Non-Intrusive Load Monitoring – Energy Disaggregation and Transfer Learning using Deep Learning Methods

Abstract— **Non-intrusive load monitoring (NILM) is a process of deducing what appliances are used in the house as well as their individual energy consumption. With an increase in the deployments of smart meters worldwide, there has been a motivational shift using from traditional source separation methods to data-driven NILM methods for dis-aggregating the individual appliance readings. In this paper, we provide a novel deep learning ensemble method to derive the individual appliance readings from the unified mains readings. Also we would be providing a generalised solution that would work effectively across different datasets spanning different countries. Such a generalised solution would be able to produce accurate results on a house that is not present in the training set, called cross domain transfer learning. We also apply appliance transfer learning in which model will be trained on a particular appliance and tested on a different appliance. Performance results of the designed network show a considerable improvement over the baseline models.**

I. Introduction

Without a solution to NILM there is no accurate way to filter consumption data, inform about consumption issues, nor present conservation solutions to the homeowner that would allow them to understand their home's consumption and take action to conserve. NILM would lead to increased cost-savings on consumer side and better energy prediction for power distribution companies. This would in-turn reduce the additional costs incurred in using secondary power sources required during unpredictable loads.

However, installing hardware sensors at each appliance level can be very expensive. Also it can give easy access for hackers to exploit the smart grid ecosystem and thus deduce a user's action by their usage. For example, a hacker may record a person's activities such as when he/she sleeps, cooks and does not stay in the house. This information could then be used by thieves to steal valuable belongings from the house. Thus, the privacy of a user is compromised in directly recording appliances readings.

Since the computation power has been rising manifold and with the advent of virtual machines, it has become easier to apply deep learning methods to derive the appliance readings from the aggregated mains readings. There appears to be a lot of scope for deep learning in this area of research where better predictions could lead to better optimization. Stacking different deep learning models gives better prediction results and reduces over-fitting.

Transfer learning can be implemented in two ways. One of the way is using the Appliance transfer learning (ATL) method and the other is Cross-domain transfer learning (CTL). In ATL, we use one of our 'complex' appliances like Washing machine to predict the 'simple' appliance like Kettle readings. In Cross-domain transfer learning, we will use model trained on one dataset against test set of another dataset. This saves a lot of computational machine time as we only need to train our model for one appliance and its learning can directly be used for another appliance.

The paper is structured as follows: In section II we discuss about the current research areas going on in the field of energy disaggregation. Section III draws out the architecture and implementation steps followed to implement the ensemble model while Section IV and Section V discusses the datasets used and results of our model. Finally, we discuss our conclusions and future work.

II. Literature Review

Appliance Loading Monitoring in a non-intrusive manner started somewhat a decade ago where the framework of NILM was introduced [1]. After that a number of different approaches have

been proposed from the hand-engineered feature detection through SIFT [2] to machine learning and deep learning models [3]. Earliest of the research began with the works of Hart [4], who introduced the practice of non-intrusive appliance load monitoring and from which many NILM algorithms using low frequency data were later designed.

Existing algorithms can be briefly classified into three approaches: Unsupervised, semi-supervised and supervised learning. *Unsupervised learning* approaches utilize only unlabeled aggregate data and consider an appliance as a hidden Markov model (HMM). Additive factorial hidden Markov models are extensions to these algorithms where the output is an additive function of all the hidden states. Accurate inferences in such models [5] is highly difficult as they use approximation techniques which are highly susceptible to local minima. Whereas, some Factorial Hidden Markov Models (FHMM) like the one by Ruoxi et al. [6] use nonparametric Bayesian models to simultaneously detect the number of appliances and merge multiple states corresponding to the identical appliance into one. *Semi-supervised learning* approaches use both labeled and unlabeled data and aim to reduce the labeling effort [7]. Finally, *supervised learning* approaches use submetered loads and/or hand-labeled observations to formulate NILM as a supervised learning. In our research, we will be using the unsupervised deep learning approach.

A wide array of research has been going for using deep learning techniques in this field. The results of comparison between shallow and deep networks [8] show a considerable improvement in using deep learning over shallow learning algorithms. In their research, using just the REDD dataset, we see that the mean squared error (MSE) of the shallow algorithm used - Support Vector Regression (SVR) lies between the less connected Deep Neural Networks (DNN) and LSTM. LSTM performs better than the conventional neural networks because of its property to selectively remember patterns for a long duration of time. Hence, it makes sense to stack different deep learning models together so that advantages of all models can be fully utilized to minimize the errors.

Recent literature on using deep learning methods largely includes the work done by Jack Kelly et al. In their paper [9], they discuss energy disaggregation using LSTM and denoising autoencoders and rectangles. Combinatorial Optimization (CO) and FHMM, discussed above are used as baselines for comparing the three neural networks. From their results, we can conclude that these neural networks do generalize very well. This establishes a strong evidence in our belief that the ensemble of the deep learning models we would be using, would also generalise well. We see that LSTM outperforms CO and FHMM on two-state appliances (kettle, fridge and microwave) but falls behind CO and FHMM on multi-state appliances (dish washer and washing machine). From their result above, we believe that our stack of models would eventually provide accurate results for all the types of appliances.

For running the above algorithms, majority of the algorithms use the Non-intrusive Load Monitoring Toolkit (NILM-TK) [10] to allow for comparison of different energy disaggregation algorithms on a variety of different parameters. It is an open-source toolkit that includes parsers for data sets, a collection of preprocessing algorithms, a set of statistics for describing data sets, two reference benchmark disaggregation algorithms and a suite of accuracy metrics. Its main aim is to provide a simple interface for the addition of new data sets and algorithms. The library has been rapidly picking up since its inception. A few significant changes have been introduced in the library [11] after its development. One such change was the introduction of a new NILMTK interface which reduces the barrier-to-entry for specifying experiments for NILM research. NILMTK-CONTRIB, a new repository compatible with the recent disaggregation algorithms, allowed users to own the implementation of their algorithm. Various algorithms in these, such as Edge Detection Algorithm, CO, Discriminative Sparse Coding, ExactFHMM, ApproxFHMM, Recurrent Neural Networks (RNN) are discussed. Authors conclude that edge detection algorithm worked better for fridges. We are extensively making use of this library for our research work.

Recent developments have been in the field of

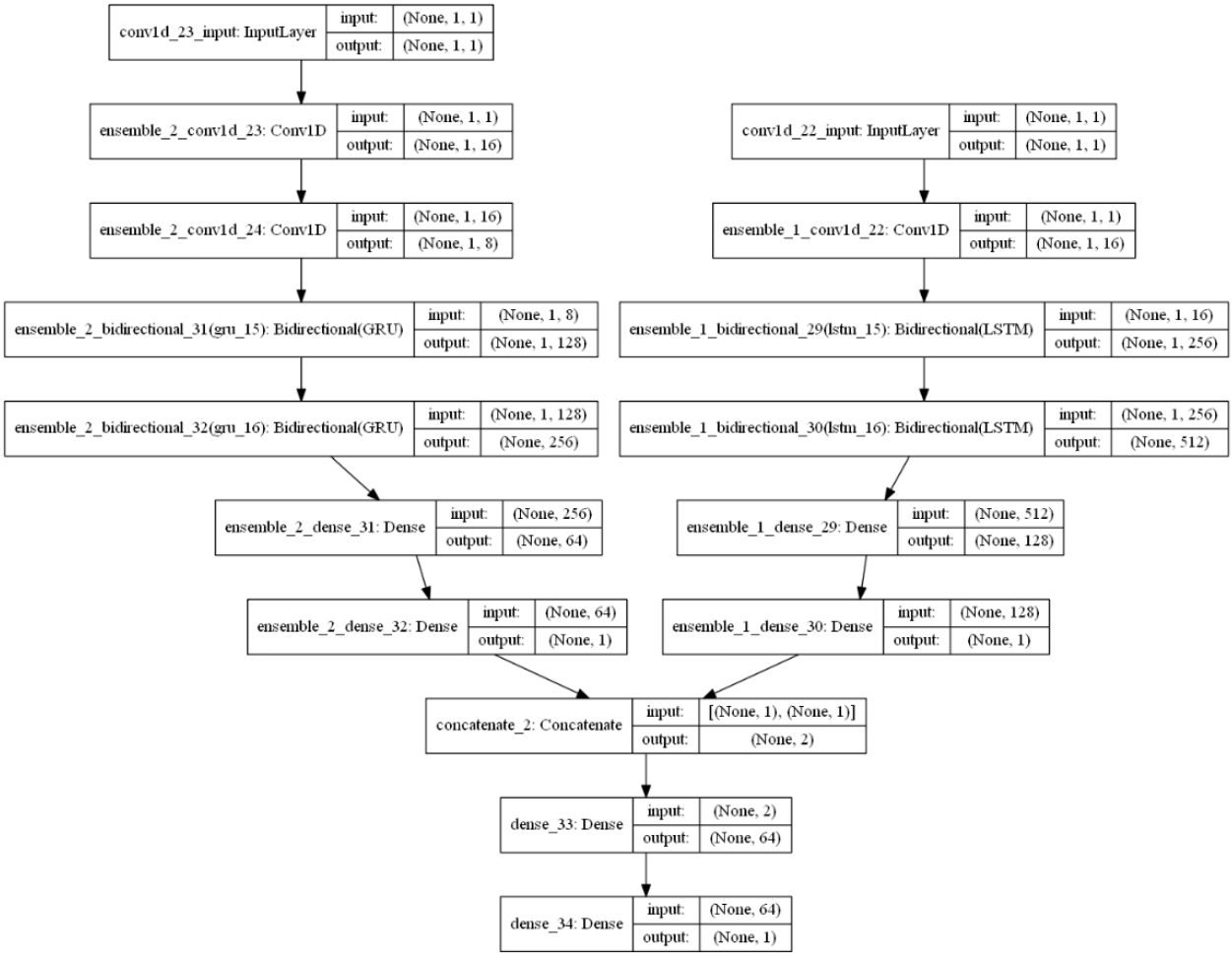

Fig. 1. Architecture of stacked ensemble model

transfer learning in which we predict results for a house not present in the training set using well-trained generalised model. David et. al in their work [12] demonstrate transferability using CNN and GRU network architectures on UK-REFIT [13] and UK-DALE [14] datsets having similar appliances and on US REDD [15] having different appliances. Another research work in this area of research presents that latent features learnt by a complex appliance like a washing machine can be transferred to a simple appliance like kettle. We believe that even our ensemble will have similar transferability properties and provide better predictions.

III. Architecture

In this section, we will look at the architecture of the ensemble deep learning model. We have looked at different literature survey and did not find any techniques which ensemble different deep learning models to predict the individual household appliance. We figured out the model suitable for our need based on the appliance data and tried different hyper parameters to find the best performing model. The model architecture is shown in **figure 1**. We have further applied this model for appliance transfer learning and cross domain transfer learning.

In the architecture, we have defined 2 models. Our **first model** is **Bidirectional GRU** and **second model** is **Bidirectional LSTM**.

In the first model, we have an Input Layer(Sequential), then we add a Conv1D layer, followed by 2 more Conv1D layers. We then have 2 Bidirectional GRU layers and finally we have 2 Dense layers.

In the second model, we have an Input Layer(Sequential), then we add a Conv1D layer and then we have 2 Bidirectional LSTM layers to it. Finally, we add 2 Dense layers.

Then we ensemble both the models and pass it through 2 Dense layers.

We have used our stack ensemble model for the purpose of the transfer learning both for the case of the Appliance transfer learning and the Cross Domain transfer learning.

IV. Experimental Setup

A. Data sets

Several open-source data sets are available for the purpose of energy disaggregation. These data were measured in household buildings from different countries. The sensors installed in these buildings read active power, but some sensors also read other information, for example, reactive power, current, and voltage. In NILM, the active power data are used. However, the main difference between the data sets is the sampling frequency. Due to this issue, pre-processing for aligning the readings need to be done before NILM algorithms are applied to the data. In literature, five appliances are usually considered for disaggregation which are kettle, microwave, fridge, dish washer and washing machine. For our project, we have particularly used the REDD and REFIT dataset.

The Reference Energy Disaggregation Data Set (REDD) is a data measured for 6 buildings in US. Measurements include mains with 1s sampling period and several appliances with 3s sampling period. High-frequency current and voltage measurements are also available at 15KHz sample frequency. The lengths of observations were between 3 and 19 days. The REFIT dataset includes cleaned electrical consumption data in Watts for 20 households in UK at aggregate and appliance level, timestamped and sampled at 8 second intervals.

B. Metrics

We have used the following metrics for evaluation in our project:

$$\textbf{recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\textbf{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{F1} = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

$$\textbf{accuracy} = \frac{TP + TN}{P + N} \tag{4}$$

$$\textbf{F1} = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

$$\textbf{accuracy} = \frac{TP + TN}{P + N} \tag{6}$$

$$\textbf{mean absolute error} = \frac{1}{T}\sum_{t=1}^{T} |\hat{y}_t - y_t| \tag{7}$$

$$\textbf{root mean squared error} = \frac{1}{T}\sum_{t=1}^{T} |\hat{y}_t - y_t|^2 \tag{8}$$

where

TP = number of true positive
FP = number of false positives
FN = number of false negatives
P = number of positives in ground truth
N = number of negatives in ground truth
$y_t^{(i)}$ = appliance i actual power at time t
$\hat{y}_t^{(i)}$ = appliance i estimated power at time t

	CO	FHMM
Appliance	RMSE	RMSE
Lights	123.84	73.62
Sockets	28.19	37.15
Microwave	235.3	545.96
Dishwasher	456.92	185.93
Fridge	107.66	100.26

Fig. 2. Metric values for CO and FHMM

V. Results

As part of the project, we are using two datasets namely REDD and REFIT datasets for prediction. REFIT contains house meter readings of UK, while REDD contains meter readings of US. The database has information of both the low and high frequency readings. The use of the two datasets is to show how transfer learning can be effectively applied using deep learning ensemble method. We used the REDD dataset, and specifically data for building 1 for training and validation for stacked

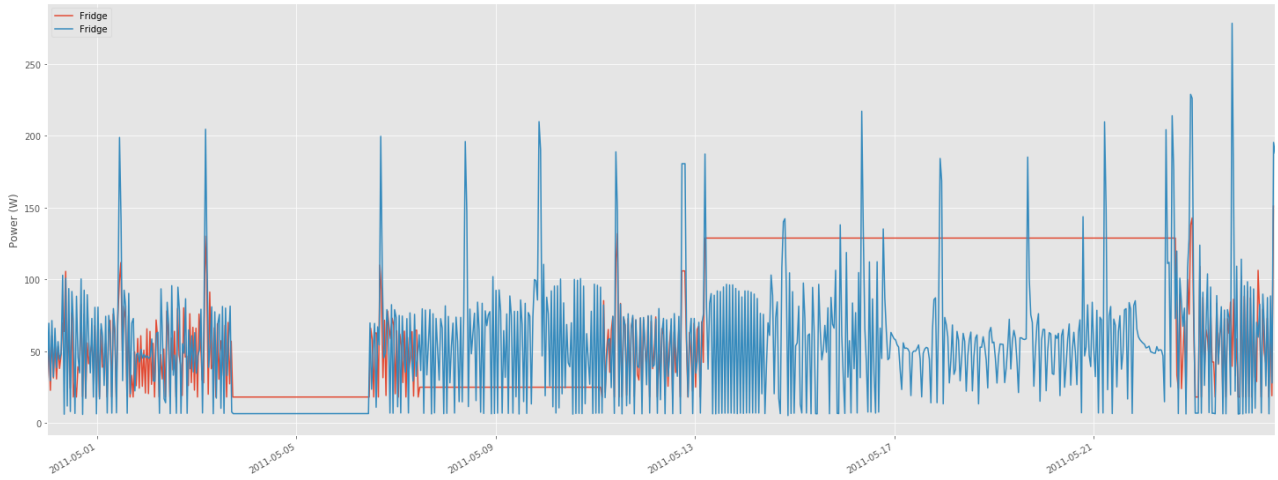

Fig. 3. Predictions(in blue) and Ground Truth(in red) for DAE

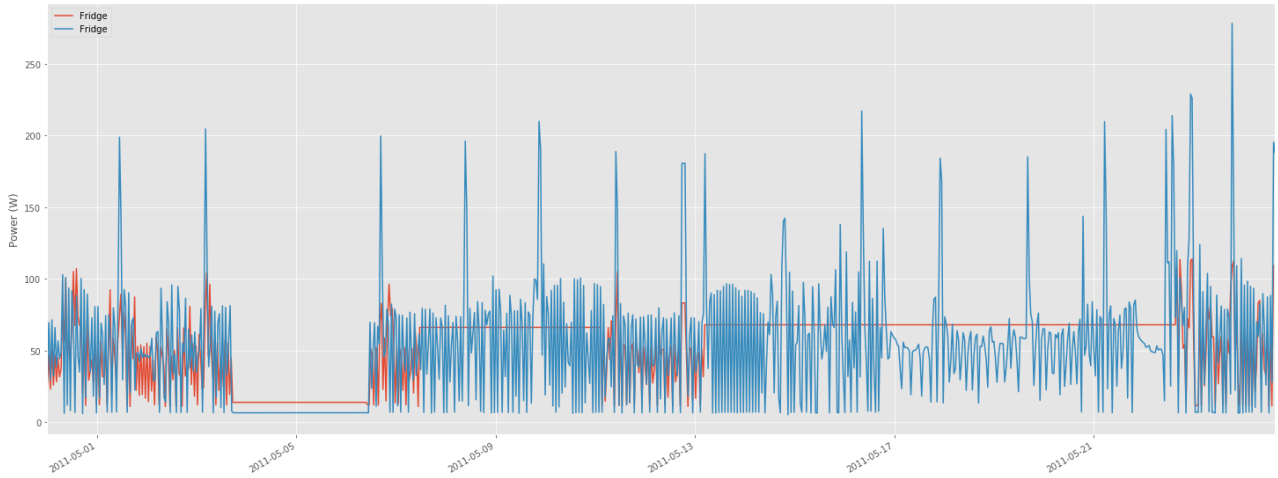

Fig. 4. Predictions(in blue) and Ground Truth(in red) for Average Ensemble

	DAE					Average Ensemble					
Appliance	**Recall**	**Precision**	**F1-score**	**MAE**	**RMSE**	**Recall**	**Precision**	**Accuracy**	**F1-score**	**MAE**	**RMSE**
Metric	0.99	0.89	0.94	38	41.79	0.85	0.92	0.8	0.88	35.03	40.73
Lights	1	0.99	0.99	15.93	18.15	1	0.99	0.99	0.99	16.07	18.3
Sockets	0.33	0.74	0.46	27.58	118.57	0.88	0.74	0.68	0.8	26.52	91.97
Microwave	1	0.7	0.85	17.37	76.72	1	0.75	0.75	0.85	30.93	78.77
Dishwasher	0.99	0.8	0.89	44.67	70.97	0.93	0.81	0.77	0.86	41.54	64.56
Fridge	0.99	0.67	0.8	32.98	213.7	0.99	0.67	0.67	0.8	42.95	208.8

Fig. 5. Metric values for DAE and Average Ensemble

ensemble. Data till April 30, 2011 was taken for training while the rest after it was used for validating and testing. The ratio of training, validation, testing was 80:10:10 respectively. We used Google Cloud Platform for training our models on a virtual Ubuntu machine with a 30 GB RAM. REFIT dataset was used for transfer learning.

	RNN					Stacked Ensemble				
Appliance	Recall	Precision	F1-score	MAE	RMSE	Recall	Precision	F1-score	MAE	RMSE
Lights	0.32	0.95	0.48	23.36	43.01	0.4	0.98	0.56	20.72	38.26
Sockets	1	0.99	0.99	5.48	10.77	1	0.99	0.99	16.69	18.97
Microwave	0.02	0.75	0.04	27	179.31	0.14	0.76	0.25	26.21	162.15
Dishwasher	0.03	0.86	0.06	6.88	79.59	0.25	0.87	0.22	5.98	67.77
Fridge	1	0.81	0.89	74.93	172.67	0.34	0.98	0.54	51.89	105.99

Fig. 6. Metric values for RNN and Stacked Ensemble

	GRU					Stacked Ensemble				
Appliance	Recall	Precision	F1-score	MAE	RMSE	Recall	Precision	F1-score	MAE	RMSE
Lights	0.3	0.96	0.46	21.63	41.18	0.4	0.98	0.56	20.72	38.26
Sockets	1	0.99	0.99	16.5	18.78	1	0.99	0.99	16.69	18.97
Microwave	0.02	0.74	0.03	32.84	209.17	0.14	0.76	0.25	26.21	162.15
Dishwasher	0.08	0.85	0.15	26.73	140.52	0.25	0.87	0.22	5.98	67.77
Fridge	0.38	0.94	0.54	60.2	128.96	0.34	0.98	0.54	51.89	105.99

Fig. 7. Metric values for GRU and Stacked Ensemble

A. Baseline Model-1

We considered Combinatorial Optimisation(CO) and Factorial Hidden Markov Model(FHMM) as our baseline models. These models are considered as legacy models and are very naive in their approach. Their implementation in the NILMTK library was used for training. The results we got for these models are as follows:

Fig. 2 shows that both CO and FHMM give very poor classification results with an RMSE as high as 545W for Microwave.

B. Baseline Model-2

Our next baseline model for comparison was the Denoising Auto Encoder (DAE) model proposed by Kelly et. al in their research paper 'Neural NILM: Deep Neural Networks Applied to Energy Disaggregation'. The DAE neural network architecture consists of 3 fully connected layers and a couple of 1D convolutional layers. Results were obtained on DAE model for 25 epochs.

Fig. 3 shows the ground truth consumption and the predicted consumption fit for fridge by DAE. For this and all the subsequent graphs, red denotes ground truth while blue denotes the prediction.

Thus we can see that DAE performs much better than CO and FHMM for the mentioned metrics.

C. Model-1 : Average Ensemble

Our first approach to segregate individual appliance consumption from the mains electricity data was using the average ensemble. This approach uses the same model as DAE but uses 5 instances of the same model for training and the final prediction is the average of individual predictions of each of these models. Fig. 4 shows the results obtained for Average Ensemble model for 25 epochs.

Fig. 5 shows the comparison between DAE and Average Ensemble for different metrics for 25 epochs.

Average Ensemble performs better than DAE for almost all metrics for all appliances.

D. Model-2 : Stacked Ensemble

Our second approach implements stacked ensemble deep learning by stacking 2 architectures - Recurrent Neural Network(RNN) which has Bidirectional LSTM at its core and Gated Recurrent Unit (GRU). The RNN network is inspired by the same research paper by Kelly et al as mentioned above. These 2 stacked models are then fed to a dense layer which again passes the learning to final dense layer. Our stacked ensemble model seems to be either performing better in most of the cases or the same for other cases. For eg. stacked ensemble

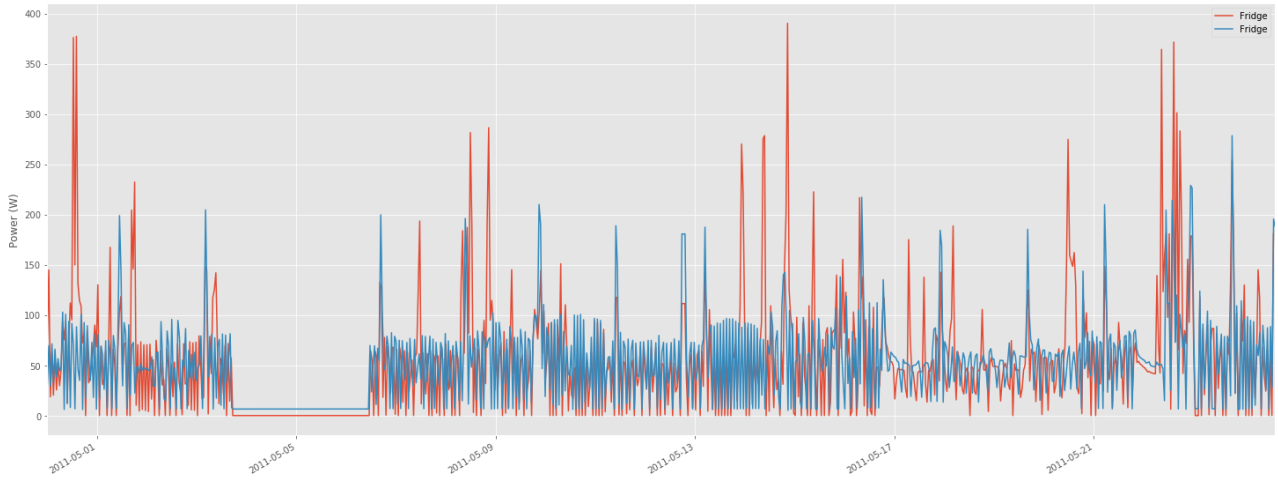

Fig. 8. Predictions(in blue) and Ground Truth(in red) for Stacked Ensemble

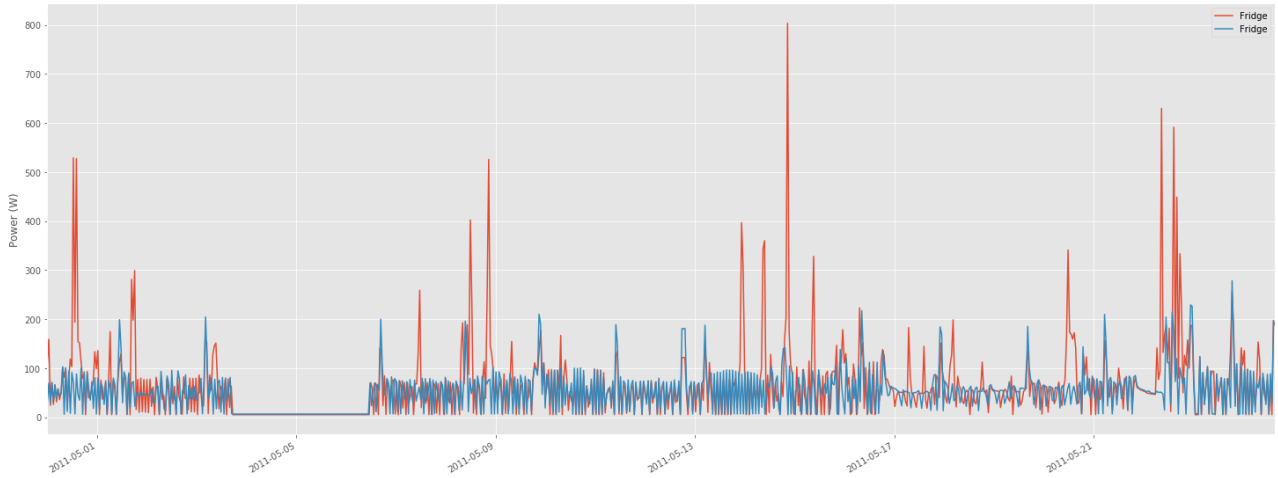

Fig. 9. Predictions(in blue) and Ground Truth(in red) for GRU

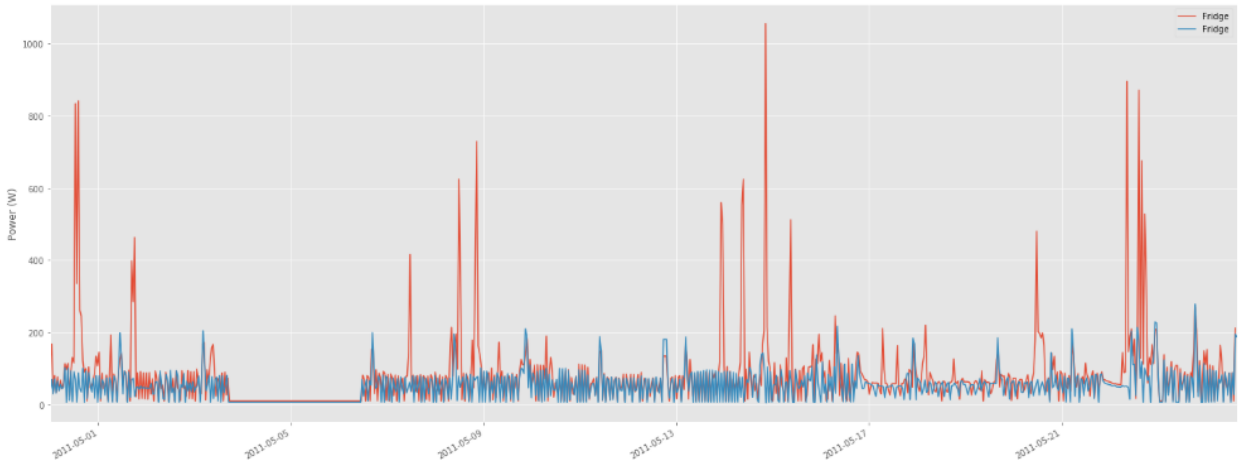

Fig. 10. Predictions(in blue) and Ground Truth(in red) for RNN

	Appliance Transfer Learning - Stacked Ensemble					
Appliance	**Recall**	**Precision**	**Accuracy**	**F1-score**	**MAE**	**RMSE**
Lights	0.28	0.88	0.33	0.43	56.51	95.08
Sockets	0.28	1	0.28	0.44	53.74	94.52
Microwave	0.29	0.75	0.4	0.41	55.09	117.25
Dishwasher	0.28	0.75	0.39	0.4	58.7	121.08

Fig. 11. Appliance Transfer Learning - Training Data (Fridge)

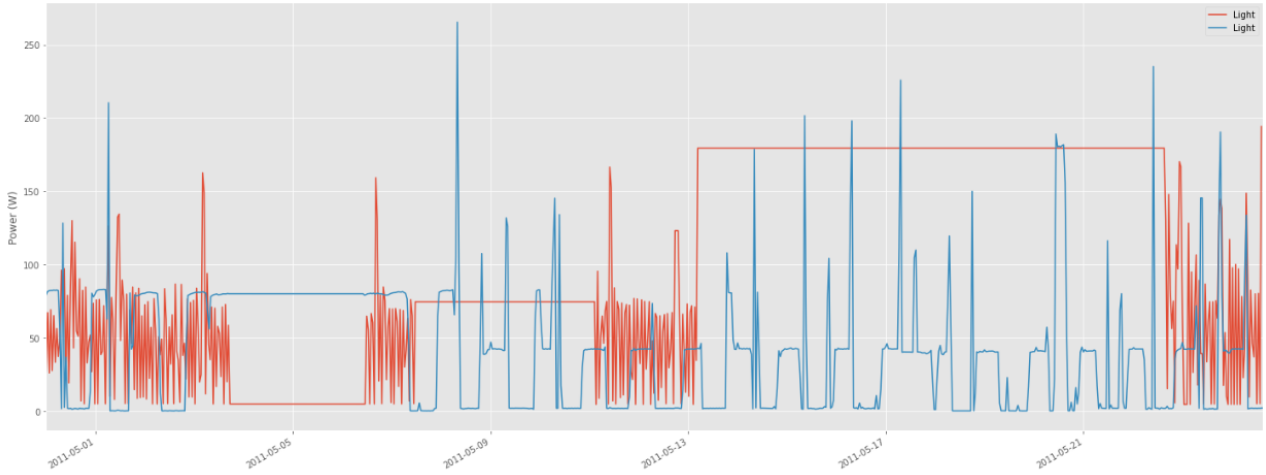

Fig. 12. Appliance Transfer Learning - Training Data(Fridge), Testing Data(Light)

	Cross Domain Transfer Learning - Stacked Ensemble					
Appliance	**Recall**	**Precision**	**Accuracy**	**F1-score**	**MAE**	**RMSE**
Fridge	0.28	0.74	0.39	0.41	54.37	108.05

Fig. 13. Cross Domain Transfer Learning - Training Data (Fridge)

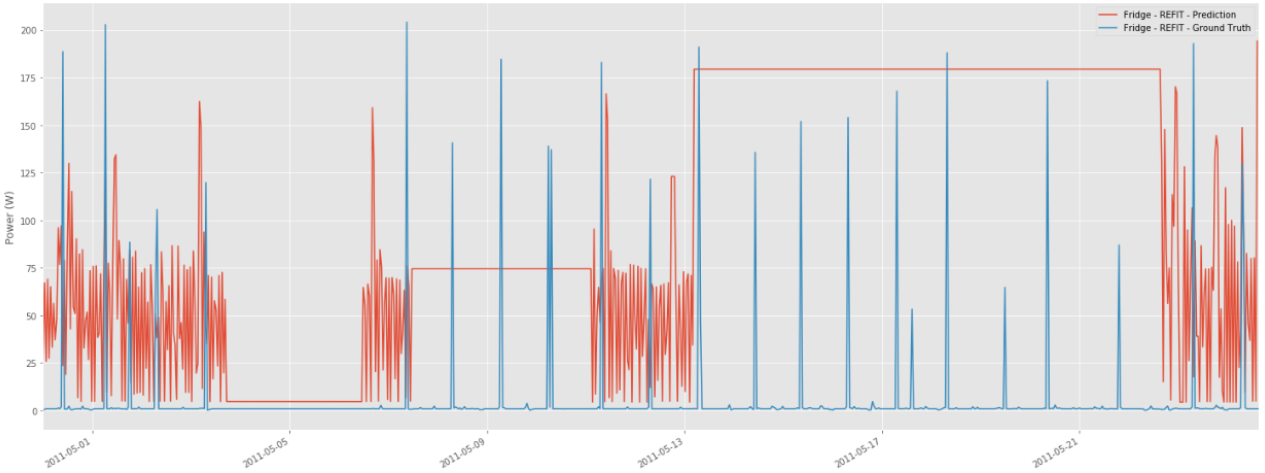

Fig. 14. Cross Domain Transfer Learning - Training Data(Fridge - REDD), Testing Data(Fridge - REFIT)

model gives a lower RMSE of 105.99 for fridge whereas RNN and GRU give a higher RMSE of

172.67 and 128.96 respectively. The tables in Fig. 6 and 7 show the classification performance for stacked model against the RNN and GRU models for 3 epochs.

The plots in Fig. 8, 9, 10 indicate the predictions for fridge for all 3 models.

It is evident from the plots that the predictions from stacked ensemble model fits the ground truth data more accurately than DAE and RNN. It is able to catch the periodic spikes in data and predicts accordingly. Considering the length of this document and for a better view of comparison, we are only displaying the plot results for fridge.

E. Transfer Learning

For the purpose of transfer learning, we had 2 tasks in hand. We implemented transfer learning by training on **Fridge** and testing on other appliances (in case of Appliance Transfer Learning) or testing on Fridge test set from REFIT database (in case of Cross Domain Transfer Learning).

1) Appliance Transfer Learning: In appliance transfer learning, we train our model on the training data of the one appliance, say appliance 'A' and then test it on the other appliance, say appliance 'B' and see if we can get good results. If this can give us good results we can save a lot of energy and computation power while training only for the some appliance and training for the rest. We only need to install sensors to get the data for some appliances. From literature survey, we know that the appliance which has high energy needs is most suited for the training model and we can use it for prediction for the low energy needs appliance. In our case, we are using **Fridge** as our training appliance and we will test the result on other appliances.

From the metric values in Fig. 11, we can see that the Training data - Fridge gives considerable result for all the 4 appliances. When we compare them with their metric values we got by training on the appliance's own dataset, we saw that the values are convincing. Surprisingly, for microwave the RMSE actually reduced when we used Fridge as the training dataset.

2) Cross Domain Transfer Learning: In the Cross Domain Transfer Learning, we want to train our model on the training set of appliance of one dataset and then test on the same appliance with data from another dataset. For this purpose we will be using Fridge as our appliance and we will train the data from REDD database and then test the data from REFIT database. We will try to infer our results from both the datasets.

The values in Fig. 13 are convincing but it is higher than what we achieve when we run on the test dataset of fridge appliance from REDD. This can happen mainly due to the fact that the databases are located in geographically different places. **REDD** database is readings of the database from US and **REFIT** database is readings from UK. The way measurement is done can differ and the type of sensors or difference in precision reading may have affected the result by large. If in future experiments, we are able to get good results, it can help in preserving user's privacy and save a lot of cost to install sensors all across the globe.

VI. Conclusion and Future Work

In this work, we introduced a stacked ensemble deep learning model to accurately predict the energy disaggregation. Despite its naivety, the model provided reasonably good results as compared to the models used till date. We also showed that stacked ensemble performs well for Appliance Transfer Learning and Cross Domain Transfer Learning too. The stacked models behave inherently better than others due to their advantage of keeping the better predictions and discarding the useless ones in successive epochs. Such predictions will go a long way in optimizing the energy consumption and ensuring user's privacy.

Future work on this includes designing a better stacked model along with appropriate hyperparameter tuning and training for more number of epochs. Since we had limited resources, we could only train our model for 3 epochs due to the large time taken for training. Using GPUs to train the model can further reduce the training time.

Task	Contributors
Literature Survey	Adnan Vasanwalla, Vishal Sarda, Abhiraj Smit
Brainstorming different ideas	Vishal Sarda, Adnan Vasanwalla, Abhiraj Smit
Designing the stacked model architecture	Vishal Sarda, Adnan Vasanwalla
Exploratory Data Analysis	Adnan Vasanwalla, Vishal Sarda, Abhiraj Smit
Implementing Stacked Ensemble	Vishal Sarda, Adnan Vasanwalla
ImplementingTransfer Learning	Abhiraj Smit

Fig. 15. Distributions of tasks

VII. Contributions

Group Members: Adnan Vasanwalla, Vishal Sarda, Abhiraj Smit

The distribution of work is shown in Fig. 15.

References

[1] Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey, Ahmed Zoha, et. al

[2] Object recognition from local scale-invariant features, D.G. Lowe

[3] Neural NILM: Deep Neural Networks Applied to Energy Disaggregation, Jack Kelly, et. al

[4] Nonintrusive Appliance Load Monitoring, George Hart

[5] Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation, Kolter et. al

[6] A Fully Unsupervised Non-intrusive Load Monitoring Framework, Ruoxi et. al

[7] Toward a Semi-Supervised Non-Intrusive Load Monitoring System for Event-based Energy Disaggregation, Karim et. al

[8] Energy Disaggregation for NILM applications using Shallow and Deep Networks, Laksmi et. al

[9] Neural NILM: Deep Neural Networks Applied to Energy Disaggregation, Jack Kelly et. al

[10] NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring, Nipun Batra et. al

[11] Towards reproducible state-of-the-art energy disaggregation, Nipun Batra et. al

[12] Transferability of Neural Network Approaches for Low-rate Energy Disaggregation, David et. al

[13] An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study, David et. al

[14] The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes, Jack Kelly el. al

[15] REDD: A Public Data Set for Energy Disaggregation Research, Kolter et. al