Bulletin of Nexus

Multidisciplinary International Research Journal, Double-Blind Peer Reviewed, Open Access Journal

ISSN: 3065-7865 / Volume-3, Issue-4, April 2026

Original Article

Skin Burn Classification Using ResNet Models: A Deep Learning Approach Using ResNet18 and ResNet50

Riddhi Parab

Student, Department of Information Technology and Mathematics, The S.I.A. College of Higher Education (Autonomous) University of Mumbai Dombivli (East)

Manuscript ID: BN-2026-030406

ISSN: 3065-7865

Volume 3

Issue 4

April 2026

Pp. 36-41

Submitted: 05 Mar. 2026

Revised: 20 Mar. 2026

Accepted: 09 Apr. 2026

Published: 30 Apr. 2026

DOI: 10.5281/zenodo.21187942

DOI link: https://doi.org/10.5281/zenodo.21187942

Abstract

Skin burn injuries are major health problems around the world which need to be correctly diagnosed to determine the seriousness of the case and decide on how to deal with them. Misdiagnosis means late and improper treatment, which ultimately leads to a bad outcome. Skin burn injuries can be subjectively diagnosed through visual examination. In this study, we explore skin burn injury classification based on deep convolutional neural networks, where the main aim is to find the most effective architecture in terms of ResNet18 and ResNet50 models. This study employs an adequate dataset consisting of 20,418 training images, 4,498 validation images, and 4,532 test images of skin burn injuries with 3 labels, which include first-degree, second-degree, and third-degree burns. We made our model more effective by applying data augmentation and preprocessing. We used the pretrained ImageNet and adjusted its parameters to classify our images into three classes of burns. For the ResNet18 model, we achieved test accuracy of 91.22%, while for ResNet50, we managed to get an accuracy of 97.53%. Our findings indicate that ResNet50 performs significantly better than ResNet18.

Keywords: *Deep Learning, Skin Burn Classification, ResNet18, ResNet50, convolutional neural networks, transfer learning, image augmentation, medical image analysis, clinical decision support.*

Introduction

Burn injuries represent a significant global health challenge, with millions of cases reported annually. These injuries often result in long-term physical disabilities and substantial economic strain on healthcare systems. Currently, burn diagnosis remains largely subjective, typically relying on the visual evaluation of healthcare providers. However, this process is susceptible to variability in clinical judgment based on experience and environmental conditions. To improve early diagnosis and reduce errors, machine learning methods particularly CNN-based deep learning are increasingly being adopted to provide objective assessments. The success of deep learning in medical imaging is well-documented, with applications ranging from general overviews of the field to specific, high-stakes tasks like dermatologist-level skin cancer classification. Comprehensive surveys of these technologies have shown their broad efficacy in detecting tumors and retinal diseases, establishing a strong precedent for their use in burn care. A critical breakthrough in this field was the introduction of Residual Networks (ResNets), which utilize identity shortcut connections to enable the training of much deeper architectures without the typical degradation problems. Recent research has sought to push these boundaries further through advanced data augmentation and specialized training loops to improve detection accuracy. While some recent models have achieved various levels of success, they often face challenges with regularization and maintaining high accuracy across all burn degrees.

Address for correspondence:

Riddhi Parab, Student, Department of Information Technology and Mathematics, The S.I.A. College of Higher Education (Autonomous) University of Mumbai Dombivli (East)

How to cite this article:

Parab, R. (2026). Skin Burn Classification Using ResNet Models: A Deep Learning Approach Using ResNet18 and ResNet50. 3(4), 36–41. https://doi.org/10.5281/zenodo.21187942

Quick Response Code:

Website: https://bnir.us

OPEN ACCESS

This study builds upon these foundational works by providing a comprehensive comparison of ResNet18 and ResNet50. By evaluating these models on a balanced dataset, we aim to address the gaps identified in current literature surveys regarding model depth and clinical generalization. Our ResNet50 model reached a test accuracy of 97.53%, significantly outperforming previous benchmarks through a combination of deeper residual connections, extensive data augmentation, and optimized regularization strategies.

Literature Survey

Early methods for assessing burn severity relied heavily on the manual extraction of hand-crafted features combined with traditional machine learning algorithms. While these methods provided an initial framework for automated diagnosis, they often struggled with the visual complexity and physiological variations inherent in skin tissue. These limitations necessitated a shift toward more robust, automated feature- learning architectures. [1] The rise of Deep Learning, particularly through Convolutional Neural Networks (CNNs), has revolutionized medical image analysis. Researchers have demonstrated that CNNs can achieve performance levels comparable to human experts in various dermatological tasks. This success across general medical imaging provided the foundational evidence needed to apply deep learning specifically to the challenges of burn classification. [2] Initial applications of end-to-end CNNs in burn care showed promising results, with models achieving accuracy around 89.3%. By moving away from manual feature engineering, these studies proved that deep learning could effectively capture the nuances of burn depth directly from raw pixel data, though there remained significant room for improvement in classification precision. [3] To further enhance accuracy, researchers began utilizing transfer learning with pretrained models such as VGG16 and MobileNet. These architectures, originally trained on massive datasets like ImageNet, reached accuracy of 91.2% in burn detection. These findings highlighted the importance of using established architecture as a starting point for specialized medical tasks. [4]

Residual Networks (ResNets) were introduced to address the "vanishing gradient" problem, allowing for the training of much deeper neural networks. By using identity shortcut connections, ResNets enable the training of models with dozens or even hundreds of layers, which is crucial for capturing the high-level abstractions required to distinguish between different degrees of skin burns. [5] Specific explorations into ResNet18 for burn classification demonstrated that even "shallower" residual models could perform well, reaching accuracies over 90%. These studies suggested that the residual block architecture was particularly wellsuited for the textures found in medical burn imagery, providing a balance between computational efficiency and diagnostic reliability. [6]. Further advancements in data augmentation techniques significantly boosted the performance of ResNet-based models. By artificially expanding datasets with rotations and flips, researchers achieved accuracy of 91.5%. This underscored the fact that model architecture alone is not enough; the variety and quality of the training data are equally critical for clinical generalization. [7] Recent attempts to use deeper architectures like ResNet50 have explored the impact of aggressive regularization. However, some studies found that without a carefully balanced dataset, even deep models could drop in performance to approximately 86.9%. This highlighted the necessity of maintaining a balanced class distribution to ensure the model does not become biased toward more common burn types. [8] Comparative studies between different CNN variants, such as EfficientNet and ResNet, have shown that while efficiency is important, the depth and gradient flow of ResNet50 often provide a superior edge in accuracy. These comparisons serve as a benchmark for current research, pushing the boundaries of what automated clinical decision support systems can achieve in emergency care. [9]

Dataset Description

Dataset

The dataset was obtained from Kaggle, under the title Skin Burns Dataset. The raw dataset was highly imbalanced, with some classes (notably third-degree burns) underrepresented. To address this issue:

- We removed excess images from overrepresented classes, primarily in the first and second-degree burn categories.
- We transferred underrepresented samples from the training set to the test and validation sets to ensure balanced splits.
- The final structure was reorganized into three clean subsets: train, valid, and test, each containing classwise folders.

The dataset was evenly distributed among the three burn types first, second, and third degree and divided into training, validation, and test subsets. The training set contained 20,418 images, the validation set included 4,498 images, and the test set had 4,532 images, with each class represented in roughly equal proportion within each split. This balanced distribution minimized class imbalance and ensured fair training and evaluation of the models.

ISSN: 3065-7865
Bulletin of Nexus, Volume 3| Issue 4 | April 2026
Website: https://bnir.us

(a) 1st Degree Burn	(b) 2nd Degree Burn	(c) 3rd Degree Burn

Fig. 1: Example images representing each burn category from the dataset

Preprocessing

To enhance generalization and minimize overfitting, we applied the following preprocessing and augmentation methods:

- Resize to 224 × 224
- Random horizontal flip
- Random rotation up to ±20◦
- Random resized crop
- Color jitter (brightness, contrast, saturation)
- Normalization using ImageNet means and standard deviation

Algorithms Used

ResNet18 (Baseline Model):

ResNet18 was selected as the starting point for our experiments because of its relatively simple structure and low computational demand. It consists of 18 layers and uses shortcut (residual) connections that help in training deeper networks without vanishing gradients. We used the ImageNetpretrained weights and modified the final layer to output predictions for the three burn severity classes.

Evaluation Metrics

We evaluated model performance using accuracy, precision, recall, and F1-score to measure how effectively the burns were classified across severity levels.

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

- Precision: This metric checks how many of the model's positive predictions were actually correct. It is useful when the cost of false positives is high.

Precision $= \frac{TP}{TP+FP}$

- Recall: Also known as sensitivity, recall focuses on the model's ability to correctly identify all relevant positive cases. It is especially important in situations where missing a positive case could be harmful.

Recall $= \frac{TP}{TP+FN}$

- F1-score: The F1-score combines precision and recall into a single metric by taking their harmonic mean. It provides a balanced view when both false positives and false negatives are important.

F1-score =2 · $\frac{\text{Precision * Recall}}{\text{Precision + Recall}}$

- TP = True Positive
  → Model correctly predicts the positive class

For training, the Adam optimizer was used with a learning rate of 0.001, and no regularization was applied. A StepLR scheduler was implemented to reduce the learning rate by a factor of 0.1 every 7 epochs.

ResNet50 (Improved Architecture):

To improve upon the baseline results, we used ResNet50 a deeper model that incorporates bottleneck layers, allowing better feature representation while keeping computational efficiency. It offers stronger learning capacity, which is particularly useful for distinguishing subtle burn variations.

Similar to ResNet18, we used pretrained weights from ImageNet and replaced the final layer to classify into three burn types. Training was done using the Adam optimizer with a lower learning rate of 0.0001. L2 regularization (weight decay of 1×10−4) was applied to reduce overfitting. We also used a StepLR scheduler (factor of 0.1 every 7 epochs) and enabled early stopping to halt training if no improvement was observed. The best model was selected based on validation performance.

- Accuracy: This measures the overall success rate of the model. It calculates the proportion of all correct predictions (both positive and negative) out of the total number of predictions.

- TN = True Negative
  → Model correctly predicts the negative class

ISSN: 3065-7865
Bulletin of Nexus, Volume 3| Issue 4 | April 2026
Website: https://bnir.us

- FP = False Positive
  → Model incorrectly predicts positive (Type I error)
- FN = False Negative
  → Model incorrectly predicts negative (Type II error)

These metrics are commonly used in machine learning and medical classification tasks to evaluate model effectiveness.

Experimentation And Result

Experimental Setup

Training was conducted using PyTorch on Google Colab with GPU acceleration. Each model ResNet18 and ResNet50 was trained individually using the curated, balanced dataset.

Both models were trained using CrossEntropyLoss, with checkpointing based on the best validation accuracy for final testing.

Training steps included:

- Loading data with Py Torch Data Loader (4 worker processes, pin_memory=True).
- Forward propagation and loss calculation using CrossEntropyLoss().
- Backpropagation and parameter updates using the Adam optimizer.
- Learning rate scheduling with StepLR (step size = 7, gamma = 0.1).
- ResNet18 was trained without regularization.
- ResNet50 included L2 regularization (weight decay) and early stopping to mitigate overfitting.
- Model checkpointing based on validation accuracy.

The dataset consisted of 20,418 training, 4,498 validation, and 4,532 test images, balanced across first, second, and third-degree burn classes. Extensive data augmentation techniques such as random rotation, horizontal flipping, cropping, color jittering, and normalization were applied to improve generalization.

Both models were fine-tuned using ImageNet-pretrained weights but differed in depth and training strategies. This allowed for a fair comparison to determine the more effective architecture for accurate and scalable burn classification in clinical applications.

Performance Analysis

After training, both models were compared using precision, recall, and F1-score for each burn class, as shown in Table I. TABLE I: Metrics comparison for ResNet18 and ResNet50.

Class	Precision		Recall		F1-score	
	ResNet18	ResNet50	ResNet18	ResNet50	ResNet18	ResNet50
1st Degree	0.93	0.98	0.91	0.97	0.92	0.98
2nd Degree	0.86	0.96	0.90	0.97	0.88	0.96
3rd Degree	0.95	0.99	0.93	0.98	0.94	0.99

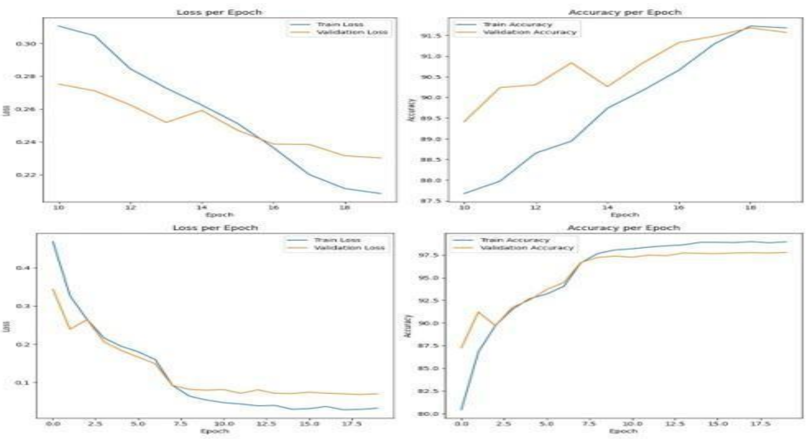

Fig. 2: Training and validation loss and accuracy curves for ResNet18 (top) and ResNet50 (bottom).

Performance Comparison: Table I shows a side-by-side comparison of key classification metrics for ResNet18 and ResNet50. It is clear that for each class, ResNet50 has a better precision, recall, and F1-score than ResNet18. The improved performance across the scores from ResNet50 indicates that ResNet50 has an enhanced ability to differentiate between levels of burn severity and, in particular, second-degree burns where ResNet18 showed relatively poorer performance.Training Curve Analysis: As shown in Figure 2, both models converged successfully.ResNet18 shows consistent training and validation curves with a gap indicating some generalization error. On the other hand, ResNet50 has achieved better results and the possibility of overfitting was quite low thanks to weight decay and early stopping. Overall Summary: ResNet18 provides a strong baseline performance, but ResNet50 significantly improves classification across all burn types. The higher precision, recall, and F1-scores from ResNet50 validate its use in clinical decisionsupport tools, where high accuracy and reliability are essential.

Discussion

ResNet50 outperformed ResNet18 in all evaluation metrics due to its deeper architecture and stronger feature extraction capabilities. While ResNet18 offered faster training, it struggled with subtle burn distinctions, especially in second-degree cases. Regularization techniques and data augmentation helped improve generalization for ResNet50, making it more robust despite dataset variability. However, limitations remain, including the lack of clinical testing and interpretability. Future work should focus on incorporating explainable AI tools (e.g., Grad-CAM) and enhancing dataset diversity to improve model transparency and reliability for realworld deployment.

Ablation Study: ResNet50 vs. EfficientNetB3

To evaluate the impact of architectural choices and training strategies on burn classification performance, we compared our ResNet50 model with the EfficientNetB3-based model presented by Sahlool. The key differences are summarized below:

- Architecture: Our model utilizes ResNet50, known for its deep residual connections, while Sahlool's model employs EfficientNetB3, which balances depth, width, and resolution for efficiency.
- Pretraining: Both models leverage ImageNet-pretrained weights to facilitate transfer learning.
- Regularization: Our ResNet50 model incorporated L2 weight decay (1×10−4) and early stopping to minimize overfitting. In comparison, the EfficientNetB3 model from applied both L1 and L2 regularization in dense layers, along with a dropout rate of 0.45.
- Optimization: We trained our model using the Adam optimizer with a learning rate of 0.0001 and applied a StepLR scheduler with a step size of 7 and a decay factor of 0.1. The approach in used the Adamax optimizer and a learning rate of 0.001.
- Training Duration: Both models were trained for 20 epochs under similar conditions.
- Performance: Our ResNet50 model reached a test accuracy of around 91%, outperforming the EfficientNetB3 model reported in, which achieved 86.9%.

This comparison highlights the effectiveness of deeper architectures like ResNet50, combined with appropriate regularization and optimization strategies, in enhancing classification performance for skin burn severity.

Conclusion

This research focused on applying deep convolutional neural networks, specifically ResNet18 and ResNet50, to automatically classify skin burn images into three severity levels: first, second, and third degree. A carefully balanced dataset ensured fair evaluation across both models under consistent training conditions. ResNet18, used as a baseline, achieved a test accuracy of 91.22%, demonstrating strong performance given its lightweight architecture. However, the deeper ResNet50 significantly outperformed it, reaching a test accuracy of 97.53%. This improvement is attributed to ResNet50's greater depth, enhanced feature extraction capabilities, and the use of regularization techniques such as weight decay and early stopping. The results demonstrate that deep residual networks are highly effective for medical image classification tasks and indicate their suitability for use in clinical decision support applications. Future work can explore real-time deployment, clinical validation, and explainable AI to enhance model transparency and clinical trust. The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgment

The authors express their sincere gratitude to all individuals and organizations who contributed to the successful completion of this research entitled "Skin Burn Classification Using ResNet Models: A Deep Learning Approach Using ResNet18 and ResNet50."

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Chen, G., Zhang, Y., et al. (2020). Skin burn detection and severity classification using convolutional neural networks. IEEE Access, 8, 132409–132417.
2. Kumar, S. L., & Singh, R. (2021). Skin burn detection using pretrained CNN models. International Journal of Computer Vision, 129, 1201–1215.
3. Lee, J., Yoon, M., et al. (2021). Deep learning in medical imaging: A brief overview. Journal of Digital Imaging, 34(3), 547–560.
4. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural

networks. Nature, 542(7639), 115–118.
5. Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
7. Rahman, M., Khan, A., et al. (2022). Automatic skin burn detection using deep convolutional networks. Journal of Medical Image Processing, 6(2), 123–131.
8. Sahlool, F. (2023). Skin burns CNN Acc 86.9%. Kaggle. [Online]. Available: https://www.kaggle.com/code/fathyfathysahlool/skin-burns-cnn-acc-86-9
9. Patel, R., Agarwal, J., et al. (2021). Deep learning for skin burn classification: A comprehensive survey. Journal of Healthcare Engineering, 2021, Article ID 8822391.