

Identifying Architectural Properties that Predict Model Performance in AutoML NILM Frameworks

Manuel Pereiro¹ , Bautista Roccella Elizalde¹ , and Nicolas A. Monzon^{1,2} 

¹ Universidad Argentina de la Empresa, CABA, Argentina
broccella@uade.edu.ar, manpereiro@uade.edu.ar, nimonzon@uade.edu.ar

² Universidad de la República, PEDECIBA-Informática, Montevideo, Uruguay

Abstract. Non-Intrusive Load Monitoring (NILM) enables the disaggregation of household energy consumption into individual appliances, offering a non-invasive path toward improved energy management. Recent work has demonstrated the potential of Automated Machine Learning (AutoML) frameworks leveraging Bayesian Optimization to automate model selection and hyperparameter tuning for NILM tasks. Still, a fundamental question remains open: given the growing number of candidate models, how can a researcher determine a priori which architectural choices are likely to perform well within such a framework, without running the full experimental pipeline? This work extends the paper of Bayesian Optimization-based AutoML for NILM by systematically investigating architectural properties of NILM models that correlate with strong performance. By closely examining the results from different models, we have identified specific, verifiable characteristics in a model's design (elements observable in its structure) that indicate whether it will be suitable for our framework. We further evaluate the generalizability of these findings on real industrial smart meter data from Argentine electric cooperatives (FACE). We aim to provide principled guidance for model selection in AutoML NILM pipelines.

Keywords: non-intrusive load monitoring, automated machine learning, bayesian optimization, machine learning pipeline, energy disaggregation.

Identificación de propiedades arquitectónicas que predicen el rendimiento de los modelos en marcos AutoML NILM

Abstract. El Monitoreo de Cargas No Intrusivo (NILM) permite la desagregación del consumo eléctrico agregado de un hogar en la contribución individual de cada artefacto al consumo total. Algunos trabajos recientes demostraron el potencial de frameworks de Aprendizaje automático automatizado (AutoML) con Optimización Bayesiana para automatizar la selección de modelos en tareas NILM. Sin embargo, todavía permanece abierta una pregunta fundamental: ¿Cómo puede un investigador determinar inicialmente qué propiedades arquitectónicas predicen un

buen rendimiento dentro de dicho framework, sin ejecutar el pipeline experimental completo? El presente trabajo extiende la investigación sobre AutoML para NILM mediante el estudio sistemático de propiedades arquitectónicas que correlacionan con un alto rendimiento. Tras examinar detenidamente los resultados de los distintos modelos, hemos identificado características específicas y verificables en el diseño de un modelo (elementos observables en su estructura) que indican si será adecuado para nuestro marco. Los hallazgos se validan sobre datos reales de cooperativas eléctricas argentinas (FACE). El resultado esperado es un conjunto de condiciones que permita orientar la selección de modelos con fundamento teórico previo a la experimentación práctica.

Keywords: monitoreo de cargas no intrusivo, aprendizaje automático automatizado, optimización bayesiana, flujo de trabajo de aprendizaje automático, desagregación energética.

1 Introducción

El Monitoreo de Cargas No Intrusivo (NILM) es el problema de desagregar el consumo eléctrico agregado de un hogar en las contribuciones individuales de cada electrodoméstico, utilizando únicamente mediciones tomadas en un único punto del circuito (Hart, 1992). A diferencia de los enfoques intrusivos, NILM ofrece una vía escalable y de bajo costo hacia el monitoreo energético detallado, con aplicaciones directas en respuesta a la demanda, auditoría de eficiencia y programas de cambio según la conducta (Angelis et al., 2022).

La creciente diversidad de arquitecturas propuestas motivó la aplicación de Aprendizaje Automático Automatizado (AutoML) al problema. Siavash & Moin propusieron AutoML4NILM³ (Siavash & Moin, 2026), un framework que aplica Optimización Bayesiana para automatizar la selección de modelos y el ajuste de hiperparámetros sobre un espacio de 11 algoritmos, demostrando que los pipelines automatizados pueden alcanzar rendimiento competitivo sin la necesidad de una configuración manual.

Sin embargo, el framework trata todos los modelos candidatos como igualmente viables previamente. No existe mecanismo para predecir, antes de ejecutar el pipeline completo, qué modelos son arquitectónicamente adecuados para condiciones específicas del problema. Este gap tiene consecuencias prácticas dada la carga computacional documentada de la búsqueda bayesiana, la exploración no informada desperdicia recursos en candidatos estructuralmente inadecuados. La evidencia de la literatura sugiere que las ventajas sistemáticas entre modelos son detectables a elecciones arquitectónicas concretas: SGN (Shin et al., 2018) reduce el MAE entre un 18% a 25% en electrodomésticos on/off mediante gating binario; TreeCNN (Jia et al., 2019) supera a modelos secuenciales cuando las magnitudes de consumo varían ampliamente; DCNN-GASF (Yin et al., 2023) muestra ventajas al convertir señales 1D en representaciones 2D. Este trabajo propone sistematizar y validar estos patrones como propiedades arquitectónicas verificables.

³ Repositorio: <https://github.com/qas-lab/AutoML-for-NILM>.

2 Antecedentes

NILM busca recuperar las señales de consumo individuales $p_i(t)$ a partir de la medición agregada:

$$P(t) = \sum_i p_i(t) + \varepsilon(t) \quad (1)$$

La tarea se descompone en la detección de estados (on/off) y estimación de potencia. La complejidad de la desagregación depende del solapamiento entre firmas de potencia. Se formaliza como coeficiente de complejidad, identificando el tipo de electrodoméstico como factor crítico que la mayoría de los sistemas no controla (Nalmpantis & Vrakas, 2019).

En materia de evaluación, se identificaron al menos cuatro métricas de exactitud energética incompatibles en uso simultáneo previo a NILMTK⁴ (librería open-source para el uso de NILM), lo que imposibilitaba la comparación entre papers (Batra et al., 2014). NILMTK estandarizó F1-score para detección de estados, Signal Aggregate Error (SAE) para error de estimación energética, y Total Energy Correctly Assigned (TECA) para la proporción de energía correctamente atribuida. AutoML4NILM evalúa exclusivamente con MAE y accuracy, quedando parcialmente alineado con este estándar y limitando la capacidad de distinguir arquitecturas según su fortaleza en cada subtarea.

3 Metodología

El trabajo se estructura en 4 etapas.

3.1 Construcción de la matriz modelos-propiedades

Se construye una matriz donde cada fila corresponde a un modelo NILM documentado en la literatura y cada columna a una propiedad arquitectónica candidata. Los modelos considerados incluyen los 11 presentes en AutoML4NILM y modelos adicionales con resultados publicados, entre ellos SGN, TreeCNN y DCNN-GASF. Cada propiedad se codifica como variable binaria: presente o ausente, determinable por inspección de la descripción del modelo sin ejecutar ningún experimento.

3.2 Módulo de evaluación multi-métrica

AutoML4NILM se extiende con un módulo que implementa F1-score, SAE y TECA junto al MAE existente, alineando el framework con los estándares de la comunidad. Esta extensión permite distinguir propiedades que afectan la detección de estados de aquellas que afectan la estimación energética, para así enriquecer el análisis de correlación.

⁴ Sitio oficial: <http://nilmtk.github.io/>. Repositorio: <https://github.com/nilmtk/nilmtk>.

3.3 Análisis de correlación y formalización

Con la matriz y los vectores de rendimiento multi-métrico, se realiza un análisis estadístico de correlación entre cada propiedad y cada métrica, desagregado por tipo de electrodoméstico. Las propiedades con correlación estadística significativa se formalizan como condiciones que se pueden verificar: enunciados precisos y chequeables en la arquitectura de cualquier modelo que se quiera probar.

3.4 Validación sobre datos industriales

Las propiedades identificadas se validan sobre datos reales de la Federación Argentina de Cooperativas Eléctricas (FACE). A diferencia de datasets públicos como UK-DALE⁵ (Kelly & Knottenbelt, 2015) o REDD⁶ (Kolter & Johnson, 2011), los datos industriales requieren preprocesamiento significativo: tratamiento de valores faltantes, corrección de inconsistencias de muestreo y estandarización de formatos. Se desarrolla un pipeline de preprocesamiento documentado y reproducible. La pregunta de validación es si los modelos que cumplen las propiedades formalizadas pueden mantener su ventaja predicha en un contexto industrial real.

4 Propiedades Arquitectónicas Candidatas

El análisis cruzado de la literatura permite identificar cuatro patrones recurrentes como candidatos a propiedades predictoras, pendientes de confirmación observacional mediante la metodología de la Sección 3.

4.1 Modelado de estados discretos

SGN demostró que un mecanismo de gating binario reduce consistentemente el MAE en electrodomésticos on/off independientemente de la arquitectura base. En este caso la condición candidata se presenta en presencia de una rama de clasificación cuya salida condiciona la estimación de potencia. Es verificable inspeccionando si el modelo produce una probabilidad de estado binario que multiplica o pondera su estimación de consumo.

4.2 Descomposición por magnitud de consumo

TreeCNN resuelve el problema de enmascaramiento (cargas de alto consumo suprimen la señal de cargas menor consumo) mediante una estructura jerárquica que separa electrodomésticos por magnitud antes de extraer características. La condición candidata se puede observar en la existencia de una estructura que aisle distintos rangos de magnitud de consumo en ramas de procesamiento independientes. Verificable comprobando si la arquitectura descompone el problema por escala de consumo antes de la extracción de características.

⁵ <https://jack-kelly.com/data/>.

⁶ <https://site.ieee.org/pes-bdaps/the-reference-energy-disaggregation-data-set-usa/>.

4.3 Representación temporal multi-escala

DCNN-GASF y los resultados usando TreCNN muestran que los modelos que operan sobre representaciones 2D de señales de potencia superan a modelos que tratan la señal como una secuencia 1D pura. Para esta propiedad la condición aspirante es el uso de una codificación 2D o multi-escala como entrada a la capa principal de extracción de características. La misma es verificable por la dimensionalidad de la representación de entrada del modelo.

4.4 Modelado del residuo desconocido

Nalmpantis y Vrakas señalan que en hogares con alta proporción de consumo (no etiquetado), los modelos que consideran explícitamente la carga residual desconocida obtienen mejor rendimiento global. Condición candidata para esta propiedad es la inclusión de un componente de salida o función de pérdida que mida la parte del consumo total que no se puede asignar a los electrodomésticos que ya conocemos. Verificable inspeccionando si el modelo produce una salida de residuo o incluye un término de residuo en su entrenamiento.

5 Contribuciones Esperadas

La principal contribución de este trabajo es generar un conjunto de condiciones arquitectónicas verificables que permitan predecir el rendimiento de un modelo en frameworks del tipo AutoML4NILM. La forma esperada de un resultado es: *"Para electrodomésticos on/off, los modelos que incorporan un mecanismo de gating binario alcanzan reducciones de MAE en el rango 15% a 25% respecto a modelos idénticos sin dicho mecanismo, independientemente de la arquitectura base."*

Esto permite a investigadores futuros razonar sobre la competencia de un modelo antes de experimentar, proporcionando un fundamento teórico para la construcción del espacio de búsqueda en AutoML para NILM. Otras contribuciones que se buscan aportar son el módulo de evaluación multi-métrica para AutoML4NILM y la obtención de resultados experimentales sobre el dataset industrial de FACE, constituyendo la primera evaluación documentada de un framework AutoML para NILM sobre datos de cooperativas eléctricas argentinas y abriendo una línea de investigación hacia el despliegue de NILM en la infraestructura energética argentina.

Durante la preparación de este trabajo, los autores utilizaron Claude Sonnet 4.6⁷ (Anthropic) únicamente con el fin de mejorar el lenguaje y la legibilidad del manuscrito. Tras utilizar esta herramienta, los autores revisaron y editaron el contenido según fuera necesario y asumen plena responsabilidad por el contenido de la publicación.

⁷ <https://www.anthropic.com/claude>.

References

- Angelis, G.-F., Timplalexis, C., Krinidis, S., Ioannidis, D., & Tzovaras, D. (2022). Nilm applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings*, 261, 111951. <https://doi.org/https://doi.org/10.1016/j.enbuild.2022.111951>
- Batra, N., Kelly, J., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., Singh, A., & Srivastava, M. (2014). Nilmtk: An open source toolkit for non-intrusive load monitoring. *Proceedings of the 5th International Conference on Future Energy Systems*, 265–276. <https://doi.org/10.1145/2602044.2602051>
- Hart, G. (1992). Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12), 1870–1891. <https://doi.org/10.1109/5.192069>
- Jia, Y., Batra, N., Wang, H., & Whitehouse, K. (2019). A tree-structured neural network model for household energy breakdown. *The World Wide Web Conference*, 2872–2878. <https://doi.org/10.1145/3308558.3313405>
- Kelly, J., & Knottenbelt, W. (2015). The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. *Scientific Data*, 2(1), 150007. <https://doi.org/10.1038/sdata.2015.7>
- Kolter, J. Z., & Johnson, M. J. (2011). Redd: A public data set for energy disaggregation research. *Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainability*, 25, 59–62.
- Nalmpantis, C., & Vrakas, D. (2019). Machine learning approaches for non-intrusive load monitoring: From qualitative to quantitative comparison. *Artificial Intelligence Review*, 52(1), 217–243. <https://doi.org/10.1007/s10462-018-9613-7>
- Shin, C., Joo, S., Yim, J., Lee, H., Moon, T., & Rhee, W. (2018). Subtask gated networks for non-intrusive load monitoring. <https://arxiv.org/abs/1811.06692>
- Siavash, N., & Moin, A. (2026). A bayesian optimization-based automl framework for non-intrusive load monitoring. <https://arxiv.org/abs/2602.05739>
- Yin, H., Zhou, K., & Yang, S. (2023). Non-intrusive load monitoring by load trajectory and multi-feature based on dcnn. *IEEE Transactions on Industrial Informatics*, 19(10), 10388–10400. <https://doi.org/10.1109/TII.2023.3240924>