IMPACTO DE *DATASETS* EXTREMADAMENTE DESBALANCEADOS EN MODELOS DE IA PARA LA SALUD: ANÁLISIS DESDE EL MONITOREO DE PERSONAS MAYORES

Amanda Milena Santacruz-Madroñero y Leonardo Betancur Agudelo
Universidad Pontificia Bolivariana - UPB-Medellin, Colombia

RESUMEN

Este trabajo explora el impacto de datasets desbalanceados en modelos de inteligencia artificial (IA) para monitorear la salud de personas mayores, una población vulnerable con alto riesgo de condiciones subdiagnosticadas. El desbalance de clases, donde las enfermedades graves están subrepresentadas, reduce la precisión y dificulta la detección de condiciones críticas. Se analizan técnicas como sobremuestreo, submuestreo, enfoques híbridos y deep learning, evaluando sus beneficios y limitaciones. Aunque estas técnicas mejoran la identificación de patrones en clases minoritarias, enfrentan desafíos como el sobreajuste y los altos costos computacionales, lo que dificulta su implementación en entornos clínicos. El trabajo enfatiza la importancia de desarrollar algoritmos robustos y accesibles para aplicar IA de manera efectiva en esta población creciente y vulnerable.

PALABRAS CLAVE

Muestreo Mal Condicionado, Datos Extremadamente Desbalanceados, Datos Sintéticos, Inteligencia Artificial

1. INTRODUCCIÓN

La Inteligencia Artificial (IA) es fundatmental en el sector salud, especialmente para monitorear a personas mayores, una población vulnerable con alta prevalencia de enfermedades crónicas y riesgo de condiciones subdiagnósticadas, según Taha et al (2023). Sin embargo, los modelos de IA, acorde con Haixiang et al (2017), enfrentan el desafío de datasets desbalanceados, donde las instancias de condiciones graves son significativamente menores que las de individuos sanos. Este desbalanceo afecta la capacidad de los modelos para identificar patrones en clases minoritarias, reduciendo su precisión y efectividad en la detección de enfermedades críticas, esenciales para la calidad de vida de los pacientes. Este documento analiza el impacto de los datasets desbalanceados en modelos de IA para el monitoreo de salud en personas mayores. Se revisan técnicas como el sobremuestreo y submuestreo, evaluando sus beneficios y limitaciones, y se sugieren mejoras para optimizar la precisión y generalización en entornos clínicos.

2. MATERIALES Y MÉTODOS

Se presenta la Tabla 1 que resume las técnicas para manejar datos desbalanceados en aplicaciones médicas, destacando características, ventajas, desafíos y algunas aplicaciones en el contexto clínico. Aquí es importante tener en cuenta que seleccionar la estrategia adecuada depende de los datos, necesidades del problema y recursos, fundamentales en contextos médicos.

Tabla 1. Técnicas para el manejo de datos desbalanceados

Técnica	Descripción	Ventajas	Desafíos	Aplicaciones en Salud
Sobremuestre SMOTE Chawla et al (2002), ADASYN He (2008).	Generación de datos sintéticos para la clase minoritaria mediante interpolación y adaptación según la dificultad	- Mejora el rendimiento en casos de desbalanceo extremo. - Aumenta la representatividad de la clase minoritaria.	- Susceptible al ruido. - Requiere supervisión cuidadosa en contextos médicos, según Satyendra y Amit (2022).	Detección de enfermedades raras o diagnósticos con datos limitados.
Submuestreo	Reducción de la clase mayoritaria eliminando instancias redundantes.	- Reduce el riesgo de sobreajuste. - Simplifica los modelos en datasets grandes.	- Pérdida de información valiosa. - Riesgo de afectar la precisión en diagnósticos críticos, según Satyendra y Amit (2022).	Equilibrio en análisis de cohortes médicas amplias.
Técnicas Híbridas (SMOTE-ENN, SMOT Tomek Links según Jolouda et al (2023)	Combinan sobremuestreo y submuestreo, eliminando ruido y mejorando la generalización del modelo, según Gholampour (2024).	- Balance efectivo de clases. - Mejora la calidad de los datos. - Reduce errores cercanos a la frontera.	- Ajuste complejo. - Sensibles a configuraciones inadecuadas en datos clínicos.	Diagnósticos multiclase con datos de calidad variable.
Modelos de Coste Sensibl	Penalizan errores en la clasificación de la clase minoritaria.	- Aumenta la sensibilidad. - Útil para reducir falsos negativos.	- Implementación compleja. - Riesgo de sobreajuste si no se calibra correctamente.	Detección de enfermedades críticas como cáncer o fallos cardíacos.
Deep Learnin (CNN, GAN)	Uso de redes neuronales para crear instancias sintéticas avanzadas de la clase minoritaria.	- Capacidad para modelar patrones complejos. - Resultados de alta calidad en contextos bien definidos.	- Alto costo computacional. - Requiere grandes volúmenes de datos. - Limitaciones en recursos, según Lee y Lee (2023)	Diagnósticos basados en imágenes médicas y señales fisiológicas

3. RESULTADOS

3.1 Aplicando Técnicas de Sobremuestreo

El sobremuestreo es ampliamente utilizado para resolver el desbalanceo de clases, según Yang et al (2024) en datasets médicos, siendo, según Li et al (2023), SMOTE la técnica más común. Genera instancias sintéticas de la clase minoritaria, mejora la precisión en casos extremos, pero enfrenta limitaciones en condiciones graves para personas mayores, como insuficiencia cardíaca o deterioro cognitivo. Variantes como KNSMOTE, combina SMOTE con clustering k-means para reducir ruido, y técnicas como WTASUWO y OREM según Zhu et al (2023), optimizan la precisión creando instancias adaptativas en áreas subrepresentadas. Si bien el sobremuestreo mejora la detección de enfermedades críticas en personas mayores, persiste el desafío de equilibrar generación de datos y evitar sobreajuste o ruido.

3.2 Con Técnicas Híbridas

Las técnicas híbridas mejoran la precisión de modelos para monitorear condiciones crónicas en personas mayores. Acorde con Nizam y Orman(2024), GASMOTEPSO_ENN combina SMOTE, algoritmos heurísticos (GA), optimización por enjambre de partículas (PSO) y submuestreo (ENN) según Zhao y Li (2020), balanceando datos y mejorando resultados. Según Santander (2023), Deep Learning Jerárquico Híbrido, aplicado con Internet de las Cosas Médicas (IoMT), predice condiciones graves con alta precisión.

Según Zeng et al (2016), SMOTE con Tomek Links genera datos sintéticos y elimina ruido, siendo útil para diabetes o insuficiencia cardíaca, mientras ROS-CNN, según Ma et al (2023), combina sobremuestreo aleatorio y redes neuronales convolucionales para clasificar condiciones como sarcopenia. Estos métodos, enfrentan desafíos de implementación, como evitar el sobreajuste y balancear falsos negativos y positivos, fundamentales en salud. Un falso negativo puede pasar una enfermedad grave desapercibida, mientras un falso positivo podría derivar en intervenciones innecesarias.

3.3 Con Modelos de Costo Sensible

Los modelos de costo-sensible penalizan más los errores en la clase minoritaria, importante en salud donde un falso negativo puede tener consecuencias graves. según Mienye y Sun (2021), Cost-sensitive XGBoost ajusta costos para enfermedades críticas como cáncer de mama, mientras Cost-sensitive Random Forest según Yang et al (2009), se utiliza en enfermedades cardíacas y pulmonares en personas mayores. según Pérez y Gutiérrez (2023), GentleBoost y RankCost según Wan et al (2014), son otros modelos aplicados en enfermedades crónicas y monitoreo de diabetes. Sin embargo, ajustar los costos de clasificación es un desafío, debido a que puede generar sobreajuste y afectar la generalización. Equilibrar falsos positivos y negativos sigue siendo un problema crítico, especialmente en personas mayores.

3.4 Con Modelos de *Deep Learning*

Según Xu et al (2019), las redes neuronales convolucionales (CNN) y LSTM (Long Short-Term Memory), mejoran la clasificación médica al generar instancias sintéticas más realistas para clases minoritarias. En el monitoreo de personas mayores, estas técnicas identifican patrones que los modelos tradicionales no detectan, mejorando el diagnóstico de enfermedades emergentes y raras. Ejemplos incluyen el uso de CNN y LSTM para predecir comportamientos anormales mediante datos de sensores en hogares inteligentes y según Almutairi et al (2022), Autoencoder-CNN-LSTM para mejorar predicciones sobre movilidad y comportamiento. Según Tufail et al (2024), los Modelos Generativos Adversarios (GANs) también mitigan el desbalanceo creando instancias sintéticas. Sin embargo, estas técnicas requieren grandes volúmenes de datos y alta capacidad computacional, limitando su aplicación en entornos clínicos con recursos reducidos. Además, la falta de interpretabilidad dificulta su adopción por profesionales médicos.

4. DISCUSIÓN

Las técnicas de IA para gestionar datasets desbalanceados en el monitoreo de la salud de personas mayores ofrecen avances significativos, pero también presentan desafíos no resueltos. Estrategias de sobremuestreo, como las variantes de SMOTE, mejoran la precisión al detectar condiciones críticas, aunque pueden introducir ruido y sobreajuste, que afecta la generalización. Los métodos híbridos, que combinan sobremuestreo y submuestreo, balancean mejor los datos, pero necesitan ajustes complejos y altos costos computacionales, lo que complica su aplicación en clínicas con recursos limitados. Los modelos de costo-sensible son buenos para reducir falsos negativos en enfermedades críticas, pero tienden a generar falsos positivos, que incrementa las intervenciones innecesarias y sobrecarga el sistema de salud. Por otro lado, las técnicas de deep learning permiten identificar patrones complejos, pero su aplicabilidad está restringida por altos requisitos computacionales y falta de interpretabilidad, que dificulta su adopción por profesionales médicos. Este campo sigue abierto a la investigación, demandando soluciones prácticas que beneficien a la creciente población mayor, vulnerable por enfermedades crónicas y subdiagnosticadas.

5. CONCLUSIONES

Las técnicas de IA han avanzado en el manejo de datasets desbalanceados, pero enfrentan retos en la práctica clínica. Métodos como SMOTE mejoran la detección de condiciones críticas, mientras los híbridos logran balances efectivos, aunque ambos enfrentan problemas de ruido y altos costos computacionales. Los modelos de costo-sensible reducen falsos negativos, esenciales en salud, pero deben calibrarse para evitar falsos positivos. Las técnicas de deep learning, en tendencia, son inaccesibles en entornos con recursos limitados.

Por lo tanto, es necesario desarrollar algoritmos robustos que equilibren precisión y generalización, especialmente para monitorear a una población vulnerable como las personas mayores.

REFERENCIAS

Almutairi, M., Gabralla, L. A., Abubakar, S., & Chiroma, H. (2022). Detecting elderly behaviors based on deep learning for healthcare: Recent advances, methods, real-world applications and challenges. *IEEE Access*, 10, 69802-69821.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Gholampour, Seifollah. (2024). Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable. *Machine Learning and Knowledge Extraction*, 6(2), 827-841.

Gutiérrez, D., Villa, W. M., & López-Lezama, J. M. (2017). Flujo óptimo reactivo mediante optimización por enjambre de partículas. *Información tecnológica*, 28(5), 215-224.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220-239.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *2008 IEEE International Joint Conference on Neural Networks (IJCNN)* (pp. 1322-1328). IEEE.

Joloudari, Javad Hassannataj, Abdolreza Marefat, Mohammad Ali Nematollahi, Solomon Sunday Oyelere, & Sadiq Hussain. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, 13(6), 4006.

Lee, Ji-Na, & Ji-Yeoun Lee. (2023). An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Applied Sciences*, 13(6), 3571.

Li, Yue, Qingyu Hu, Guilan Xie, & Gong Chen. (2023). Prediction of the Health Status of Older Adults Using Oversampling and Neural Network. *Mathematics*, 11(24), 4985.

Ma, W., Gou, C., & Hou, Y. (2023). Research on adaptive 1DCNN network intrusion detection technology based on BSGM mixed sampling. *Sensors*, 23(13), 6206.

Mienye, Domor, & Sun, Yanxia. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25, 10.1016/j.imu.2021.100690.

Nizam-Ozogur, H., & Orman, Z. (2024). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, e13596.

Pérez-Velasco, S., & Gutiérrez-Tobal, G. C. (2023). Assessment of Residual Deep Neural Networks and AdaBoost to predict adherence to digital-based active and healthy aging interventions.

Santander Baños, F. (2023). Optimización de hiperparámetros de una red neuronal convolucional a través de un algoritmo metaheurístico para mejorar la clasificación de arritmias cardíacas.

Satyendra Singh Rawat & Amit Kumar Mishra. (2022). Review of Methods for Handling Class-Imbalanced in Classification Problems. *Eprint: 2211.05456.*

Taha Shiwani, Samuel Relton, Ruth Evans, Aditya Kale, Anne Heaven, & Andrew Clegg. (2023). Ageing Data Research Collaborative (Geridata). Horizons in Artificial Intelligence in the Healthcare of Older People. *Age and Ageing*, 52(12), afad219.

Tufail, H., Ahad, A., Puspitasari, I., Shayea, I., Coelho, P. J., & Pires, I. M. (2024). Deep Learning in Smart Healthcare: A GAN-based Approach for Imbalanced Alzheimer's Disease Classification. *Procedia Computer Science*, 241, 146-153.

Wan, X., Liu, J., Cheung, W. K., & Tong, T. (2014). Learning to improve medical decision making from imbalanced data without a priori cost. *BMC Medical Informatics and Decision Making*, 14(1), 1-9.

Xu, J., He, Z., & Zhang, Y. (2019). CNN-LSTM combined network for IoT enabled fall detection applications. In *Journal of Physics: Conference Series* (Vol. 1267, No. 1, p. 012044).

Yang, F., Wang, H. Z., Mi, H., Lin, C. D., & Cai, W. W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, 10(1), 1-14.

Yang, Yuxuan, Akbarzadeh, Hadi, & Aickelin, Uwe. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. *Health Informatics*, Volume 6 – 2024.

Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)* (pp. 225-228).

Zhao, S., & Li, J. (2020). ELS: a fast parameter-free edition algorithm with natural neighbors-based local sets for k nearest neighbor. *IEEE Access*, 8, 123773-123782.

Zhu, T., Liu, X., & Zhu, E. (2023). Oversampling With Reliably Expanding Minority Class Regions for Imbalanced Data Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(6), 6167-6181.