energies

MDPI

Article

An Ensemble Method for Non-Intrusive Load Monitoring (NILM) Applied to Deep Learning Approaches

Silvia Moreno [1,2,*], Hector Teran [3], Reynaldo Villarreal [1], Yolanda Vega-Sampayo [1], Jheifer Paez [1], Carlos Ochoa [3], Carlos Alejandro Espejo [4], Sindy Chamorro-Solano [5] and Camilo Montoya [6]

1 Centro de Investigación, Desarrollo Tecnológico e Innovación en Inteligencia Artificial y Robótica AudacIA, Universidad Simón Bolívar, Barranquilla 080005, Colombia; reynaldo.villarreal@unisimon.edu.co (R.V.); yolanda.vega@unisimon.edu.co (Y.V.-S.); jheifer.paez@unisimon.edu.co (J.P.)
2 Systems Engineering Department, Universidad del Norte, Barranquilla 081007, Colombia
3 Faculty of Engineering, Universidad Simón Bolívar, Barranquilla 080005, Colombia; hector.teran@unisimon.edu.co (H.T.); carlos.ochoa@unisimon.edu.co (C.O.)
4 Centro de Crecimiento Empresarial e Innovación Macondolab, Universidad Simón Bolívar, Barranquilla 080005, Colombia; carlos.espejo@unisimon.edu.co
5 Centro de Investigación e Innovación en Biodiversidad y Cambio Climático Adaptia, Universidad Simón Bolívar, Barranquilla 080005, Colombia; sindy.chamorro@unisimon.edu.co
6 Solenium S.A.S., Medellin 050031, Colombia; camilom@solenium.co
* Correspondence: silvia.moreno@unisimon.edu.co or scmoreno@uninorte.edu.co; Tel.: +57-3005555132

Abstract: Climate change, primarily driven by human activities such as burning fossil fuels, is causing significant long-term changes in temperature and weather patterns. To mitigate these impacts, there is an increased focus on renewable energy sources. However, optimizing power consumption through effective usage control and waste recycling also offers substantial potential for reducing energy demands. This study explores non-intrusive load monitoring (NILM) to estimate disaggregated energy consumption from a single household meter, leveraging advancements in deep learning such as convolutional neural networks. The study uses the UK-DALE dataset to extract and plot power consumption data from the main meter and identify five household appliances. Convolutional neural networks (CNNs) are trained with transfer learning using VGG16 and MobileNet. The models are validated, tested on split datasets, and combined using ensemble methods for improved performance. A new voting scheme for ensembles is proposed, named weighted average confidence voting (WeCV), and it is used to create combinations of the best 3 and 5 models and applied to NILM. The base models achieve up to 97% accuracy. The ensemble methods applying WeCV show an increased accuracy of 98%, surpassing previous state-of-the-art results. This study shows that CNNs with transfer learning effectively disaggregate household energy use, achieving high accuracy. Ensemble methods further improve performance, offering a promising approach for optimizing energy use and mitigating climate change.

Keywords: NILM; convolutional neural networks; climate change; energy consumption optimization

Citation: Moreno, S.; Teran, H.; Villarreal, R.; Vega-Sampayo, Y.; Paez, J.; Ochoa, C.; Espejo, C.A.; Chamorro-Solano, S.; Montoya, C. An Ensemble Method for Non-Intrusive Load Monitoring (NILM) Applied to Deep Learning Approaches. *Energies* **2024**, *17*, 4548. https://doi.org/10.3390/en17184548

Academic Editors: John Boland, Shunli Wang, Jiale Xie and Guang Wang

Received: 4 July 2024
Revised: 3 August 2024
Accepted: 11 August 2024
Published: 11 September 2024

1. Introduction

Climate change is one of the most pressing concerns worldwide. It involves long-term changes in temperature and weather patterns. Since the 19th century, human activity has been the main driver of climate change, mainly due to burning fossil fuels such as coal, petroleum, and gas. This generates greenhouse gas emissions that act like a blanket around the Earth, capturing the heat of the sun and raising temperatures. The consequences of climate change include intense droughts, water shortages, serious fires, sea level rise, floods, melting of polar ice, catastrophic storms, and diminished biodiversity [1].

Due to the urgent need to reduce the negative impact of fossil fuels, research has been directed toward energy production, with many resources being invested in renewable energy sources such as wind, solar, and water-based generation. However, less attention has been directed towards the optimization of power consumption, where controlling usage and recycling waste can have a major impact on the electrical requirements of the globe [2]. To optimize energy consumption, measuring the power used by different appliances and devices in homes is necessary. However, in most locations, there is only the report of a single meter that measures total power usage. Recent research has focused on the estimation of disaggregated energy consumption in non-invasive ways from a single meter. This task is known as non-intrusive load monitoring (NILM).

This concept was proposed by Hart [3], and his study is considered the pioneer of this domain. Solutions proposed previously to this work had a strong hardware component. Intrusive monitoring points were installed in every home appliance and connected to a central data collector. Hart proposed an approach that used simple hardware and sophisticated software for data analysis, thus eliminating permanent intrusion in homes.

In more recent works, NILM has been addressed as a machine learning problem. Supervised and unsupervised methods have been applied to solve it. Supervised methods are based on datasets with the consumption data of every device and the aggregated signal. This approach seeks to generate models that learn how to disaggregate the appliance signals from the aggregated signal. Common techniques include Bayesian learning and neural networks. Unsupervised methods seek to find signatures of possible devices in the aggregated signal without prior knowledge of the devices in the circuit. The more commonly used unsupervised techniques are hidden Markov models (HMMs), which define several hidden states that the model can transition to, thus representing the operational condition of the device (on, off, and intermediate states). They then relate these states to observable results based on the analyzed consumption data [4].

In the last five years, there has been ongoing research to solve NILM, with varying methods being applied to handle this problem. For example, in the work presented by Lazzaretti et al. [5], a multi-agent architecture for an NILM solution was presented and evaluated. Five load event detection agents, feature extraction agents, and classification agents were studied to implement the best combinations of agents in LMMs. To evaluate the proposed system, the COOLL dataset and the LIT-Dataset were used. Performance improvements were detected in all scenarios, with power-ON and power-OFF detection improving by up to 13%, while classification accuracy improved by up to 9.4%.

Another approach is the one proposed by Biansoongnern and Plangklang [6], who presented an alternative low-cost embedded NILM system for household energy conservation with a low sampling rate. Four symmetry pattern features were extracted, containing information on the value of active power change, the value of reactive power change, and the number of intersection points between the active power data and the reference line as well as an estimation of an equation for the starting characteristics of the electrical equipment. The validity of the tests was checked for 1 month in three houses to analyze the results. The proposed method was able to detect 91.3% of total events, and the average accuracy of the system in disaggregating devices was 0.897.

Other authors have addressed the problem of large numbers of loads running simultaneously. In the work presented by Li et al. [7], the authors proposed a new NILM method based on dynamic time warping (DTW) optimization and event detection. Firstly, a feature extraction algorithm, STFT-SSAE, was constructed by using short-time Fourier transform (STFT) to extract time–frequency features from the load and then using a sparse stack autoencoder (SSAE) to extract important features from time–frequency information. Secondly, the above features were input into Bi-LSTM and DTW models, respectively, and a new probabilistic model was established. A Bi-LSTM-DTW load recognition architecture was built by combining the two models. Finally, the load identification model of SSAE-Bi-LSTM based on DTW optimization (DOSL) was trained by the preset combined data, which ensured the high confidence of the DOSL model in various complex operating scenarios. This method achieved an accuracy of 0.9412 in the PLAID dataset and an accuracy of 0.9306 in the UK-DALE dataset.

Another more recent work is the one presented by Pan et al. [8], which introduces the MUSENILM model, a non-intrusive load decomposition model incorporating a parallel multi-scale attention mechanism to enhance energy monitoring and management in smart grids. The core innovation of the proposed model is its ability to extract multi-scale features, enhancing the model's understanding of time series data and achieving significant performance improvements on the UK-DALE and REDD public datasets. Specifically, when MUSENILM identified the fridge electricity consumption pattern on the UK-DALE public dataset, compared to previous models, the accuracy improved from 88% to 91% and the F1 score increased from 87% to 90%. This model had an overall accuracy on the UK-DALE dataset of 0.98.

Other authors have focused their research on the management of the uncertainty inherent in energy consumption data. In the work presented by Li et al. [9], a promising approach to address this uncertainty is proposed by applying stochastic optimization. This approach considers stochastic programming based on scenarios to integrate the variability of photovoltaic (PV) energy production into active distribution network (ADN) optimization, applying methods such as the alternating direction method of multipliers (ADMM) to solve the optimization model in a distributed manner and protect data privacy. The incorporation of this approach improved the accuracy and robustness of NILM, enabling better load management and greater integration of renewable energy sources.

Another work that applied stochastic optimization techniques in the management of multi-agent energy systems is the one presented by Ding et al. [10]. These authors presented a promising direction for addressing uncertainty in the integration of renewable energy sources and the operation of smart buildings. In their paper, the authors highlighted the application of constrained random programming (CCP) to handle uncertainties related to wind and solar power generation as well as outdoor temperature using an adaptive alternating direction method of multipliers (ADMM) to solve income and payment subproblems in a cooperative and distributed context. The challenges for future research include the adaptation and improvement of these methods in non-intrusive load monitoring (NILM), specifically to optimize energy management and data privacy in scenarios with high penetration of renewable energy and smart buildings.

While the previous works used traditional machine learning algorithms, other authors have tried to apply deep learning techniques, such as convolutional neural networks (CNNs). In the work presented by Edmonds and Abdallah [2], instead of using the traditional approach of dealing with electricity data as time series, the IMG-NILM approach was used to transform time series into heatmaps, with higher electricity readings represented by 'hotter' colors. The image representation was then used in a CNN to detect the signature of an appliance from aggregated data. The proposed approach attained a test accuracy of up to 93% on the UK-DALE dataset within a single house, and in more challenging settings where electricity data were collected from different houses, IMG-NILM attained an average accuracy of 85%.

Another study that applied CNNs is the one presented by Nolasco et al. [11]. In this work, an integrated method for the detection, feature extraction, and classification of high-frequency NILM signals for the publicly available LIT-Dataset was presented. In terms of detection, the results were above 90% for most cases. For classification, the final accuracies were around 97%. These authors also included a multi-label procedure to avoid the disaggregation stage, indicating the loads connected at a given time and increasing the recognition of multiple loads.

A different approach was proposed by Machlev et al. [12]. These authors proposed solving NILM as a multi-objective optimization problem instead of a classic single-objective function. The main idea was to model each NILM feature as an objective function and to mutually minimize these objectives based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II). The presented algorithms can operate in real time using low sampling rates (0.25 Hz and lower) without training the system, just using information on the average power signatures of each appliance. MO-NILM version 1 achieved an average accuracy of 0.9626 on the REDD dataset with six appliances, and version 2 achieved an average F-measure of 0.938 on the AMPds dataset.

Another approach is the one proposed by Ma et al. [13]. This article proposed a multi-chain disaggregation method for NILM (MC-NILM). MC-NILM integrates the models generated by existing algorithms and considers the relation among these models to improve the performance of energy disaggregation. Given the high time complexity of searching for the optimal MC-NILM structure, this article proposed two methods to reduce the time complexity: the k-length chain method and the graph-based chain generation method. Finally, the authors used the Dataport and UK-DALE datasets to evaluate the feasibility, effectiveness, and generality of the MC-NILM approach. Different versions of this algorithm achieved F1-scores of 0.581, 0.633, and 0.620 in the UK-DALE dataset.

Another work that applied deep learning techniques is the study performed by Xu et al. [14]. In this study, a model was proposed that integrates the power and on/off states to simultaneously disaggregate the power and device on/off states. The model comprises two main modules: a power encoding module for power disaggregation and a convolutional state module (CSM) for on/off state disaggregation. The power encoding module utilizes BERT-LSTM and long short-term memory networks for initial energy disaggregation. In contrast, the CSM employs convolutional neural networks for device state disaggregation. The output of the power encoding module is multiplied by the probability of on/off states to obtain the final power. The proposed model was evaluated using the REDD and UK-DALE datasets. Compared to the baseline models, the results showed an improvement in the average accuracy of device state classification from 0.948 to 0.957 and a decrease in the average error between the real power and disaggregated power from 26.356 W to 25.108 W.

Other authors have applied deep learning to solve the issue of privacy in NILM architectures. In the work presented by Wang et al. [15], an NILM approach based on a pyramid network with a two-dimensional convolutional neural network (2D-CNN) was designed, and privacy was protected using homomorphic encryption and secure multiparty computation technology. Privacy-preserving protocols were designed for operators of pyramid networks, such as convolution, full connection, batch normalization, average pooling, ReLU, and upsampling, and were combined to construct a privacy-preserving 2D-CNN pyramid network. The entire process does not restore the original information contained in the data or the intermediate results, thereby protecting the privacy of both parties. The experimental results on the UK-DALE dataset showed that the pyramid network based on a 2D-CNN could perform well, with an accuracy of 95.81%, and that the privacy-preserving 2D-CNN pyramid network could maintain the inference performance of the 2D-CNN pyramid network while protecting the privacy of the client data and server model parameters with consistent accuracy and recall. Table 1 summarizes the previous works on NILM mentioned in this paper.

These and other works applied machine learning and deep learning techniques to solve NILM, but there has been no attempt yet to combine different models through ensembles. An ensemble model is created by generating multiple models and combining them to produce an output classification. To combine the different models, a voting process is performed among them to determine the result. There are different types of voting; the most common is average voting, in which the average of the probabilities for each class of all the models is computed, and then a classification is performed based on the average probability [16].

In this paper, images are generated from intervals of the total power signals with single appliance labels obtained from the UK-DALE dataset and used to train different deep learning models. Two pre-trained networks, VGG16 and MobileNet, are fine-tuned with additional CNN layers, applying transfer learning to classify the signal images into five different appliances. Then, several models are combined in an ensemble to improve the performance. Two voting algorithms are applied for the ensembles: average voting and a new method proposed in this paper, weighted average confidence voting (WeCV).

Standards metrics such as the overall accuracy and true positive rate of each appliance are computed and compared, showing that the use of ensembles can improve the base performance and outperform the state of the art.

Table 1. Previous work summary.

Authors	Dataset	Techniques	Performance
Lazzaretti et al. [5]	COOLL, LIT	Multi-agent machine learning architecture	0.998 accuracy
Biansoongnern and Plangklang [6]	Custom dataset	Symmetry features and classification	0.897 accuracy
Li et al. [7]	PLAID, UK-DALE	Dynamic time warping (DTW) optimization	0.9412 accuracy in PLAID, 0.9306 accuracy in UK-DALE
Pan et al. [8]	UK-DALE, REDD	multi-scale features, 1-D Convolutional Neural Networks (CNNs)	0.91 on fridge, 0.98 overall accuracy
Edmonds and Abdallah [2]	UK-DALE	Heatmaps and convolutional neural networks (CNNs)	0.93 accuracy
Nolasco et al. [11]	LIT	1-D convolutional neural networks (CNNs)	0.97 accuracy
Machlev et al. [12]	REDD, AMPds	Multi-objective optimization	0.962 accuracy in REDD, 0.938 F-measure in AMPds
Ma et al. [13]	UK-DALE	Multi-chain disaggregation	0.633 F1-score
Xu et al. [14]	REDD, UK-DALE	LSTM and CNNs	0.957 accuracy
Wang et al. [15]	UK-DALE	2D-CNN pyramid network	0.9581 accuracy

2. Materials and Methods

To predict the disaggregated loads of different home appliances from a single whole-house power meter, the UK-DALE dataset was used [17]. This dataset records the power demand from five houses. In each house, the whole-house mains power demand is recorded every six seconds as well as the power demand from individual appliances every six seconds. To load and manipulate the time series data, the Python library NILMTK was used [18]. This is an open-source specialized toolkit for non-intrusive load monitoring. Using NILMTK, the power time series from a single house was loaded. This series has information on the power consumption during the whole year of 2013. The dataset was split into training and test sets by creating a time window from January to the beginning of October for the training dataset, and data from the rest of the year comprised the test dataset. Then, the first household appliance's information from the training dataset was considered (fridge). The power time series of the fridge was selected, and a filter was applied to select the indexes of the observations where the fridge's active power was greater than zero. Afterwards, these indexes were used to select observations of the active power of the mains series in the time intervals where the fridge was on. Then, this time series was divided into a hundred intervals and plotted. These images were saved to be used as the input for a convolutional neural network (CNN).

This same process to generate appliance images was repeated for five appliances: fridge, washer/dryer, kettle, dishwasher, and HTPC. For all images, the same plotting parameters were applied, especially the scales of the axes. The images had a resolution of 300 × 300 pixels and three channels (RGB). Then, the same process was applied to the test dataset, but in this case, the mains time series was divided into twenty intervals.

After the training and test images were generated, they were used for a convolutional neural network scheme. The Python library Keras was utilized with a TensorFlow backend to generate, train, and evaluate the CNN. We used Keras version 2.9.0 and TensorFlow version 2.9.1. Then, transfer learning was applied for this study. First, two pre-trained neural networks, MobileNet V2 [19] and VGG16 [20], with weights from training with ImageNet were used as the initial layers of the CNN and frozen. Then, a few more layers were added according to two different architectures, as presented in Figures 1 and 2, and a process of fine-tuning was applied.

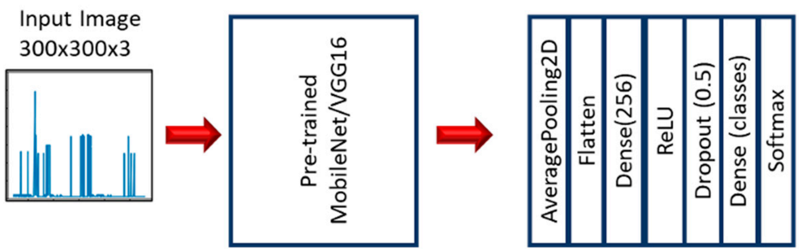

Figure 1. CNN architecture scheme Model 1.

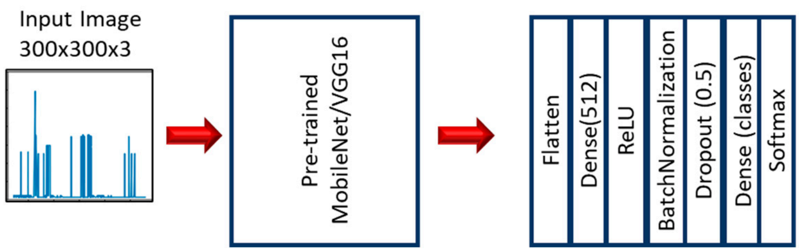

Figure 2. CNN architecture scheme Model 2.

The CNNs were trained with the images of the five appliances mentioned before, making this a multiclass problem with five classes. Several combinations of parameters were used. Different optimizers (Adam and SGD) and different numbers of epochs were tested with this architecture. First, the models were trained with the training dataset. This dataset was divided into training (90%) and validation (10%) sets for this process. After the models were trained and saved, they were applied to the test dataset, where the accuracy of the models was assessed.

The CNNs were trained in a Lenovo PC acquired in Barranquilla, Colombia with an AMD Ryzen 7 5800H processor with Radeon Graphics 3.20 GHz and 16 GB of RAM. The average training time for 100 epochs was approximately 31.13 min, and the average execution time of the prediction on the test dataset with the saved model was 11.138 s.

After these base models were trained and tested, an approach with ensembles was applied. Two voting schemes were employed: average voting and a new voting scheme proposed in this paper, weighted average confidence voting (WeCV). In the average voting scheme, the pseudo-probabilities for each class from the different models are summed and divided by the number of models:

$$P_c = \frac{\sum_{i=1}^{N} p_{c,i}}{N} \tag{1}$$

where P is the probability for class c (there are five classes in this case), N is the number of models in the ensemble, and $p_{c,i}$ is the probability for class c in the model number i.

After computing the final probability for each class with the ensemble, the class with the highest probability is selected. In our new proposed scheme, WeCV, a weighted average among the different models is calculated. The weight for each model varies according to the confidence level of the prediction of the model. The confidence level depends on the pseudo-probability for a certain class for each model. The weight for each model is determined in the following way:

$$Difference_{c,i} = \left| \frac{1}{N} - p_{c,i} \right| \quad (2)$$

$$Weight_{c,i} = \frac{Difference_{c,i}}{\sum_i^N Difference_{c,i}} \quad (3)$$

where c is the class (five classes or appliances in this case), $p_{c,i}$ is the pseudo-probability for class c on model i, and N is the number of models in the ensemble. After the weights for each model are calculated, the final probability for class c on the ensemble is calculated in the following way:

$$P_c = \sum_i^N \left(weight_{c,i} \times p_{c,i} \right)$$

After the probability for each class is computed, the class with the highest probability is selected. During the second stage of the experiment, ensembles of the best 3 and 5 models based on their performance in the training dataset were selected, and average voting and WeCV were applied to combine the probabilities of the models and find the final classification.

3. Results

In the first stage of the experiment, the base models of pre-trained networks combined with complementary architectures with a process of fine-tuning were tested. The best results in terms of overall accuracy and true positive rate (TPR) for each appliance are presented in Tables 2 and 3.

In this first experiment, we observed that five models achieved an accuracy higher than 95%, and two models had an overall accuracy of 97%. The best results were obtained with the combination of MobileNet, Model 2, and Adam, and with two models with VGG16, Model 2, and Adam but with different numbers of epochs. Considering the performance of each appliance, we can see that for two of the appliances, the kettle and washer/dryer, we obtained a perfect classification with a 1.0 TPR. Other appliances such as the fridge and the HTPC achieved a poorer performance but were always above 90%, which indicates an adequate sensitivity. The appliance with the lowest performance was the dishwasher, with a TPR that varied between 0.700 and 0.950, but in the best models, it achieved a decent performance.

Table 2. Base models best results, overall accuracy.

Code	Base Network	Fine-Tuning Model	LR	Epochs	Optimizer	Test Accuracy
1	MobileNet	Model 2	0.0005	50	Adam	0.93
2	MobileNet	Model 2	0.0005	100	Adam	0.95
3	MobileNet	Model 1	0.0005	50	Adam	0.91
4	MobileNet	Model 1	0.0005	100	Adam	0.92
5	MobileNet	Model 2	0.0005	50	SGD	0.91
6	VGG16	Model 2	0.0005	50	Adam	0.95
7	VGG16	Model 2	0.0005	100	Adam	0.95
8	MobileNet	Model 2	0.0005	5	Adam	0.97
9	MobileNet	Model 2	0.0005	100	Adam	0.97

Table 3. Base models' best results for appliance performance.

Code	Overall Accuracy	Dishwasher TPR	Fridge TPR	HTPC TPR	Kettle TPR	Washer/Dryer TPR
1	0.93	0.700	1.000	0.950	1.000	1.000
2	0.95	0.850	0.950	0.950	1.000	1.000
3	0.91	0.650	0.950	0.950	1.000	1.000
4	0.92	0.700	0.950	0.950	1.000	1.000
5	0.91	0.700	0.950	0.900	1.000	1.000
6	0.95	0.850	1.000	0.900	1.000	1.000
7	0.95	0.900	0.900	0.950	1.000	1.000
8	0.97	0.950	0.950	0.950	1.000	1.000
9	0.97	0.950	0.950	0.950	1.000	1.000

Next, ensembles of the best three and five models were tested with two voting schemes: average voting and WeCV. The best results in terms of the overall accuracy and true positive rates of each appliance are shown in Table 4.

Table 4. Ensembles' best results.

Ensemble Scheme	Overall Accuracy	Dishwasher TPR	Fridge TPR	HTPC TPR	Kettle TPR	Washer/Dryer TPR
Ensemble average (1, 2, 3, 4, 6)	0.92	0.700	0.950	0.950	1.000	1.000
Ensemble average (1, 2, 6)	0.92	0.800	0.850	0.950	1.000	1.000
Ensemble average (2, 6, 7)	0.96	1.000	0.850	0.950	1.000	1.000
Ensemble WeCV (6, 8, 9)	0.97	1.000	0.900	0.950	1.000	1.000
Ensemble WeCV (6, 8, 9, 2, 7)	**0.98**	**1.000**	**0.950**	**0.950**	**1.000**	**1.000**

In this experiment with ensembles, we observed that by combining the five best models mixing MobileNet and VGG16-based architectures and applying our proposed voting scheme, WeCV, we achieved the best result with an accuracy of 0.98. With this voting scheme, we also achieved a perfect classification for the dishwasher, which was the class with the lowest performance in the base model tests. The kettle and the washer/dryer also maintained a perfect performance, and the dishwasher and the fridge had a TPR of 0.95, which is a very good result. These results indicate that WeCV significantly improves the performance of the base models. In Table 5, we can see a comparison of the performances of several recent models with the UK-DALE dataset.

Table 5. Performance of several models with the UK-DALE dataset.

Authors	Accuracy
Li et al. [7]	0.9306
Pan et al. [8]	**0.98**
Edmonds and Abdallah [2]	0.93
Wang et al. [15]	0.9581
Xu et al. [14]	0.957
Our model	**0.98**

In this table, we can see that the highest accuracy obtained with the UK-DALE dataset is 0.98 and was achieved by the model proposed by Pan et al. [8] and our model. However, in their paper, Pan et al. stated that they achieved an accuracy of 0.91 on the fridge prediction, and in our model, we achieved an accuracy of 0.95 for the fridge and a perfect performance for three appliances. The performance for all appliances was always above 95%, which indicates that our model has a balanced and adequate sensitivity for all appliances.

4. Discussion

The results of this study indicate that the approach of transfer learning with pre-trained CNNs works quite well for NILM, since all the best models achieved a high performance over 90%. In almost every best result, Adam was used as the optimizer, which indicates that this algorithm may work better for NILM with pre-trained CNNs. Additionally, better results were obtained with Model 2 as a complementary architecture, especially combined with VGG16, whereas no results over 90% were obtained with Model 1. Between VGG16 and MobileNet, better results were obtained with the latter. Regarding individual appliance performance, a perfect classification was obtained with the kettle and the washer/dryer, which probably indicates that the images of these time series are simpler. From the appliances, the dishwasher seemed to have the poorest performance of all the appliances, but in the best models, its performance was above 90%, which is still adequate.

In the experiment with ensembles, the best performance was achieved with the five best models and WeCV as the voting scheme. The best model achieved an overall accuracy of 0.98. As for the performance for each appliance, it was observed that three of the appliances (dishwasher, kettle, and washer/dryer) achieved a perfect classification, which is remarkable especially for the dishwasher, which had the lowest performance with the base models. For the fridge and the HTPC, although their performance was not perfect, they still had a TPR of 0.95, which is a very good result. Particularly for the fridge, this performance is better than that obtained by Pan et al. [8]. Although the improvement compared to the best base models was modest, it was still significant, and the ensemble matched or outperformed the state of the art with the UK-DALE dataset, where the best result is 98% [8].

5. Conclusions

This work explored non-intrusive load monitoring (NILM) with the well-known UK-DALE dataset. A new approach was applied, where images of the main meter series were generated, and pre-trained convolutional neural networks (CNNs) were applied after a process of fine-tuning to classify between five appliances. Then, ensembles of the best models were applied, and a new voting scheme proposed in this paper, WeCV, was implemented. With the base models, a high accuracy of 0.97 was achieved by two of the trained models, thus already outperforming other state-of-the-art works, such as that of Edmonds and Abdallah [2] which achieved an accuracy of 93% with the same dataset. We observed that the best results in the base models were always obtained with the complementary architecture number 2 (Model 2) and using Adam as an optimizer, which suggests that these parameters are better suited for this problem.

In the second stage, we tested ensembles of the best models, and with the five best models and WeCV, we achieved an even higher accuracy of 0.98. This result matches the best in the state of the art and surpasses the performance of the model for the fridge appliance. These results show that there is a promising future for performing NILM by applying CNNs with images of the main meter, and that better results can be achieved by using ensemble schemes. For future work, other pre-trained CNNs such as ResNet or Inception may be tested, as well as other less traditional ensemble voting schemes. Furthermore, this same approach may be applied to other public datasets for NILM.

Author Contributions: Conceptualization, all authors; methodology, S.M., R.V. and S.C.-S.; software, S.M., H.T., R.V. and C.M.; validation, J.P., C.O. and C.A.E.; formal analysis, S.M., S.C.-S. and Y.V.-S.; investigation, S.M., C.O. and J.P.; resources, C.A.E. and C.M.; data curation, Y.V.-S., H.T. and C.M.; writing—original draft preparation, S.M., H.T. and Y.V.-S.; writing—review and editing, S.C.-S., J.P. and C.O.; visualization, S.M. and C.A.E.; supervision, R.V. and C.A.E.; project administration, R.V., C.A.E. and C.M.; funding acquisition, Y.V.-S. and S.C.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Colombian Ministry of Science, MINCIENCIAS, grant number 91661.

Data Availability Statement: The dataset for this study can be found at https://jack-kelly.com/data/ (accessed on 20 October 2023).

Acknowledgments: This research could not have been performed without the support of the Colombian Ministry of Science, Simon Bolivar University, and the company Solenium S.A.S, which inspired this work.

Conflicts of Interest: Author Camilo Montoya was employed by the company Solenium S.A.S. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. United Nations What Is Climate Change? Available online: https://www.un.org/es/climatechange/what-is-climate-change (accessed on 18 May 2022).
2. Edmonds, J.; Abdallah, Z.S. IMG-NILM: A Deep Learning NILM Approach Using Energy Heatmaps. In Proceedings of the Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, Tallinn, Estonia, 27–31 March 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1151–1153.
3. Hart, G.W. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* **1992**, *80*, 1870–1891. [CrossRef]
4. Chavat, J.; Nesmachnow, S.; Graneri, J. Non-Intrusive Energy Disaggregation by Detecting Similarities in Consumption Patterns. *Rev. Fac. Ing. Univ. Antioq.* **2021**, *98*, 27–46. [CrossRef]
5. Lazzaretti, A.E.; Renaux, D.P.B.; Lima, C.R.E.; Mulinari, B.M.; Ancelmo, H.C.; Oroski, E.; Pöttker, F.; Linhares, R.R.; da Silva Nolasco, L.; Lima, L.T.; et al. A Multi-Agent NILM Architecture for Event Detection and Load Classification. *Energies* **2020**, *13*, 4396. [CrossRef]
6. Biansoongnern, S.; Plangklang, B. An Alternative Low-Cost Embedded NILM System for Household Energy Conservation with a Low Sampling Rate. *Symmetry* **2022**, *14*, 279. [CrossRef]
7. Li, M.; Tu, Z.; Wang, J.; Xu, P.; Wang, X. Dynamic Time Warping Optimization-Based Non-Intrusive Load Monitoring for Multiple Household Appliances. *Int. J. Electr. Power Energy Syst.* **2024**, *159*, 110002. [CrossRef]
8. Pan, G.; Wang, H.; Tian, T.; Luo, Y.; Xia, S.; Li, Q. Research on Non-Intrusive Load Decomposition Model Based on Parallel Multi-Scale Attention Mechanism and Its Application in Smart Grid. *Energy Build.* **2024**, *312*, 114210. [CrossRef]
9. Li, Z.; Su, S.; Jin, X.; Xia, M.; Chen, Q.; Yamashita, K. Stochastic and Distributed Optimal Energy Management of Active Distribution Networks Within Integrated Office Buildings. *CSEE J. Power Energy Syst.* **2024**, *10*, 504–517. [CrossRef]
10. Ding, B.; Li, Z.; Li, Z.; Xue, Y.; Chang, X.; Su, J.; Jin, X.; Sun, H. A CCP-Based Distributed Cooperative Operation Strategy for Multi-Agent Energy Systems Integrated with Wind, Solar, and Buildings. *Appl. Energy* **2024**, *365*, 123275. [CrossRef]
11. Nolasco, L.D.S.; Lazzaretti, A.E.; Mulinari, B.M. DeepDFML-NILM: A New CNN-Based Architecture for Detection, Feature Extraction and Multi-Label Classification in NILM Signals. *IEEE Sens. J.* **2022**, *22*, 501–509. [CrossRef]
12. Machlev, R.; Belikov, J.; Beck, Y.; Levron, Y. MO-NILM: A Multi-Objective Evolutionary Algorithm for NILM Classification. *Energy Build.* **2019**, *199*, 134–144. [CrossRef]
13. Ma, H.; Jia, J.; Yang, X.; Zhu, W.; Zhang, H. Mc-Nilm: A Multi-Chain Disaggregation Method for Nilm. *Energies* **2021**, *14*, 4331. [CrossRef]
14. Xu, W.; Jiang, C.; Zhang, Q.; Zheng, J. An Intelligent Non-Intrusive Load Monitoring Model Based on Power Encoding and Convolutional State Modules. *Meas. Sci. Technol.* **2024**, *35*, 086210. [CrossRef]
15. Wang, Y.; Zhou, P.; Ye, W.; Qi, W. Non-Intrusive Load Monitoring and Its Privacy-Preserving Scheme Based on Pyramid Network. *Jisuanji Gongcheng/Comput. Eng.* **2024**, *50*, 182–189. [CrossRef]
16. Moreno, S.; Bonfante, M.; Zurek, E.; Cherezov, D.; Goldgof, D.; Hall, L.; Schabath, M. A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC. *Tomography* **2021**, *7*, 154–168. [CrossRef] [PubMed]
17. Kelly, J.; Knottenbelt, W. The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes. *Sci. Data* **2015**, *2*, 150007. [CrossRef] [PubMed]
18. Batra, N.; Kelly, J.; Parson, O.; Dutta, H.; Knottenbelt, W.; Rogers, A.; Singh, A.; Srivastava, M. NILMTK: An Open Source Toolkit for Non-Intrusive Load Monitoring. In Proceedings of the 5th International Conference on Future Energy Systems, Cambridge, UK, 11–13 June 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 265–276.

19. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
20. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.