Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Energy Disaggregation at Billing Resolution: Metering Interval, Meter Channels and Calibration Length for Appliance-Level Feedback from Smart Meters

Rounak Panda
Independent Researcher
B.Tech. in Computer Science and Engineering,
Maulana Abul Kalam Azad University of Technology, West Bengal
Kolkata, West Bengal, India

***Abstract*—India is installing smart meters at a scale that will soon make interval consumption data available for most households, and appliance-level feedback derived from that data is one of the benefits routinely claimed for the programme. Work on disaggregation, which would supply such feedback, is evaluated almost entirely at the resolution of individual metering intervals. A household bill, and any feedback built on it, is an aggregate over a month. This paper asks whether the two agree. The evidence comes from four years of minute-level whole-house measurements with three submetered circuits. Four predictors are compared, from a constant-share baseline to gradient-boosted trees, across metering intervals from 1 to 60 minutes, three meter feature sets and seven calibration lengths. Both per-interval estimation accuracy and monthly kWh attribution error are reported for every configuration. The two metrics rank configurations differently on every circuit, in each of three independent test years: in all twelve circuit-years the configuration that wins on per-interval accuracy is not the one that wins on monthly error. Choosing by the conventional metric costs 4.9 percentage points of monthly error on average, and up to 8.5. The rank correlation between the two metrics is itself unstable, changing sign between test years for two of the four circuits, so the relationship cannot be summarised by a single coefficient. Coarsening the meter from 1 to 30 minutes costs the kitchen circuit 21.5 per cent of its per-interval accuracy. It changes the water heater, the laundry circuit and the unmetered remainder by less than seven per cent. On that evidence the 30-minute block profile mandated by Indian distribution utility specifications is adequate for billing-scale feedback, and inadequate only for event-level applications. Adding reactive power, voltage and current to the load profile improves per-interval accuracy by at most 0.045 and is close to worthless for the two largest circuits. Per-interval accuracy saturates after about a week of submetered calibration data, while monthly attribution error needs most of a year, because it depends on seasonal coverage rather than on waveform variety. A calibration performed once held for three subsequent years, and did so more stably than the constant-share heuristic it replaces.**

***Keywords*—non-intrusive load monitoring; energy disaggregation; smart metering; block load profile; evaluation metrics; consumer feedback; gradient boosting; demand-side management**

I. INTRODUCTION

India is in the middle of the largest smart metering programme ever attempted. Under the Revamped Distribution Sector Scheme, 19.79 crore consumer meters have been sanctioned, and 3.77 crore had been installed by 31 December 2025, with 5.28 crore installed nationally across all schemes [1]. Each of those meters records a block load profile. Distribution utility specifications set the capture period at a default of 1800 seconds, programmable to 900 seconds, in line with the load survey parameters of IS 15959 (Part 2) [2], [3]. In other words, the country is building a national archive of household consumption at half-hourly resolution.

One benefit routinely claimed for that archive is appliance-level feedback. If a household could be told that the water heater accounted for a third of last month's bill, it might act on the information. Evidence that disaggregated feedback changes behaviour is mixed but not negligible, with reported savings of a few per cent when the feedback is specific and timely [4], [5], [6]. Delivering it means solving the non-intrusive load monitoring problem: recovering appliance-level consumption from a single whole-house measurement. Hart posed that problem in 1992 [7], and a large literature has grown around it since [8], [9].

There is a mismatch between how that literature is evaluated and what a feedback product needs. Disaggregation is assessed at the resolution of individual metering intervals, using per-sample error or classification measures [10], [11], [12]. A bill is not a sequence of intervals. It is one number per circuit per month, and the errors a model makes within a month may cancel or may compound. Nobody appears to have asked whether a model chosen by the per-interval metric is the model a billing application should use. That is the first question this paper takes up.

Two further questions matter to whoever writes the meter specification. The load profile interval is a procurement choice with a direct cost in memory, communication and head-end storage, and 15 minutes costs roughly twice what 30 minutes costs. What does the finer interval buy? Similarly, a meter can be asked to log reactive energy, voltage and current alongside active energy. Does that help enough to justify the channels? And because any disaggregation model has to be trained against

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

real submetered ground truth, somebody has to install submeters in a sample of homes for some period. How long is long enough?

These are not open-ended research questions. They are answerable by measurement, and the answers are numbers a utility can put in a specification. This paper supplies them for one household with four years of data, and is explicit throughout about how far a single household licenses generalisation.

A. Contributions

Five contributions follow.

1) A billing-resolution evaluation. Alongside the usual per-interval estimation accuracy I report monthly kWh attribution error and monthly share error, which are what a feedback product actually displays.

2) Evidence that the two metrics disagree. The full grid of 57 configurations is evaluated on three independent test years. In all twelve circuit-years the best configuration under one metric is not the best under the other, and the cost of choosing wrongly is quantified. The rank correlation between the metrics changes sign between years for two circuits, which is a result about the metrics rather than about the models.

3) A metering interval curve. Accuracy is traced from 1 to 60 minutes for each circuit, separating loads for which the Indian 30-minute default is nearly free from those for which it is expensive.

4) A meter channel ablation. Temporal context and the reactive, voltage and current channels are valued separately, which turns a specification argument into a number.

5) Calibration length and durability. How many days of submetering a disaggregator needs, measured under both metrics, and whether one calibration survives three later years.

Section II reviews related work, Section III sets out the method, Section IV the protocol, Section V the results, and Section VI what follows from them.

II. RELATED WORK

A. Disaggregation methods

Hart framed non-intrusive load monitoring as the recovery of appliance states from aggregate power, and built the first edge-detection solution [7]. Later work moved to probabilistic state-space models, notably factorial hidden Markov models and their approximate inference [13], and then to neural sequence models: Kelly and Knottenbelt applied denoising autoencoders and recurrent networks [10], and Zhang et al. introduced sequence-to-point learning, which remains a strong baseline [14]. Surveys by Zoha et al. [8] and later reviews [9] cover the space in more detail than is useful to repeat here.

Almost all of this work targets high-frequency or minute-level data. A national rollout creates a different regime, coarser by one to two orders of magnitude. Section V-A finds that coarse sampling is survivable for large loads. Physically that is unsurprising: a water heater runs long enough for a half-hour bin to resolve it, and a microwave does not.

B. Datasets and evaluation

Public datasets have shaped the field: REDD [15], UK-DALE [16] and the household measurements used here [17]. Batra et al. built NILMTK to make comparisons reproducible [11], and later showed how fragile published comparisons had been [12]. Makonin and Popowich examined NILM performance metrics directly and argued that several in common use are misleading, particularly for appliances with low duty cycles [18]. Their concern is close to mine, but it stays within per-interval evaluation. What I do here is compare per-interval evaluation against an aggregate evaluation over a billing period, and show that the choice between them changes which configuration wins.

C. Feedback and the Indian metering context

Darby's review of feedback on energy consumption remains the standard reference for what direct and indirect feedback achieve [4], with later syntheses reaching broadly similar conclusions [5], [6]. On the Indian side, the rollout position is documented by the Ministry of Power [1], the meter requirements by IS 16444 and IS 15959 [2], [19], and distribution utility tender specifications set the operative load profile interval [3]. Commentary on the programme's progress and its data-use questions is available from independent energy policy groups [20].

III. METHOD

A. Problem statement

Let the meter report energy E_t consumed in interval t of length Δ. The house contains C circuits with true energies y_{tj} summing to E_t. A disaggregator produces $\hat{y}_{tj}$ from features available at the head-end. Two quantities are then of interest. The per-interval error compares $\hat{y}_{tj}$ with y_{tj} directly. The billing error compares the monthly totals, Σ over t in month m, which is what a consumer-facing product reports.

B. Features

Three nested feature sets separate what the meter must expose from what merely helps.

Set A, total only. Interval energy E_t plus calendar encodings: hour of day as sine and cosine, day of week, month of year, and a weekend flag. This is the minimum any block load profile provides.

Set B, plus temporal context. Set A plus E at lags and leads of one to four intervals, centred rolling mean, maximum and minimum over approximately one hour, the first difference, and the daily mean. All of this is derivable at the head-end from the profile itself, at no metering cost.

Set C, plus extra meter channels. Set B plus interval reactive energy, mean voltage, mean current and a derived power factor.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

These require the meter to log additional channels, so the gain from B to C is the quantity a specification argument turns on.

C. Models

Four predictors span the range from a heuristic a utility could deploy tomorrow to a learned model.

The constant-share baseline assigns each circuit a fixed fraction of E_t, estimated from the training period. It is trivial, it needs no features beyond the total, and it is exactly what a utility would do with no disaggregation at all. Ridge regression is a linear model over the standardised feature set. Gradient-boosted regression trees are fitted per circuit with histogram binning [21], [22], early stopping on a held-out tenth of the training data, at most 200 boosting iterations. The neural network has two hidden layers of 64 and 32 rectified units, trained with Adam and early stopping. It is deliberately small. Including it serves comparison rather than any bid for the state of the art.

Predictions are clipped at zero. Because the meter total is known exactly, predictions may also be rescaled so that they sum to E_t, and I report both forms.

D. Metrics

Per-interval quality is reported as the estimation accuracy standard in the disaggregation literature,

$$A_j = 1 - \Sigma_t \, |\hat{y}_{tj} - y_{tj}| \, / \, (2 \, \Sigma_t \, y_{tj}) \qquad (1)$$

which is bounded above by one and penalises both over- and under-assignment. Normalised root mean squared error and the coefficient of determination were also computed and tell the same story. Billing quality is reported as the mean absolute percentage error of monthly totals,

$$M_j = (1/|\mathcal{M}|) \, \Sigma_m \, |\hat{Y}_{mj} - Y_{mj}| \, / \, Y_{mj} \qquad (2)$$

with Y_{mj} the true energy of circuit j in month m. I also report the share error, meaning the mean absolute difference in percentage points between the predicted and the true share of the monthly bill. Of the two it is the more forgiving, and it is included because a feedback screen usually shows proportions rather than absolute energy.

E. Statistical treatment

Agreement between the two evaluation regimes is assessed with the Spearman rank correlation [23] computed across all configurations, per circuit. Where a model has a random seed, five seeds are run and reported with 95 per cent percentile bootstrap intervals [24]. Everything else is deterministic given the data split, so no further resampling is applied and none is implied.

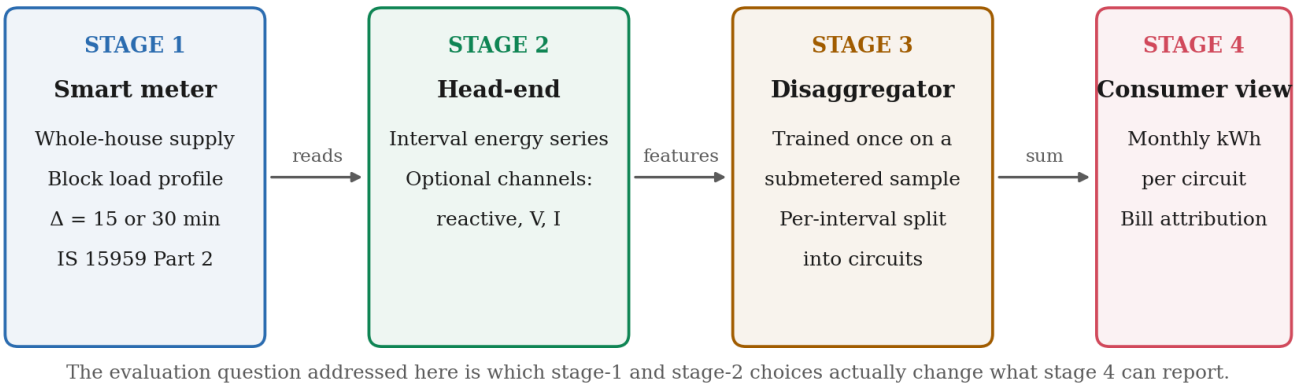

Fig. 1. The path from meter to consumer feedback. Stage 1 fixes the metering interval and the logged channels, and is a procurement decision. Stage 3 is trained once against submetered ground truth collected from a sample of homes. Stage 4 is what the household sees, and it is an aggregate over a billing period rather than a sequence of intervals. This paper asks which stage-1 and stage-2 choices actually change stage 4.

IV. EXPERIMENTAL PROTOCOL

A. Data

Measurements come from the individual household electric power consumption series archived at the UCI Machine Learning Repository [17]. They cover one house near Paris from 16 December 2006 to 26 November 2010 at one-minute resolution: 2,075,259 records, of which 1.25 per cent are missing. Each record carries global active and reactive power, voltage, current, and three submeters. Submeter 1 covers a kitchen with dishwasher, oven and microwave; submeter 2 a laundry area with washing machine, tumble drier, refrigerator and a light; submeter 3 an electric water heater and an air conditioner.

The three submeters do not cover the whole house. I therefore treat the difference between total consumption and the sum of the submeters as a fourth target, labelled unmetered, and it is the largest of the four at 51.2 per cent of energy. This is not an appliance. It is everything else in the house, and predicting it is the same problem a utility faces when it wants to tell a consumer how much of the bill is not explained by the loads it can identify. A small number of minutes, 1,050 of 2.05 million, give a slightly negative remainder through rounding, and are clipped at zero.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Minute records are aggregated into bins of $\Delta \in \{1, 5, 15, 30, 60\}$ minutes by summing energy. Any bin containing a missing minute is dropped, so that every retained bin is an exact energy sum rather than an interpolation. Table I gives the resulting series.

TABLE I
DATA USED, AFTER BINNING AND REMOVAL OF INCOMPLETE BINS

Quantity	Δ=1	Δ=15	Δ=30	Δ=60
Bins	2,049,280	136,551	68,240	34,085
Energy (kWh)	37,284	37,254	37,223	37,165
Kitchen share	6.2%	6.2%	6.2%	6.2%
Laundry share	7.1%	7.1%	7.1%	7.1%
Heater / AC share	35.5%	35.5%	35.5%	35.5%
Unmetered share	51.2%	51.2%	51.2%	51.2%

Shares are of total measured energy and are stable across binning because binning only sums. Bins at $\Delta = 30$ minutes number 17,367 in 2007, 17,549 in 2008, 17,361 in 2009 and 15,232 in the partial year 2010.

B. Splits

Models are trained on calendar year 2007. Headline numbers are reported on 2009, a two-year gap chosen deliberately so that no result depends on temporal adjacency. The comparison of evaluation regimes in Section V-C repeats the entire grid on 2008, 2009 and the partial year 2010, giving three test years that share no data. Section V-F uses the same three years to measure durability. The calibration-length study of Section V-E truncates the training set to the first N days of 2007 and leaves the test set untouched.

C. Implementation

Data handling uses pandas and NumPy [25]; the ridge, tree and network models use scikit-learn [21]. In full the grid is 5 metering intervals × 3 feature sets × 4 models, less the neural network at one-minute resolution, which was omitted on cost grounds. That gives 57 configurations. Every experiment here runs in about three minutes on one CPU core. It is worth saying so, because it means each claim below is cheap for a reader to check.

V. RESULTS

A. What the metering interval costs

Fig. 2 and Table II trace accuracy against Δ for the best model, boosted trees on feature set C. The four circuits behave in three distinct ways. Kitchen accuracy falls steadily, from 0.663 at one minute to 0.521 at thirty and 0.464 at sixty, a loss of 21.5 per cent of the one-minute value by the Indian default interval. The water heater and air conditioner circuit loses 2.2 per cent over the same range. The laundry circuit and the unmetered remainder do not lose anything: they gain 6.2 and 0.7 per cent respectively.

Event structure explains the pattern. A microwave or oven draws a large load for a few minutes, so a thirty-minute bin dilutes it into a background that the model cannot separate. A water heater runs long enough that a coarse bin still captures it. Slow or continuous loads dominate the laundry and unmetered series. For those, coarser bins average away minute-scale noise and make prediction slightly easier.

For a programme whose deliverable is monthly feedback, this comes close to an argument that the 30-minute default is the right choice. Circuits that survive coarse sampling carry 86.7 per cent of the energy in this house. The circuit that suffers carries 6.2 per cent. Even there, monthly attribution error at thirty minutes is 11.6 per cent against 12.8 at one minute, so the coarse meter is no worse. Finer intervals earn their cost in event-level applications, such as detecting that an appliance has been left running. They do not earn it in a monthly bill split.

TABLE II
METERING INTERVAL SWEEP (BOOSTED TREES, FEATURE SET C, TEST YEAR 2009)

Δ	Kitchen	Laundry	Heater / AC	Unmetered	metric
1 min	0.663	0.477	0.844	0.866	A_j
5 min	0.626	0.454	0.838	0.870	A_j
15 min	0.556	0.476	0.832	0.871	A_j
30 min	0.521	0.507	0.825	0.872	A_j
60 min	0.464	0.504	0.823	0.873	A_j
1 min	12.8	14.9	16.4	13.0	M_j
15 min	14.0	19.2	16.2	12.2	M_j
30 min	11.6	16.4	16.1	12.0	M_j
60 min	12.5	16.8	16.7	11.7	M_j

Upper block: per-interval estimation accuracy A_j, higher is better. Lower block: monthly attribution error M_j in per cent, lower is better. Predictions rescaled to the known meter total.

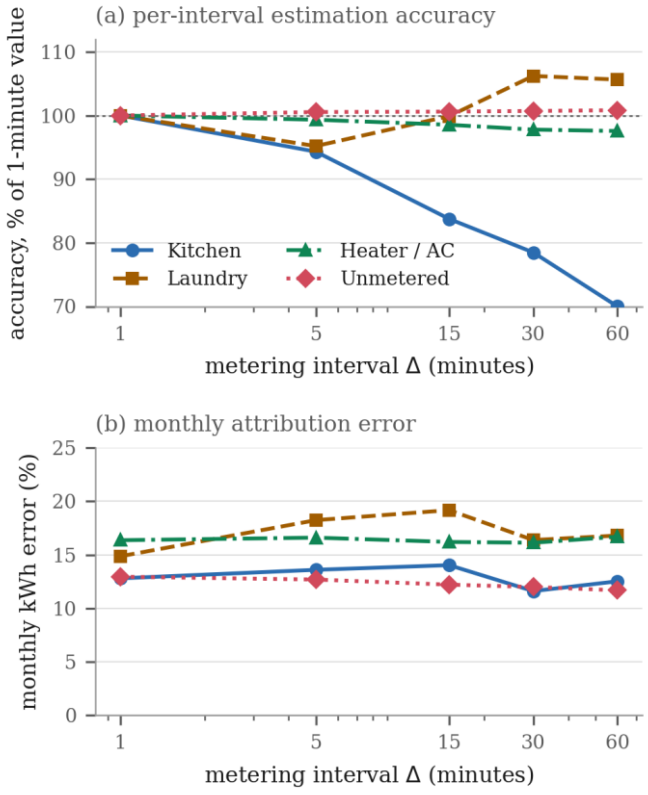

Fig. 2. Effect of the metering interval, relative to one-minute data. (a) Only the kitchen circuit degrades materially; the laundry circuit and the unmetered remainder improve slightly. (b) Monthly attribution error is nearly flat in the metering interval for every circuit. Logarithmic horizontal axis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

B. Which meter channels are worth logging

Fig. 3 and the ablation at thirty minutes separate the two increments. Adding temporal context, which the head-end can compute from the profile itself at no metering cost, raises kitchen accuracy by 0.073, laundry by 0.034, the heater circuit by 0.020 and the remainder by 0.009. Adding reactive energy, voltage and current on top of that raises kitchen by a further 0.012, laundry by 0.045, and the two largest circuits by 0.003 and 0.004.

For the circuits that dominate the bill, the reading is unambiguous. Extra electrical channels are close to worthless for the water heater and for the unmetered remainder, and the free head-end features are worth several times more than the channels that cost money. If a utility is choosing what to log for the purpose of consumption feedback, the answer is active energy and nothing else. Reactive power has other uses, in loss accounting and power quality, and nothing here argues against logging it for those reasons.

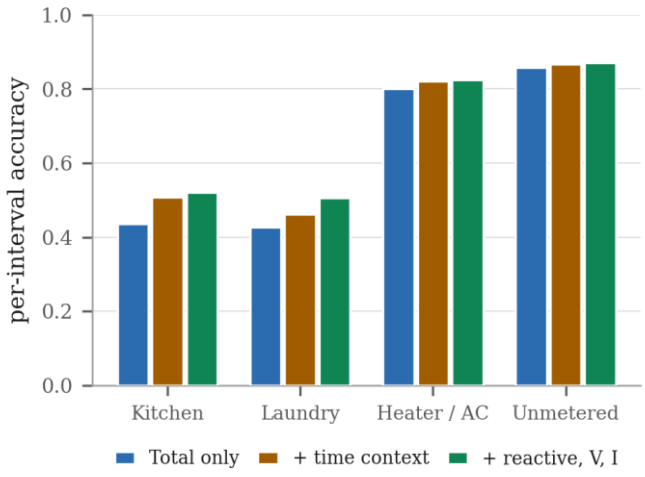

Fig. 3. Feature set ablation at a 30-minute interval. Temporal context derived from the load profile itself is worth more than the additional electrical channels, and for the two largest circuits the extra channels are worth almost nothing.

C. The two evaluation regimes disagree

This is the central result, and it is the one claim in the paper worth replicating before believing. The full grid was therefore evaluated three times, on 2008, 2009 and 2010. Fig. 4 plots monthly attribution error against per-interval accuracy for all 57 configurations in each year, one panel per circuit.

Start with what does not replicate. Only the laundry circuit shows a stable relationship, with ρ of −0.848, −0.855 and −0.869 across the three years: there, better per-interval models reliably give better monthly numbers. Everywhere else the correlation moves. The kitchen circuit gives −0.798, −0.178 and −0.786. The heater circuit changes sign outright, from +0.656 in 2008 to −0.632 and −0.643 in the two later years. The unmetered remainder, which is over half the bill, gives −0.067, +0.273 and −0.074, with bootstrap intervals spanning zero in every year. Table IV collects these with their intervals.

A single test year would have supported a stronger claim, and the wrong one. Taken alone, the 2009 value of +0.273 reads as evidence that the two metrics point in opposite directions for the largest circuit. Neither of the other two years bears that out, so the reading has to be abandoned. What all three years do support is weaker in form and stronger in evidence: the relationship between the two metrics is not stable enough to be described by a correlation at all. A practitioner cannot look up a coefficient and convert one metric into the other. Configuration choice is where this bites, and here the evidence is unanimous. In all twelve circuit-years the configuration that wins on per-interval accuracy is not the configuration that wins on monthly error. Not once in twelve does the conventional metric select the right model for a billing product. Take the unmetered remainder in 2009. Its best per-interval configuration is boosted trees on hourly data with the full channel set, giving 11.7 per cent monthly error. Its best monthly configuration is ridge regression on one-minute data with calendar features only, at 3.2 per cent. That second model scores 0.777 per interval, low enough that the usual evaluation would have discarded it before anyone looked at its monthly numbers.

The cost of choosing wrongly can be put in the units the product reports. Averaged over the twelve circuit-years, selecting by per-interval accuracy rather than by monthly error inflates monthly error by 4.9 percentage points, with a median of 4.8 and a maximum of 8.5. In ratio terms the penalty runs from 1.1 to 3.7 times the attainable error. A plausible mechanism is bias cancellation: a model with high per-interval accuracy may be consistently biased in one direction, and a month of consistent bias accumulates, whereas a noisier model whose errors change sign across the day aggregates closer to the truth. None of this makes per-interval evaluation wrong. It answers a different question from the one a billing application asks. The practical consequence is narrower and harder to argue with: a paper reporting only per-interval numbers cannot be used to choose a configuration for such an application.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

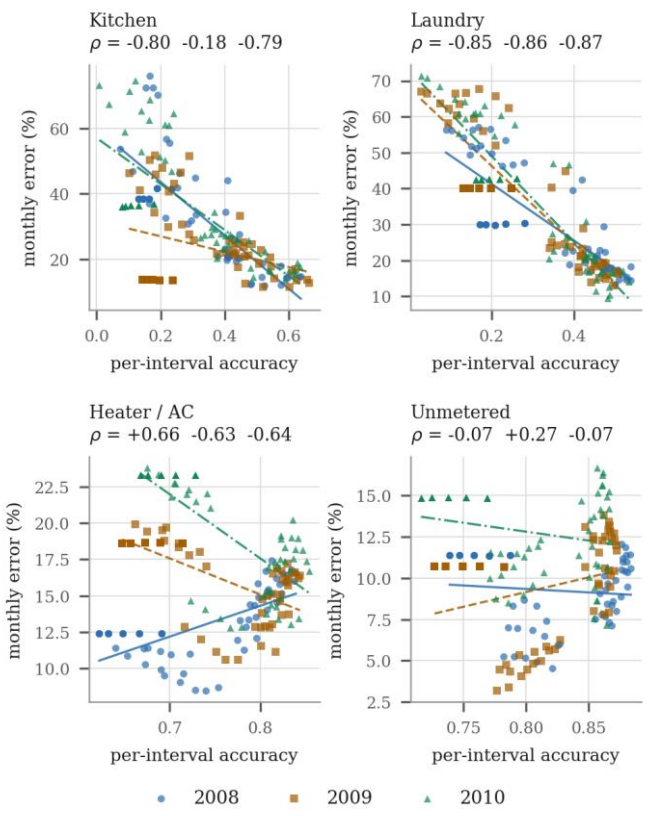

Fig. 4. Monthly attribution error against per-interval estimation accuracy, for all 57 configurations in each of three test years, with a least-squares guide line per year. The three ρ values above each panel are for 2008, 2009 and 2010 in that order. Only the laundry circuit shows a stable relationship. The heater circuit reverses sign between 2008 and the later years, visible as a guide line that slopes the wrong way.

TABLE IV
AGREEMENT BETWEEN THE TWO EVALUATION REGIMES, BY TEST YEAR

Circuit	ρ 2008	ρ 2009	ρ 2010	sign	penalty
Kitchen	−0.80	−0.18	−0.79	stable	2.7
Laundry	−0.85	−0.86	−0.87	stable	4.4
Heater / AC	+0.66	−0.63	−0.64	reverses	5.9
Unmetered	−0.07	+0.27	−0.07	reverses	6.6

ρ is the Spearman rank correlation between per-interval estimation accuracy and monthly attribution error, over all 57 configurations, computed separately in each test year. Bootstrap 95 per cent intervals span zero for the unmetered remainder in all three years and for the kitchen circuit in 2009; every other value excludes zero. “Penalty” is the mean increase in monthly attribution error, in percentage points, from selecting a configuration by per-interval accuracy instead of by monthly error, averaged over the three years.

D. How much better than a constant share?

It is worth asking what disaggregation buys over the null policy of telling every household that its bill divides in the average proportions. Table III and Fig. 5 compare the four predictors at thirty minutes on feature set C.

On per-interval accuracy the learned models are far ahead, as expected: boosted trees reach 0.521 on the kitchen circuit against 0.196 for the constant share. On monthly attribution the picture is less flattering. Boosted trees beat the constant share on the kitchen circuit, 11.6 against 13.5 per cent, and on laundry, 16.4 against 40.1. But on the unmetered remainder the constant share is better, 10.7 against 12.0 per cent, and on the heater circuit it is close, 18.6 against 16.1. In share-of-bill terms the constant share is within 0.7 percentage points on the kitchen circuit and 4.9 on the remainder.

So for the two circuits that make up 86.7 per cent of consumption, a model that learns nothing at all is competitive with gradient boosting at the resolution a bill is reported. That should temper claims about what disaggregation adds to billing feedback specifically. Where the learned models clearly earn their place is the laundry circuit, where the constant share is wrong by 40 per cent of the monthly total, and in any application finer than a month.

TABLE III
PREDICTORS AT Δ = 30 MINUTES, FEATURE SET C, TEST YEAR 2009

Predictor	Kitchen	Laundry	Heater/AC	Unmetered	metric
Constant share	0.196	0.198	0.690	0.763	A_j
Ridge	0.262	0.210	0.762	0.816	A_j
Boosted trees	0.521	0.507	0.825	0.872	A_j
Neural net	0.415	0.496	0.823	0.867	A_j
Constant share	13.5	40.1	18.6	10.7	M_j
Ridge	46.6	65.7	10.6	5.8	M_j
Boosted trees	11.6	16.4	16.1	12.0	M_j
Neural net	25.6	12.9	13.7	9.5	M_j
Constant share	0.7	2.4	7.5	4.9	share
Boosted trees	0.7	1.0	6.4	5.6	share

A_j is per-interval estimation accuracy, higher is better. M_j is monthly attribution error in per cent and “share” is the mean absolute error of the monthly share of the bill in percentage points, both lower is better. Rescaling predictions to the known meter total changed A_j by at most 0.012 and M_j by at most 3.3 percentage points, so it is a minor effect and is not tabulated separately.

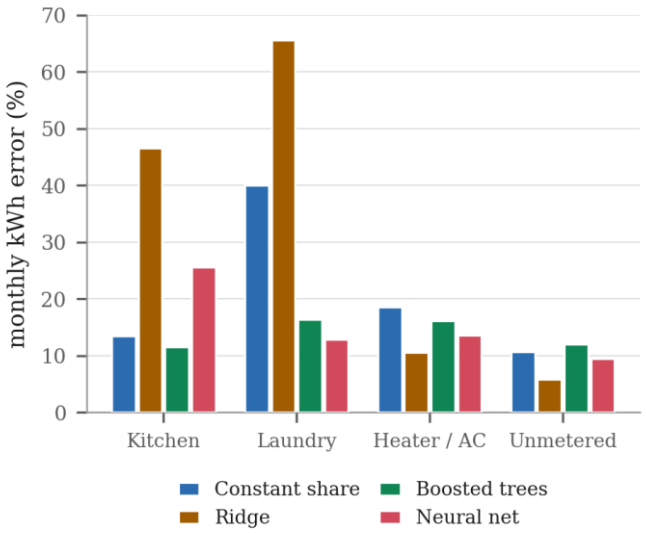

Fig. 5. Monthly attribution error by predictor at a 30-minute interval. The constant-share heuristic is competitive on the two circuits that carry most of the energy and badly wrong on the laundry circuit.

E. How long the submetering campaign has to run

Any deployment needs ground truth from somewhere, which in practice means temporarily submetering a sample of homes. Fig. 6 shows what different campaign lengths buy, training boosted trees on the first N days of 2007 and testing on 2009.

Per-interval accuracy saturates almost immediately. Seven days of data give 0.506 on the kitchen circuit against 0.521 for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

a full year, and 0.795 on the remainder against 0.872. By any per-interval reading, a week is nearly enough. Monthly attribution error tells a different story: at seven days the kitchen circuit is wrong by 78.2 per cent of the monthly total, falling to 48.2 at fourteen days, 25.7 at 120 days and 11.6 at a full year. The remainder follows the same shape, 23.2 per cent at a week against 12.0 at a year.

Why the two diverge is not mysterious. Per-interval accuracy needs enough examples of each load's waveform, and a week supplies those. Monthly attribution needs the seasons, and a model calibrated in January knows nothing about what the water heater does in July. This is the same theme for the third time. The metric chosen decides the engineering conclusion, and here it moves the recommended campaign length by a factor of fifty.

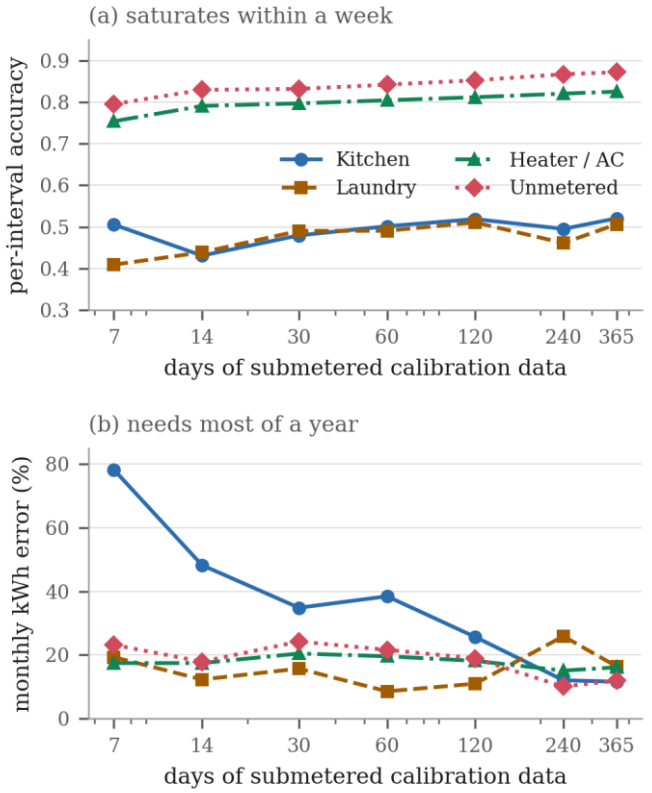

Fig. 6. Effect of the length of the submetered calibration campaign, boosted trees at a 30-minute interval, tested on 2009. (a) Per-interval accuracy is close to its ceiling after one week. (b) Monthly attribution error keeps improving until most of a year of data is available. Logarithmic horizontal axis.

F. Does one calibration last?

Training on 2007 and testing on each later year gives a direct measure of durability. Boosted trees held up: per-interval accuracy on the heater circuit was 0.818, 0.825 and 0.833 for 2008, 2009 and 2010, and monthly attribution error on the kitchen circuit was 12.5, 11.6 and 14.6 per cent. Three years after calibration, performance had not meaningfully decayed.

Far less stable was the constant-share baseline. Its monthly error on the kitchen circuit was 41.6 per cent in 2008, 13.5 in 2009 and 36.4 in 2010, a spread of 28 percentage points against 3 for the learned model. The household's average proportions moved between years while the relationship between the aggregate signal and the circuits did not. That is a point in favour of the learned model which the headline metrics in Table III do not show, and it matters for a deployment that will be recalibrated rarely if ever.

G. Seed stability and what it implies for model selection

Five seeds at the operating point separate the two learned models sharply, and not in the way a single run would suggest. Boosted trees are almost deterministic: kitchen accuracy 0.513 with a 95 per cent interval of [0.507, 0.518] and monthly error 10.7 per cent [10.0, 11.4]. The neural network is not: kitchen accuracy 0.439 [0.405, 0.467] and monthly error 22.7 per cent [17.4, 28.2], an interval nearly eleven points wide.

A single-seed comparison between these two models could therefore report almost any ordering on the monthly metric. Since much of the disaggregation literature reports single runs, this is worth stating plainly, and it is the reason every model comparison above uses the same seed and the seed study is reported separately rather than folded into the headline numbers.

VI. DISCUSSION

A. What a utility should take from this

Four conclusions are specific enough to act on. The 30-minute block load profile already mandated by distribution utility specifications is adequate for monthly appliance feedback, and moving to 15 minutes is not justified by disaggregation quality alone. For feedback purposes a meter needs to log active energy; reactive energy, voltage and current add almost nothing for the circuits that dominate a bill. A submetering campaign intended to calibrate a disaggregator should be scoped in seasons rather than weeks, because the metric that governs a billing product needs most of a year even though per-interval accuracy saturates in seven days. And a disaggregator should be evaluated on the quantity the product displays, since choosing a configuration by per-interval accuracy would select a different, and in one case a worse, configuration for every circuit studied.

B. Limitations

Severe limitations apply here, and they belong before any conclusion is carried elsewhere. All measurements come from a single household in France between 2006 and 2010. Its appliance mix, tariff, climate and occupancy have nothing in common with an Indian household, and nothing in this paper establishes that the numbers transfer. What plausibly does transfer is the structure of the findings, because the mechanisms behind them are generic: short high-power events suffer from coarse binning, aggregate metrics forgive errors that cancel, and seasonal coverage governs seasonal accuracy. Before any of these numbers enters a specification it should be reproduced on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Indian data. The methods and scripts are written so that doing so is a small piece of work rather than a large one.

Four further limitations apply. The correlation instability of Section V-C cuts both ways: it undermines any single-coefficient summary, including a favourable one, and with three years from one household I cannot say whether the instability is a property of the metrics or of this house. The unmetered remainder is a mixture of loads rather than an appliance, so its accuracy is not comparable with appliance-level figures reported elsewhere. Only four predictors are compared. A sequence-to-point model of the kind now standard in the literature [14] would very likely improve the per-interval numbers. That would not disturb Section V-C, whose argument concerns the relationship between the two metrics rather than the ceiling of either. And the evaluation is retrospective: no household saw any of these numbers and no behaviour change was measured, so nothing here speaks to whether the feedback would work.

C. Future work

Replication on Indian interval data is the obvious next step, either from a distribution utility's load profile archive with a submetered subsample, or from an instrumented set of homes. A second is to test whether a model calibrated on submetered homes transfers to unsubmetered ones. Every deployment assumes it does, and a single-household dataset cannot examine the question at all. A third is to extend the metric comparison to the sequence models that dominate current work, and to weekly and daily aggregation as well as monthly, since a feedback product may report at several resolutions at once.

VII. CONCLUSION

Appliance-level feedback is one of the more concrete benefits claimed for a national smart metering programme, and the disaggregation literature it would draw on is evaluated at a resolution that does not match what such a product reports. Measuring both regimes across 57 configurations and three independent test years, I found that they select different configurations in all twelve circuit-years, at an average cost of 4.9 percentage points of monthly error. The correlation between the two metrics proved too unstable to summarise, changing sign between years for two of four circuits. Four smaller findings came out of the same experiments. The metering interval mandated by Indian utility specifications is adequate for this purpose. The additional electrical channels a meter can log are not worth their cost. Per-interval accuracy saturates after a week of calibration data, while billing accuracy needs most of a year. And a constant-share heuristic, which learns nothing, is competitive with gradient boosting on the two circuits that carry most of the energy.

Running through all of it is a single point: the evaluation metric, not the model, decided most of these engineering questions. Anyone specifying a disaggregation system should decide what the product reports before deciding how to measure the model, because in this study those two decisions were not separable.

DATA AND CODE AVAILABILITY

The measurements used here are publicly available from the UCI Machine Learning Repository [17]. The preparation, experiment and figure scripts, together with the result files from which every number in this paper is taken, are available from the author on request.

REFERENCES

[1] Press Information Bureau, Government of India, “Progress on smart meter installation under RDSS,” New Delhi, India, Feb. 2, 2026.
[2] Bureau of Indian Standards, IS 16444 (Part 1): 2015, A.C. Static Direct Connected Watthour Smart Meter Class 1 and 2 — Specification. New Delhi, India: BIS, 2015.
[3] BSES Rajdhani Power Limited, “Technical specification for smart meters,” tender document, New Delhi, India.
[4] S. Darby, “The effectiveness of feedback on energy consumption,” Environmental Change Institute, University of Oxford, Oxford, U.K., 2006.
[5] C. Fischer, “Feedback on household electricity consumption: a tool for saving energy?,” Energy Efficiency, vol. 1, no. 1, pp. 79–104, 2008.
[6] K. Ehrhardt-Martinez, K. A. Donnelly, and J. A. Laitner, “Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities,” American Council for an Energy-Efficient Economy, Report E105, Washington, DC, USA, 2010.
[7] G. W. Hart, “Nonintrusive appliance load monitoring,” Proceedings of the IEEE, vol. 80, no. 12, pp. 1870–1891, 1992.
[8] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, “Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey,” Sensors, vol. 12, no. 12, pp. 16838–16866, 2012.
[9] P. A. Schirmer and I. Mporas, “Non-intrusive load monitoring: a review,” IEEE Transactions on Smart Grid, vol. 14, no. 1, pp. 769–784, 2023.
[10] J. Kelly and W. Knottenbelt, “Neural NILM: deep neural networks applied to energy disaggregation,” in Proc. 2nd ACM Int. Conf. Embedded Systems for Energy-Efficient Built Environments (BuildSys), 2015, pp. 55–64.
[11] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, “NILMTK: an open source toolkit for non-intrusive load monitoring,” in Proc. 5th Int. Conf. Future Energy Systems (e-Energy), 2014, pp. 265–276.
[12] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, “Towards reproducible state-of-the-art energy disaggregation,” in Proc. 6th ACM Int. Conf. Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys), 2019, pp. 193–202.
[13] J. Z. Kolter and T. Jaakkola, “Approximate inference in additive factorial HMMs with application to energy disaggregation,” in Proc. 15th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2012, pp. 1472–1482.
[14] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, “Sequence-to-point learning with neural networks for non-intrusive load monitoring,” in Proc. 32nd AAAI Conf. Artificial Intelligence, 2018, pp. 2604–2611.
[15] J. Z. Kolter and M. J. Johnson, “REDD: a public data set for energy disaggregation research,” in Proc. SustKDD Workshop on Data Mining Applications in Sustainability, 2011.
[16] J. Kelly and W. Knottenbelt, “The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes,” Scientific Data, vol. 2, art. 150007, 2015.
[17] G. Hébrail and A. Bérard, “Individual household electric power consumption,” UCI Machine Learning Repository, 2012.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

[18] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation," Energy Efficiency, vol. 8, no. 4, pp. 809–814, 2015.
[19] Bureau of Indian Standards, IS 15959 (Part 2), Data Exchange for Electricity Meter Reading, Tariff and Load Control — Companion Specification. New Delhi, India: BIS.
[20] Prayas (Energy Group), "Smart metering in India: a work in progress," Pune, India.
[21] F. Pedregosa et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[23] C. Spearman, "The proof and measurement of association between two things," American Journal of Psychology, vol. 15, no. 1, pp. 72–101, 1904.
[24] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman and Hall, 1993.
[25] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, pp. 357–362, 2020.