Severe classes.

IV. Methodology

The acne severity grading model was developed as a three-class image classification system designed to categorize facial images into Clear Skin, Mild/Moderate, and Severe classes. The goal of the methodology was to create a practical visual classification pipeline that could be integrated into PurelyBiome's Kit Activation workflow and provide an additional signal for skincare personalization.

All images were first organized according to their final class labels. Acne-positive images from ACNE04 were grouped into broader severity categories, while FFHQ images were used to represent the Clear Skin class. Representative PurelyBiome customer images were kept separate as held-out customer data to evaluate how well the model performed on real user-submitted images.

Before training, images were resized to a fixed input resolution of 224 $\times$ 224 pixels and formatted for compatibility with the convolutional neural network. Standard image preprocessing was applied so that all inputs shared a consistent shape and numerical format. This ensured that images from different sources could be passed through the same model pipeline.

Data augmentation was applied to the training images to improve robustness and reduce overfitting. The augmentation pipeline included random horizontal flipping, rotation, zooming, translation, contrast adjustment, and brightness adjustment. These transformations were selected to simulate realistic variation in user-submitted facial images, such as differences in camera angle, framing, lighting, and image brightness. Augmentation was applied only to the training dataset, while the validation dataset was left unmodified to provide a consistent evaluation of generalization performance.

Because the dataset was imbalanced, with fewer images in the Severe class than in the Clear Skin and Mild/Moderate classes, class weighting was applied during training. Balanced class weights were computed from the training labels and used to increase the contribution of underrepresented classes during optimization. This helped reduce bias toward the larger classes and encouraged the model to learn features associated with the smaller Severe class.

The classification model was built using a convolutional neural network with a pretrained EfficientNetB0 backbone. Transfer learning was used so that the model could benefit from visual features learned from large-scale image datasets, such as edges, textures, shapes, and facial structures. A custom classification head was added on top of the backbone to map extracted image features to the three acne severity categories.

Training was performed in two stages. In the first stage, the pretrained backbone was frozen and only the classification layers were trained. This allowed the model to learn the acne severity task while preserving the general visual representations from the pretrained network. In the second stage, selected deeper layers of the backbone were unfrozen and fine-tuned using a lower learning rate. This allowed the model to adapt higher-level visual features to acne-specific patterns while reducing the risk of overfitting.

Model performance was evaluated using validation accuracy, macro F1-score, weighted F1-score, per-class precision, per-class recall, and a confusion matrix. Accuracy was used as the primary overall metric because the task was framed as a multi-class classification problem. However, because the dataset was imbalanced, macro F1-score was also included to give equal importance to each class regardless of class frequency. Weighted F1-score was reported as an additional summary metric that accounts for class support. Performance was assessed on validation data and on held-out PurelyBiome customer images to measure both benchmark performance and relevance to the company's real-world Kit Activation workflow.

The final trained model produced class probabilities for each input image. The class with the highest predicted probability was selected as the model output. This output can be used as a visual severity signal alongside microbiome testing results, helping PurelyBiome personalize skincare recommendations based on both biological and image-based information.

V. Results

The model achieved strong performance across the three acne severity categories: Clear Skin, Mild/Moderate, and Severe. After training and fine-tuning, the final model reached a validation accuracy of 95.56%, a weighted F1-score of 95.30%, and a macro F1-score of 90.81%. The model also achieved 100% accuracy on 50 held-out PurelyBiome customer images. These results demonstrate that the model was able to learn visually meaningful differences between clear skin, mild/moderate acne, and severe acne, while also generalizing to representative user-submitted images from PurelyBiome's Kit Activation workflow.

Figure 2 shows the training and validation accuracy across the two training phases. During Phase 1, the pretrained EfficientNetB0 backbone was frozen while the classification head was trained. Accuracy increased rapidly in the early epochs, with validation accuracy rising from approximately 73% to above 90%. By the end of Phase 1, both training and validation accuracy stabilized around 93%, indicating that the model had learned a strong initial representation for the three-class classification task.

In Phase 2, selected deeper layers of the backbone were unfrozen and fine-tuned. This stage further improved performance, with training accuracy reaching approximately 96% and validation accuracy remaining consistently around 95%. The validation curve remained close to the training curve throughout fine-tuning, suggesting that the model improved without substantial overfitting. The final validation accuracy of 95.56% was selected as the representative benchmark result.

Because the dataset was imbalanced, accuracy was supplemented with macro F1-score and weighted F1-score. Macro F1-score is especially important because it gives equal weight to each class regardless of class frequency, making it more informative for evaluating performance on the smaller Severe class. As shown in Table I, the model performed strongly on the Clear Skin and Mild/Moderate classes, achieving F1-scores

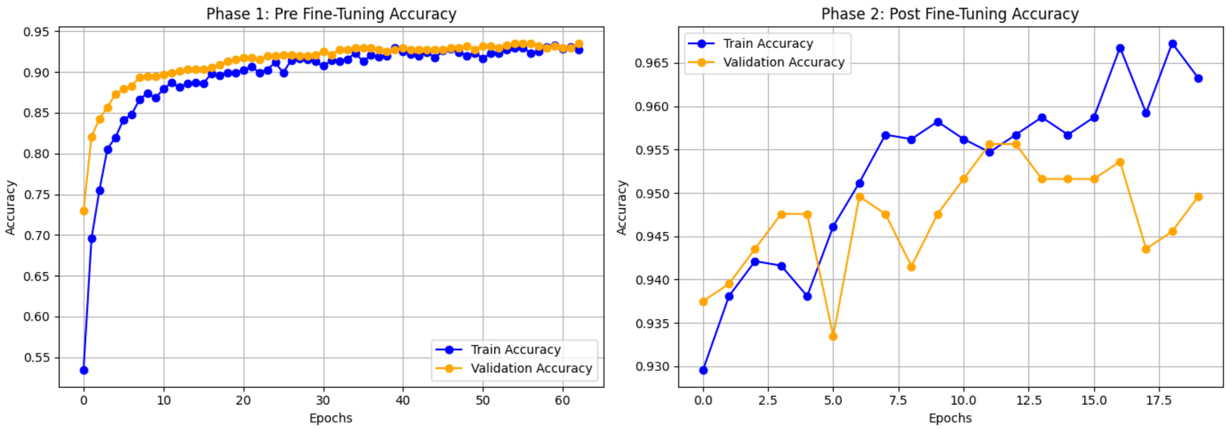


Fig. 2: Training and validation accuracy across the two-stage training process. Phase 1 trained the classification head with the pretrained backbone frozen, while Phase 2 fine-tuned selected deeper layers of the backbone.

of 0.9977 and 0.9524, respectively. The Severe class achieved a precision of 0.9231 and an F1-score of 0.7742, although its recall was lower at 0.6667.

TABLE I: Validation performance of the acne severity classification model. Macro F1-score is included to account for class imbalance across the three severity categories.

Class	Precision	Recall	F1-score	Support
Clear Skin	1.0000	0.9954	0.9977	219
Mild/Moderate	0.9205	0.9865	0.9524	223
Severe	0.9231	0.6667	0.7742	54
Accuracy	–	–	0.9556	496
Macro Avg.	0.9479	0.8829	0.9081	496
Weighted Avg.	0.9559	0.9556	0.9530	496

Table II shows the validation confusion matrix. The model correctly classified 218 of 219 Clear Skin images and 220 of 223 Mild/Moderate images. For the Severe class, the model correctly classified 36 of 54 images, while 18 Severe images were classified as Mild/Moderate. This suggests that most remaining errors occurred along the boundary between Mild/Moderate and Severe acne rather than between clear skin and acne-present categories.

TABLE II: Validation confusion matrix for the three-class acne severity model. Rows represent true labels and columns represent predicted labels.

True / Predicted	Clear Skin	Mild/Moderate	Severe
Clear Skin	218	1	0
Mild/Moderate	0	220	3
Severe	0	18	36

Figure 3 shows an example inference result on a representative PurelyBiome customer image. The model classified the image as Severe with a confidence score of 62.85%. The full confidence breakdown was 1.24% for Clear Skin, 35.91% for Mild/Moderate, and 62.85% for Severe. This example illustrates how the model outputs both a final predicted class and interpretable class probabilities, which can be used as a visual severity signal within PurelyBiome's Kit Activation workflow.

Fig. 3: Example model inference on a representative PurelyBiome customer image.

Overall, these results show that the proposed image-based acne severity classifier can provide a reliable visual signal for PurelyBiome's personalization workflow. By assigning each submitted facial image to an interpretable severity category, the model can support skincare recommendations that are better aligned with the visible condition of different users' skin.

VI. Discussion

The results demonstrate that a transfer learning-based convolutional neural network can effectively classify facial images into three acne severity categories: Clear Skin, Mild/Moderate, and Severe. The model achieved a validation accuracy of 95.56%, a weighted F1-score of 95.30%, and a macro F1-score of 90.81%. It also correctly classified all 50 held-out

PurelyBiome customer images, indicating strong potential for use as part of PurelyBiome's Kit Activation workflow.

A key strength of the approach is its practical class structure. Rather than predicting many fine-grained acne grades, the model outputs three interpretable categories that are easier to connect to skincare personalization. This is important for a consumer-facing workflow, where predictions should be simple enough to support product recommendations and routine guidance. The Clear Skin, Mild/Moderate, and Severe categories provide a useful visual signal that can complement microbiome test results and help personalize recommendations based on both biological and image-based information.

The two-stage training process also contributed to strong performance. During the first phase, the frozen pretrained backbone allowed the model to quickly learn the classification task using general visual features. During the second phase, fine-tuning selected deeper layers allowed the model to better adapt to acne-specific image patterns. The training curves show that validation accuracy remained close to training accuracy, suggesting that the model generalized well and did not show substantial overfitting during fine-tuning.

Because the dataset was imbalanced, macro F1-score provides an important additional measure beyond accuracy. While the overall validation accuracy was high, the Severe class had lower recall than the Clear Skin and Mild/Moderate classes. The confusion matrix showed that most Severe errors occurred when Severe images were classified as Mild/Moderate. This suggests that the model learned strong separation between clear skin and acne-present images, but that the boundary between Mild/Moderate and Severe acne remains more challenging. This is expected because the Severe class contained fewer training examples and because acne severity can exist along a visual spectrum rather than as sharply separated categories.

The customer inference example further illustrates the practical value of the model. In addition to producing a final severity class, the model provides a confidence distribution across all three categories. This probability breakdown is useful because acne severity can be visually ambiguous, especially for images that fall between Mild/Moderate and Severe. In the example shown, the model predicted Severe while still assigning a meaningful probability to Mild/Moderate, reflecting the uncertainty that can occur in real-world user-submitted images.

An important implication of this work is that image-based acne severity labels can serve as practical targets for connecting visible skin presentation with microbiome data. By classifying user-submitted facial images into Clear Skin, Mild/Moderate, and Severe categories, PurelyBiome can begin linking specific microbiome profiles to visible acne severity. Over time, this creates the opportunity to use microbiome measurements as an early signal that a user's skin system may be trending toward acne-prone states, even before visible severity increases. In this way, the image model does not only provide a standalone visual assessment; it also helps generate structured acne severity targets that can be paired with microbiome features to support earlier and more personalized skincare recommendations.

Despite these promising results, several limitations remain. First, the dataset combines images from multiple sources, including ACNE04, FFHQ, and PurelyBiome customer images. Differences in lighting, camera quality, pose, background, and image resolution may affect model behavior. Although this diversity can improve robustness, it may also introduce dataset-specific biases if certain image sources are strongly associated with particular labels. Second, the Severe class was smaller than the Clear Skin and Mild/Moderate classes, which likely contributed to lower Severe recall. Increasing the number and diversity of Severe examples would likely improve sensitivity for the most underrepresented category. Third, the held-out customer set provides an important real-world validation signal, but a larger customer test set would be needed to more fully evaluate performance across different skin tones, lighting conditions, acne presentations, and image capture environments.

Another limitation is that the current model performs image-level classification rather than region-level analysis. While image-level severity is useful for a first version of the product workflow, acne can vary across different areas of the face. A user may have more visible acne on the cheeks while having clearer skin on the forehead or chin. As a result, a single whole-face prediction may not fully capture localized acne patterns that could be important for skincare personalization.

Future work should extend the current image-level classification model into a segmentation-based framework. A segmentation model could identify and localize acne-affected areas directly within the image, enabling more detailed region-level severity analysis. This would allow PurelyBiome to move beyond whole-face classification and support more targeted skincare personalization, such as recommending products or routines for specific areas of the face. Segmentation outputs could also provide more granular labels for linking visible acne patterns with microbiome profiles, further improving the ability to detect when a user's skin system may be trending toward acne-prone states.

Overall, this project shows that an image-based acne severity classifier can provide a strong visual signal for skincare personalization. When combined with PurelyBiome's microbiome testing data, the model can help create a more complete understanding of each user's skin profile and support recommendations that are informed by both visible skin presentation and underlying biological signals.

VII. Conclusion

This project developed a deep learning-based acne severity grading model for classifying facial images into three interpretable categories: Clear Skin, Mild/Moderate, and Severe. Using a combination of ACNE04 acne-positive images, FFHQ clear-skin images, and held-out PurelyBiome customer images, the model was designed to support PurelyBiome's Kit Activation workflow and provide an additional visual signal for skincare personalization.

The final model achieved 95.56% validation accuracy, a 95.30% weighted F1-score, and a 90.81% macro F1-score across the three acne severity classes. It also achieved 100% accuracy on 50 held-out PurelyBiome customer images. These results show that transfer learning with an EfficientNetB0 backbone can effectively learn acne-related visual patterns and generalize to representative user-submitted images. The two-stage training approach, consisting of frozen-backbone training followed by selective fine-tuning, enabled the model to adapt pretrained visual features to the acne severity classification task.

Beyond image classification alone, this work establishes a foundation for connecting visible acne severity with microbiome data. By producing structured acne severity targets from user-submitted images, the model creates an opportunity to study how microbiome profiles relate to acne-prone skin states. Over time, this may allow PurelyBiome to use microbiome measurements as an early signal that a user's skin system is trending toward acne, enabling earlier and more personalized skincare recommendations.

Future work should expand the size and diversity of the held-out customer dataset, improve sensitivity for the Severe class, evaluate performance across different lighting conditions and skin tones, and extend the model toward acne segmentation. A segmentation-based approach would allow acne-affected regions to be localized within the face, supporting more targeted recommendations by facial area. Overall, the proposed model demonstrates strong potential as a practical visual AI component within PurelyBiome's broader microbiome-based personalization platform.

References

[1] Xiaolei Shen, Jiachi Zhang, Chenjun Yan, and Hong Zhou. An automatic diagnosis method of facial acne vulgaris based on convolutional neural network. *Scientific reports*, 8(1):5839, 2018.

[2] QT Huynh, PH Nguyen, HX Le, LT Ngo, NT Trinh, MTT Tran, HT Nguyen, NT Vu, AT Nguyen, K Suda, et al. Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. diagnostics. 2022; 12 (8): 1879, 1879.

[3] Hao Wen, Wenjian Yu, Yuanqing Wu, Jun Zhao, Xiaolong Liu, Zhexiang Kuang, and Rong Fan. Acne detection and severity evaluation with interpretable convolutional neural network models. *Technology and Health Care*, 30(1_suppl):143–153, 2022.

[4] Xiaoping Wu, Ni Wen, Jie Liang, Yu-Kun Lai, Dongyu She, Ming-Ming Cheng, and Jufeng Yang. Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10642–10651, 2019.