

Artículo Original

<https://doi.org/10.22463/0122820X.4276>

Geometric Transformations vs Noise Induction: Comparison of data augmentation techniques for dermoscopic image analysis

Transformaciones geométricas vs Inducción de ruido: Comparación de técnicas de aumentado de datos para análisis de imágenes dermoscópicas

Carlos Vicente Niño-Rondón^{1*}, Manuel Guillermo Forero-Vargas², Sergio Alexander Castro-Casadiago³

¹MSc. (c) en Ciencia de Datos, carlosvicenteninor@javerianacali.edu.co, <https://orcid.org/0000-0002-3781-4564>, Facultad de Ingeniería y Ciencia, Pontificia Universidad Javeriana Cali, Cali, Colombia.

²PhD. en Ingeniería Biomédica, manuel.forero@unibague.edu.co, <https://orcid.org/0000-0001-9972-8621>, Facultad de Ingeniería, Universidad de Ibagué, Ibagué, Colombia.

³MSc. en Ingeniería Electrónica, sergio.castroc@ufps.edu.co, <https://orcid.org/0000-0003-0962-9916>, Facultad de Ingeniería, Universidad Francisco de Paula Santander, Cúcuta, Colombia.

Cómo citar: : C. V. Niño-Rondón, M. G. Forero-Vargas, y S. A. Castro-Casadiago, "Transformaciones geométricas vs Inducción de ruido: Comparación de técnicas de aumentado de datos para análisis de imágenes dermoscópicas", *Respuestas*, vol. 28, n.º 3, pp. 48-57, sep. 2023. <https://doi.org/10.22463/0122820X.4276>

Received on May 16, 2023 - Approved on August 06, 2023.

RESUMEN

Palabras clave:

Conjunto de datos,
Cáncer de piel,
Transformaciones
Geométricas, Ruido
Gaussiano, Análisis de
Componentes Principales.

El conjunto de datos HAM10000, una colección de imágenes dermatoscópicas de lesiones cutáneas, se ha convertido en un recurso valioso para la investigación en dermatología y aprendizaje automático. Este estudio se enfoca en evaluar la eficiencia de dos técnicas de aumento de datos aplicadas a imágenes del conjunto HAM10000 de cáncer de piel. Las técnicas evaluadas en este contexto fueron las transformaciones geométricas e inducción de ruido gaussiano. En la fase metodológica, se implementó la técnica de Análisis de Componentes Principales (PCA) para comparar las imágenes originales con aquellas aumentadas por cada enfoque. Este análisis permitió una comprensión más profunda de las modificaciones introducidas por cada técnica, ofreciendo percepciones sobre la preservación de características relevantes para la clasificación de lesiones cutáneas. Los resultados obtenidos revelaron un rendimiento superior al emplear la técnica de inducción de ruido gaussiano. Esta técnica demostró ser especialmente eficaz en mejorar la calidad del conjunto de datos, contribuyendo positivamente a las tareas de diagnóstico de cáncer de piel. El análisis a través de PCA no solo respaldó la eficacia de la técnica de inducción de ruido gaussiano, sino que también proporcionó una visión detallada de cómo esta técnica conserva información crucial durante el proceso de aumento de datos. Este estudio no solo destaca la relevancia del conjunto de datos HAM10000 en la investigación dermatológica, sino que también enfatiza la importancia de seleccionar técnicas de aumento de datos adecuadas, con la inducción de ruido gaussiano emergiendo como una opción altamente eficiente para mejorar la precisión de modelos de aprendizaje automático aplicados a imágenes médicas en el contexto del cáncer de piel.

ABSTRACT

Keywords:

Dataset,
Skin Cancer,
Geometric
Transformations,
Gaussian Noise,
Principal Components
Analysis

The HAM10000 dataset, a collection of dermoscopic images of skin lesions, has become a valuable resource for research in dermatology and machine learning. This study focuses on evaluating the efficiency of two data augmentation techniques applied to images from the HAM10000 skin cancer dataset. The techniques evaluated in this context were geometric transformations and Gaussian noise induction. In the methodological phase, the Principal Component Analysis (PCA) technique was implemented to compare the original images with those augmented by each approach. This analysis allowed a deeper understanding of the modifications introduced by each technique, offering insights on the preservation of relevant features for skin lesion classification. The results obtained revealed superior performance when employing the Gaussian noise induction technique. This technique proved to be particularly effective in improving the quality of the data set, contributing positively to skin cancer diagnostic tasks. The analysis through PCA not only supported the efficacy of the Gaussian noise induction technique, but also provided detailed insight into how this technique preserves crucial information during the data augmentation process. This study not only highlights the relevance of the HAM10000 dataset in dermatological research, but also emphasizes the importance of selecting appropriate data augmentation techniques, with Gaussian noise induction emerging as a highly efficient option for improving the accuracy of machine learning models applied to medical imaging in the context of skin cancer.

*Corresponding author.

E-mail Address: carlosvicenteninor@javerianacali.edu.co (Carlos Vicente Niño-Rondón)



Peer review is the responsibility of the Universidad Francisco de Paula Santander.
This is an article under the license CC BY-NC 4.0

Introducción

El cáncer de piel es una de las enfermedades que mayor número de casos aporta a las cifras de muerte a nivel mundial. Se estima que el 10 % de los decesos globales son debidos a cáncer de piel o enfermedades cutáneas ligadas [1], [2]. Para detectar estas enfermedades, adicional al diagnóstico de los especialistas en dermatología, se emplean herramientas de diagnóstico asistido por computador, que, respecto a estructuras entrenadas, predicen la naturaleza de la lesión según su apariencia.

El dataset HAM10000 es empleado generalmente para el estudio de enfermedades cutáneas. Este conjunto de datos contiene 10015 imágenes de lesiones cutáneas obtenidas a través de dermatoscopios, que son dispositivos especializados para examinar la piel en detalle [3]. Las imágenes están etiquetadas con información clínica y en su gran mayoría se encuentran validadas por histopatología. El HAM10000 contiene 7 tipos de lesiones cutáneas: queratosis actínicas y carcinoma intraepitelial/Enfermedad de Bowen (akiec), carcinoma de células basales (bcc), lesiones benignas similares a queratosis (bkl), dermatofibromas (df), melanoma (mel), nevus melanocíticos (nv), y lesiones vasculares (vasc). De las 10015 imágenes, 6705 se encuentran etiquetadas como nv, 1113 como mel, 1099 como bkl, 514 como bcc, 327 como akiec, 142 como vasc, y 115 etiquetadas como df. Para contrarrestar el desbalanceo de datos se emplean técnicas de sobremuestreo, con el fin de generar imágenes artificiales en las clases minoritarias, equilibrando con la clase mayoritaria [4].

Las transformaciones geométricas son técnicas aplicadas al procesamiento de imágenes que implican cambios en la geometría o la posición de los elementos visuales en la imagen. Estas transformaciones se utilizan comúnmente en el aumento de datos para imágenes médicas, como en el HAM10000 [5]. Las transformaciones geométricas incluyen operaciones como rotación, traslación, escala y volteo. Estas operaciones modifican la posición, la orientación y el tamaño de la imagen, generando variantes de la imagen original. El objetivo principal de aplicar transformaciones geométricas es mejorar la capacidad del modelo de aprendizaje automático para generalizar y reconocer patrones en diferentes orientaciones, tamaños y posiciones. Asimismo, el aumento de datos por inducción de ruido gaussiano implica la adición de variaciones aleatorias a los píxeles de una imagen siguiendo una distribución normal, lo que significa que las variaciones introducidas son aleatorias y pueden tener valores positivos o negativos [6], [7]. La magnitud de estas variaciones está determinada por la desviación estándar de la distribución gaussiana. El principal objetivo de esta técnica es mejorar la robustez del modelo al exponerlo a variaciones sutiles en la apariencia de las lesiones. Además, puede simular condiciones menos ideales de captura de imágenes, como problemas de iluminación o artefactos presentes en situaciones del mundo real [8].

El Análisis de Componentes Principales (PCA) es una técnica estadística que se utiliza para reducir la dimensionalidad de un conjunto de datos mientras conserva la mayor cantidad posible de su variabilidad [9]. Estos componentes principales pueden interpretarse como patrones representativos de las variaciones presentes en las imágenes [10]. Al seleccionar un subconjunto de las componentes principales, es posible reducir la dimensionalidad de las imágenes, lo que facilita la comparación visual y cuantitativa entre conjuntos [11].

En este documento se presenta la comparación de las técnicas de transformación geométrica y de inducción de ruido gaussiano para la generación de datos artificiales del dataset HAM10000 [12]. La comparación se realiza mediante el estudio de la preservación de la naturaleza de los datos mediante la técnica de análisis

de componentes principales. Además, se hace una comparación de demanda de hardware requerido para el proceso de aumentado de datos. La comparativa se realiza sobre lenguaje de programación Python, y se emplea un computador con procesador AMD Ryzen 7 4700U y 16 GB de RAM instalada.

Materiales y Métodos

Se propone una metodología basada en etapas. En la primera etapa se analiza la naturaleza del conjunto de datos HAM10000, mientras que en la etapa 2 se aplican los enfoques de aumentado de datos, y finalmente, en la tercera etapa, se compara la eficiencia de los enfoques del aumentado de datos.

Naturaleza del HAM10000 Dataset

El conjunto de datos HAM10000 comprende una variedad de imágenes dermoscópicas de lesiones cutáneas, totalizando 10,015 imágenes de alta resolución [3]. Estas imágenes abarcan diversas categorías, incluyendo lesiones benignas y malignas, con metadatos detallados como diagnóstico clínico, diagnóstico histopatológico y ubicación en el cuerpo.

Las imágenes tienen un tamaño general de 600x600 píxeles, aunque algunas pueden ser más grandes. Utilizan la profundidad de color con canales rojo, verde y azul, y muestran características distintivas que reflejan diversas etapas de las lesiones. Pueden contener artefactos como marcas de instrumentos médicos, reflejos de luz y sombras [13].

La versión preliminar del conjunto de datos HAM10000 presenta 7 categorías principales de lesiones cutáneas. La mayor cantidad de imágenes corresponde al Nevus (nv) con 6705 imágenes, seguido de Actinic Keratoses and Intraepithelial Carcinoma (akiec) con 327 imágenes, y dermatofibromas con 115 imágenes. Las lesiones tipo vascular lesions tienen 142 imágenes. Las clases benignas, que incluyen Nevus, representan un total de 7289 imágenes, mientras que las malignas, como melanoma, Basal Cell Carcinoma y Keratosis-like Lesions, suman 2726 imágenes. Destaca que hay más de 4500 imágenes de tipo benigno en comparación con las malignas [14].

Técnicas para el aumentado de datos

Inducción de ruido

Se utiliza un conjunto de datos en formato CSV que describe las imágenes del HAM10000 dataset. Estas imágenes se encuentran en un formato de píxeles 28x28 y presentan información en los canales de color rojo, verde y azul (RGB). El conjunto de datos consta de 2352 columnas que representan la intensidad de color en ubicaciones específicas para cada una de las 10,015 imágenes, todas en un tamaño de 28x28 píxeles [15]. Además, hay una columna adicional que contiene las etiquetas representando las clases a las que pertenecen las imágenes.

El sobremuestreo se lleva a cabo a través de la adición de una variable aleatoria, representada por Z , que sigue una distribución normal con media cero y desviación estándar σ [16]. Así, la expresión del aumento de datos se formula como $Y=X+N(0,\sigma^2)$, donde X es una variable aleatoria que representa un píxel en la imagen original y Y es la variable aleatoria correspondiente al píxel aumentado.

Desde la perspectiva del valor esperado, $E[Y]$ se simplifica a $E[X]$, lo que significa que el aumento de datos no tiene un impacto significativo en la expectativa del píxel, preservando la información esencial de la clase original. Sin embargo, la variabilidad se incrementa debido al ruido introducido [17]. El propósito de esta estrategia es generar sutiles fluctuaciones que puedan manifestarse en diversas instancias de la misma clase, mejorando la capacidad del modelo para generalizar a nuevos datos y reduciendo el riesgo de sobreajuste.

Transformaciones geométricas

La transformación geométrica para el aumento de imágenes en el conjunto de datos HAM10000 implica la aplicación de operaciones como rotación, traslación, escalado, inversión horizontal o vertical, y cizallamiento de manera aleatoria a cada imagen durante el entrenamiento del modelo. Estas operaciones modifican la geometría de las imágenes originales, generando variaciones que enriquecen el conjunto de datos. La rotación simula diferentes orientaciones de las lesiones, la traslación introduce variabilidad en la posición, el escalado adapta el modelo a diferentes tamaños de lesiones, la inversión simula cambios en la orientación, y el cizallamiento distorsiona líneas y formas. Estas transformaciones buscan enriquecer el conjunto de datos sin alterar esencialmente la información de las lesiones, permitiendo que el modelo aprenda patrones más robustos y mejore su capacidad de generalización a nuevas instancias de imágenes cutáneas. Para la rotación, se emplea un ángulo aleatorio entre -10 y 10 grados, mientras que la traslación se realiza con desplazamientos aleatorios en ambas direcciones hasta 10 píxeles. En cuanto al escalado, se aplica un factor aleatorio entre 0.8 y 1.2. La inversión horizontal se ejecuta con una probabilidad del 50%, y el cizallamiento se realiza con un factor aleatorio entre -0.1 y 0.1.

Comparación de los enfoques de aumentado de datos

La comparación entre las técnicas de aumento de datos se lleva a cabo mediante el análisis de Componentes Principales (PCA) [18], una técnica de reducción de dimensionalidad. El objetivo es evaluar la capacidad de cada método para preservar la esencia y la variabilidad inherentes a las imágenes originales del conjunto de datos HAM10000 [19]. El enfoque consiste en aplicar PCA a las imágenes antes y después de la aplicación de las técnicas de aumento, permitiendo así cuantificar cómo las transformaciones geométricas afectan la estructura global de las imágenes. Durante el proceso, se extraen los componentes principales, que representan las direcciones de máxima variabilidad en el conjunto de datos [20]. La comparación se basa en la retención de varianza explicada por estos componentes antes y después del aumento, proporcionando una medida cuantitativa de la capacidad de cada técnica para conservar la información esencial de las lesiones cutáneas. Este enfoque permite evaluar la calidad del aumento de datos en términos de preservación de la variabilidad estructural, lo que es crucial para el entrenamiento eficaz de modelos de aprendizaje automático. Adicionalmente, se pondera el tiempo de ejecución de la tarea, así como la cuantificación de recursos de hardware a nivel de RAM y CPU utilizada en el proceso de aumentado de datos [21].

Resultados y Discusión

En la figura 1 se muestran imágenes originales del conjunto de datos HAM10000 en resolución de 28x28 píxeles. Se observan imágenes con pigmentación marcada en la zona de lesión, y tonalidades más claras a los costados de la lesión, reflejando la sección de piel capturada en la imagen.

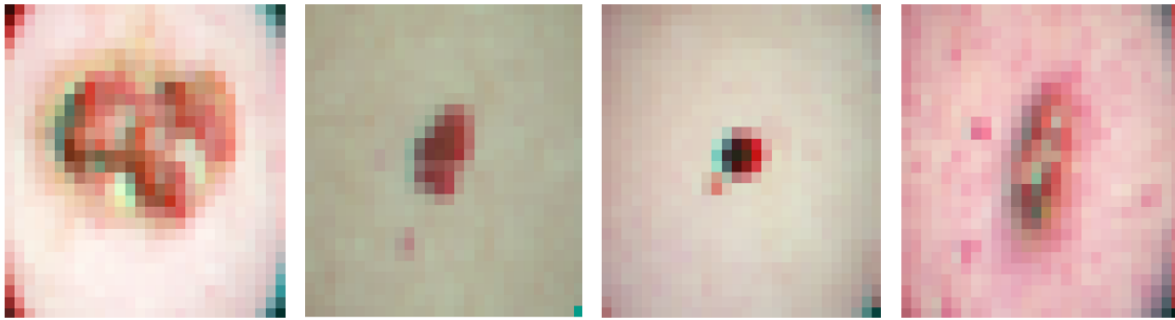


Figura 1. Imágenes originales del HAM10000 Dataset en resolución de 28x28

De igual forma, en la Tabla I se muestra la información de los píxeles para las primeras 5 imágenes del HAM10000 dataset. La información de los píxeles se refiere a los primeros cinco descriptores y a los últimos cinco descriptores.

Tabla I. Descriptores de los píxeles de las primeras 5 imágenes del HAM10000 Dataset.

Imagen	Píxeles									
	0001	0002	0003	0004	0005	2348	2349	2350	2351	2352
00001	192	153	193	195	155	147	166	185	154	177
00002	25	14	30	68	48	14	28	25	14	27
00003	192	138	153	200	145	124	142	136	104	117
00004	38	19	30	95	59	12	17	25	12	15
00005	158	113	139	194	144	135	149	109	78	92

Asimismo, en la figura 2 se muestran imágenes aumentadas mediante la aplicación de la técnica de inducción de ruido. Como se logra percibir, se preserva en la zona central de las imágenes los píxeles que demarcan la lesión mediante una pigmentación más oscura, mientras que, en los costados, se preserva los píxeles con tonalidades claras referente a la piel. De esta forma se infiere que, a nivel visual, se preserva la naturaleza de las imágenes originales del HAM10000 dataset.

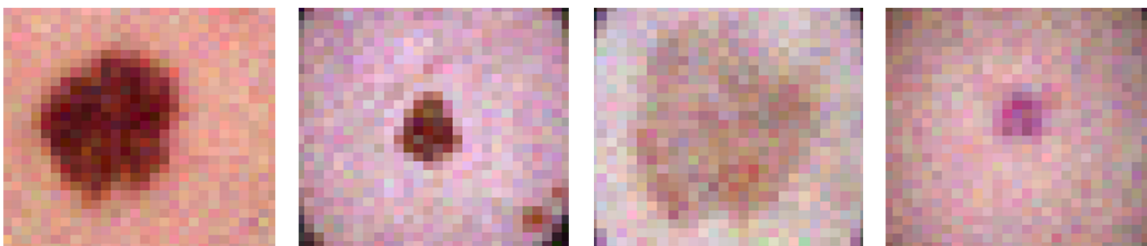


Figura 2. Imágenes obtenidas por aumentado de datos mediante inducción de ruido.

Asimismo, en la figura 3 se muestran algunas imágenes aumentadas obtenidas mediante transformaciones geométricas de las imágenes. Si bien se observan imágenes más claras, en términos generales no se presenta una forma clara en las lesiones, y se presenta una gran variabilidad a nivel visual en la pigmentación de la lesión cutánea.

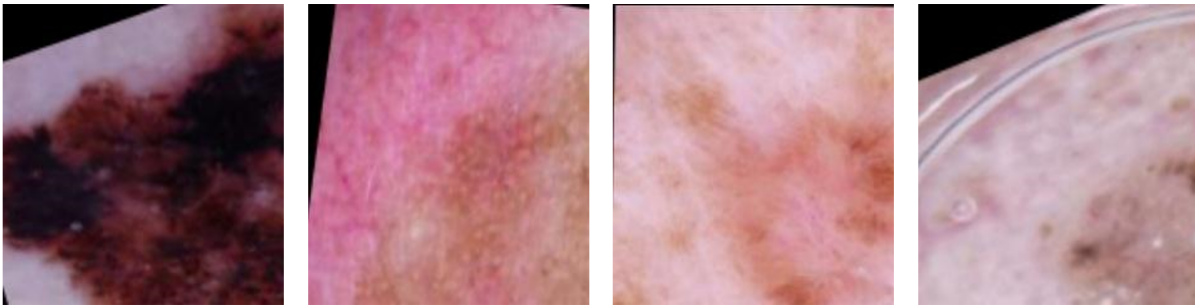


Figura 3. Imágenes obtenidas por aumento de datos mediante transformaciones geométricas.

Para validar y comparar los enfoques de aumento de datos se aplica PCA. En la figura 4a se muestra la distribución de características de componentes para el conjunto de imágenes originales. Asimismo, en la sección 4b se muestra la distribución para el conjunto de datos aumentados mediante inducción de ruido gaussiano, mientras que en la sección C se presenta la distribución para las imágenes aumentadas mediante transformaciones geométricas. Como se observa, las imágenes aumentadas por inducción de ruido (amarillo) presenta una similitud mayor en la distribución de la naturaleza del conjunto de datos respecto a las imágenes del conjunto de datos original (verde). Además, se preserva la cantidad de información entre estos dos datasets relacionados.

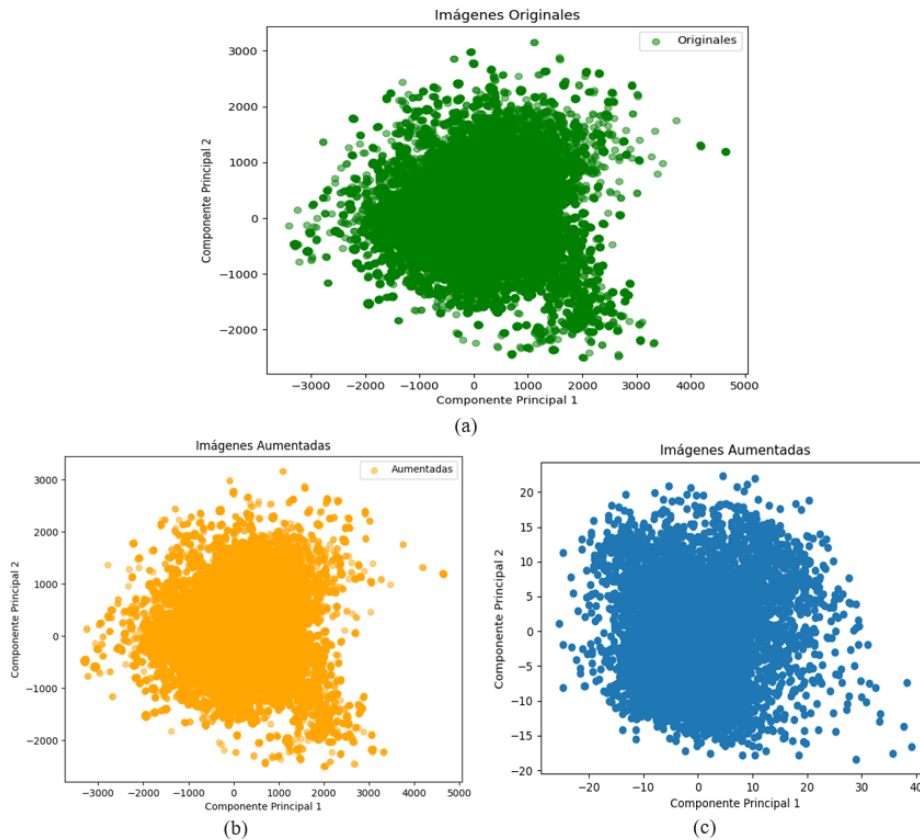


Figura 4. Distribuciones para los conjuntos de datos. (a) imágenes originales, (b) imágenes aumentadas por inducción de ruido, (c) imágenes aumentadas por transformaciones geométricas.

Asimismo, en la figura 5 se presenta la comparación entre estas técnicas respecto al tiempo de ejecución en (a), CPU utilizada (b) y RAM utilizada (c). La técnica de inducción de ruido muestra consistentemente tiempos de ejecución más reducidos en comparación con las transformaciones geométricas en todas las fases del proceso de desarrollo. Aunque la dispersión de los tiempos de ejecución puede variar en diferentes etapas, en promedio, la técnica de inducción de ruido exhibe una mayor eficiencia temporal. En cuanto al uso de la RAM, se observa que la inducción de ruido generalmente requiere un menor porcentaje de memoria en comparación con las transformaciones geométricas. La variabilidad en el uso de la RAM puede depender de la complejidad de las operaciones, pero en líneas generales, la inducción de ruido parece gestionar la memoria de manera más eficiente. Similarmente, en relación con el uso de la CPU, la técnica de inducción de ruido tiende a emplear menos recursos de procesamiento en comparación con las transformaciones geométricas. Aunque la variabilidad en el uso de la CPU puede depender de la carga computacional específica, en términos generales, la inducción de ruido parece tener un impacto menor en la CPU.

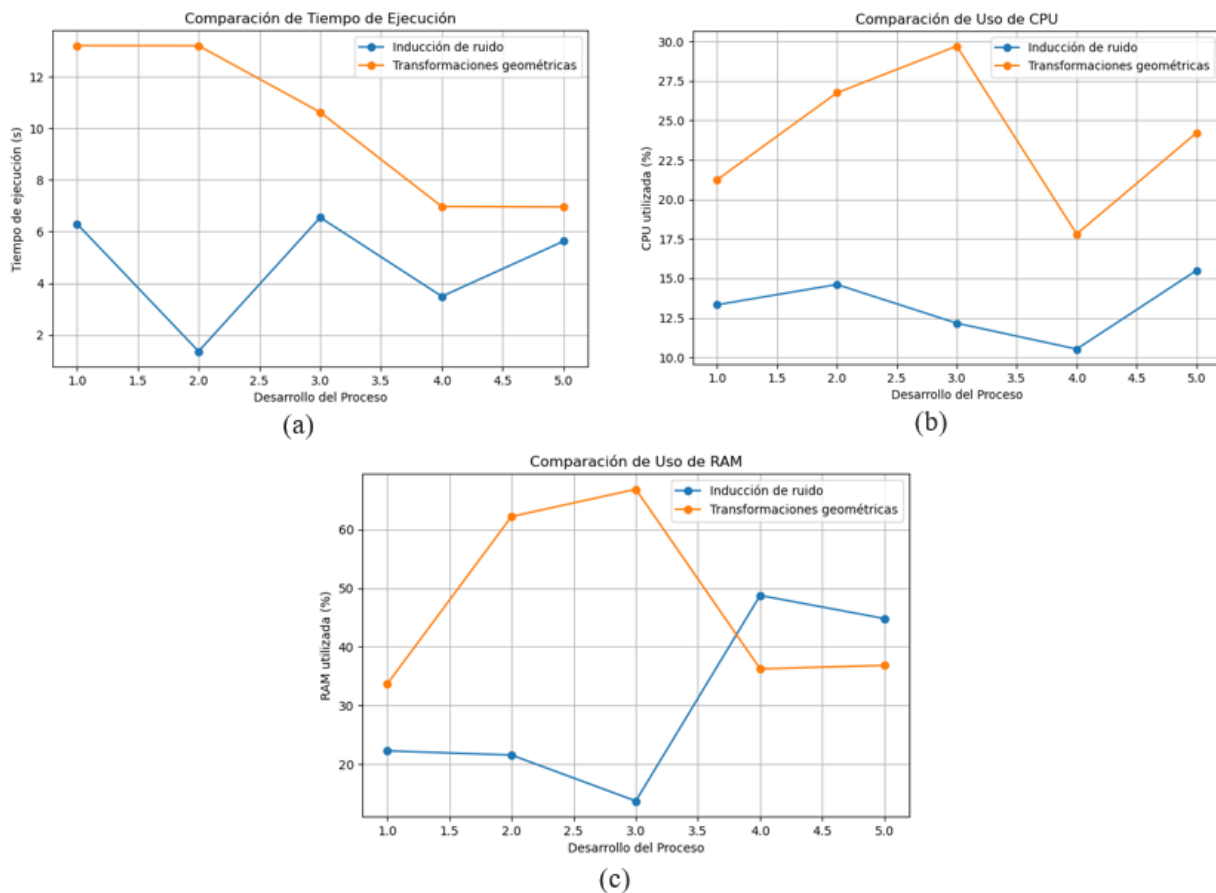


Figura 5. Comparación de las técnicas de aumentado de datos. (a) tiempo de ejecución, (b) CPU empleada, (c) RAM utilizada.

Conclusiones

Se puede inferir que la técnica de inducción de ruido aplicada al conjunto de datos HAM10000 demuestra ser efectiva en preservar la naturaleza de las imágenes originales. Visualmente, las imágenes aumentadas mediante inducción de ruido conservan las características esenciales de las lesiones cutáneas, manteniendo la pigmentación central y las tonalidades claras en los costados. En contraste, el aumento de datos mediante

transformaciones geométricas muestra una variabilidad visual significativa en la pigmentación de las lesiones y no presenta una forma clara.

El análisis de componentes principales (PCA) respalda visualmente la calidad de las imágenes aumentadas por inducción de ruido, mostrando una distribución más similar a la de las imágenes originales y preservando la información del conjunto de datos. Además, en términos de eficiencia computacional, la técnica de inducción de ruido exhibe tiempos de ejecución consistentemente más bajos, un menor uso de la RAM y una menor carga en la CPU en comparación con las transformaciones geométricas. Estos hallazgos sugieren que la inducción de ruido no solo preserva la calidad visual de las imágenes, sino que también presenta ventajas en términos de eficiencia y recursos computacionales.

En términos de eficiencia computacional, la inducción de ruido muestra consistentemente tiempos de ejecución más bajos, una menor utilización de la RAM y una carga de CPU reducida en comparación con las transformaciones geométricas. Este rendimiento más eficiente sugiere que la técnica de inducción de ruido no solo es visualmente superior, sino que también presenta ventajas prácticas al requerir menos recursos durante el proceso de aumento de datos. En resumen, la elección de la inducción de ruido como estrategia de aumento de datos en el contexto del conjunto HAM10000 ofrece beneficios tanto en calidad visual como en eficiencia computacional.

Referencias

- [1] N. Razmjoo et al., “Computer-aided Diagnosis of Skin Cancer: A Review,” *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 16, no. 7, pp. 781–793, Jan. 2020, doi: 10.2174/1573405616666200129095242.
- [2] D. N. Dorrell and L. C. Strowd, “Skin Cancer Detection Technology,” *Dermatol Clin*, vol. 37, no. 4, pp. 527–536, 2019, doi: 10.1016/j.det.2019.05.010.
- [3] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, pp. 1–9, 2018, doi: 10.1038/sdata.2018.161.
- [4] A. S. Jahn et al., “Over-Detection of Melanoma-Suspect Lesions by a CE-Certified Smartphone App: Performance in Comparison to Dermatologists, 2D and 3D Convolutional Neural Networks in a Prospective Data Set of 1204 Pigmented Skin Lesions Involving Patients’ Perception,” *Cancers (Basel)*, vol. 14, no. 15, Aug. 2022, doi: 10.3390/cancers14153829.
- [5] T. Saba, M. A. Khan, A. Rehman, and S. L. Marie-Sainte, “Region Extraction and Classification of Skin Cancer: A Heterogeneous framework of Deep CNN Features Fusion and Reduction,” *Journal of Medical Systems* 2019 43:9, vol. 43, no. 9, pp. 1–19, Jul. 2019, doi: 10.1007/S10916-019-1413-3.
- [6] H.-W. Huang, B. W.-Y. Hsu, C.-H. Lee, and V. S. Tseng, “Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers,” *J Dermatol*, vol. 48, no. 3, pp. 310–316, Mar. 2021, doi: 10.1111/1346-8138.15683.

- [7] A. Kumar and A. Vatsa, "Untangling Classification Methods for Melanoma Skin Cancer," *Front Big Data*, vol. 5, no. March, pp. 1–11, 2022, doi: 10.3389/fdata.2022.848614.
- [8] N. Hameed, A. Shabut, and M. A. Hossain, "A Computer-Aided diagnosis system for classifying prominent skin lesions using machine learning," in *2018 10th Computer Science and Electronic Engineering Conference, CEEC 2018 - Proceedings*, IEEE, 2019, pp. 186–191. doi: 10.1109/CEEC.2018.8674183.
- [9] R. Baig, M. Bibi, A. Hamid, S. Kausar, and S. Khalid, "Deep Learning Approaches Towards Skin Lesion Segmentation and Classification from Dermoscopic Images - A Review," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 16, no. 5, pp. 513–533, Jan. 2019, doi: 10.2174/1573405615666190129120449.
- [10] A. Gupta and A. Barbu, "Parameterized principal component analysis," *Pattern Recognit*, vol. 78, pp. 215–227, 2018, doi: 10.1016/j.patcog.2018.01.018.
- [11] A. K. Seghouane, N. Shokouhi, and I. Koch, "Sparse Principal Component Analysis with Preserved Sparsity Pattern," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3274–3285, Jul. 2019, doi: 10.1109/TIP.2019.2895464.
- [12] S. Wang and M. Hamian, "Skin Cancer Detection Based on Extreme Learning Machine and a Developed Version of Thermal Exchange Optimization," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/9528664.
- [13] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," *Healthcare (Switzerland)*, vol. 10, no. 7, Jul. 2022, doi: 10.3390/healthcare10071183.
- [14] G. Campanella et al., "Deep Learning for Basal Cell Carcinoma Detection for Reflectance Confocal Microscopy," *Journal of Investigative Dermatology*, vol. 142, no. 1, pp. 97–103, Jan. 2022, doi: 10.1016/J.JID.2021.06.015.
- [15] V. Nyemeesha and B. M. Ismail, "Implementation of noise and hair removals from dermoscopy images using hybrid Gaussian filter," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 10, no. 1, Dec. 2021, doi: 10.1007/s13721-021-00318-2.
- [16] O. O. Abayomi-Alli, R. Damaševičius, S. Misra, R. Maskeliūnas, and A. Abayomi-Alli, "Malignant skin melanoma detection using image augmentation by oversampling in nonlinear lower-dimensional embedding manifold," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 8, pp. 2600–2614, 2021, doi: 10.3906/elk-2101-133.
- [17] J. B. Li and H. Gao, "Sparse data-dependent kernel principal component analysis based on least squares support vector machine for feature extraction and recognition," *Neural Comput Appl*, vol. 21, no. 8, pp. 1971–1980, 2012, doi: 10.1007/s00521-011-0600-z.
- [18] A. Lasisi and N. Attoh-Okine, "Principal components analysis and track quality index: A machine

learning approach,” *Transp Res Part C Emerg Technol*, vol. 91, no. April 2018, pp. 230–248, 2018, doi: 10.1016/j.trc.2018.04.001.

- [19] Y. Zhu, C. Zhu, and X. Li, “Improved principal component analysis and linear regression classification for face recognition,” *Signal Processing*, vol. 145, pp. 175–182, 2018, doi: 10.1016/j.sigpro.2017.11.018.
- [20] N. Bakshi and V. Prabhu, “Face recognition system for access control using principal component analysis,” in *ICCT 2017 - International Conference on Intelligent Communication and Computational Techniques*, 2018, pp. 145–150. doi: 10.1109/INTELCCT.2017.8324035.
- [21] R. Cheripelli and K. R. Sri, “Evaluation of machine Learning Models for Credit Scoring,” *Ingeniería Solidaria*, vol. 16, no. 2798, pp. 2798–2805, 2020.