Investigating the Effect of Intraclass Variability in Temporal Ensembling

Siddharth Vohra1, Manikandan Ravikiran2
1University of California, San Diego
2Research & Development Center, Hitachi India Pvt Ltd., Bangalore, India
sivohra@ucsd.edu
manikandan@hitachi.co.in

Abstract
Temporal Ensembling is a semi-supervised approach that allows training deep neural network models with a small number of labeled images. In this paper, we present our preliminary study on the effect of intraclass variability on temporal ensembling, with a focus on seed size and seed type, respectively. Through our experiments we find that (a) there is a significant drop in accuracy with datasets that offer high intraclass variability, (b) more seed images offer consistently higher accuracy across the datasets, and (c) seed type indeed has an impact on the overall efficiency, where it produces a spectrum of accuracy both lower and higher. Additionally, based on our experiments, we also find KMNIST to be a competitive baseline for temporal ensembling.

1 Introduction
Deep neural networks have seen broad applications across vision speech and language in recent times. Yet this success is contingent on acquiring a large number of labeled datasets, which is expensive and time-consuming. Further, labeling is mostly manual, done by humans, due to its higher meticulousness. Recently, to address this concern of manual labeling variety of approaches have been designed, including Semi-Supervised learning algorithms (Gordon and Hernández-Lobato, 2017; Laine and Aila, 2017; Lee, 2013), which typically proffer with higher results with a small number of labeled examples (seeds). Notable among this is the Temporal Ensembling (Laine and Aila, 2017), which uses an ensemble of the earlier outputs of a neural network as an unsupervised target label and achieved high accuracy on SVHN and CIFAR-10 with just 500 and 4000 labeled samples with both naturally offering lower intraclass variances. Besides, to the best of our knowledge, there is no explicit study of temporal ensembling in the context of datasets with large intraclass variability. As such, in this work, we attempt to investigate this gap by answering the following research questions.

– RQ1: Does intraclass variability impact the accuracy of temporal ensembling? Here the intention is to check (a) how accuracy varies and (b) if there is any unique observable behavior with temporal ensembling under different intraclass variances. The assumption here being that the intraclass variability is a spectrum with a range of low to high intraclass variation. To this end, we experiment on datasets of Fashion-MNIST (Xiao et al., 2017) and KMNIST (Clanuwat et al., 2018) to find that there is a sheer drop in performance when using temporal ensembling.
– RQ2: Under settings of intraclass variability, how does seed size impact temporal ensembling?. Here we hypothesize and verify that upon increasing seed size, there is an improvement in performance.
– RQ3: What’s the effect of seed selection on temporal ensembling? More specifically, we see if the diversity of seeds have an impact on results. The preliminary experimental results show that performance is lower with some category of seeds over others.

The rest of the paper is organized as follows. In section 2, we review related research in semi-supervised learning. In section 3 we introduce the temporal ensembling approach in brief. In section 4 we present dataset, experimental setup, and answer each of the research questions with analysis in section 5. Finally, we conclude in section 6 with possible implication on future works.

*Work performed while interning at Hitachi R&D India
2 Related Work

Semi-Supervised learning has seen an extensive assortment of works originating with pseudo-labeling (Lee, 2013), which assigns pseudo-labels to unlabelled data using the model forecasts at each epoch and training on both labeled and pseudo-labeled data. Then there are variational autoencoders (Kingma and Welling, 2014), which uses a generative deep learning model (Kingma et al., 2014) for semi-supervised learning. Similarly, there are Ladder Networks (Rasmus et al., 2015), which uses noise to the input of each layer in the neural network and combines it with denoising function. Ladder network combines denoising loss with the supervised loss during final training. More recently, there is Virtual Adversarial Training (VAT) (Miyato et al., 2015), which uses perturbations generated by adversarial learning. Temporal Ensembling (Laine and Aila, 2017), which is like pseudo-labeling, except it matches the output of previous models and Mean-Teacher (Tarvainen and Valpola, 2017), which addresses the drawback of the storage requirement of temporal ensembling by using an exponential moving average. There are many other notable works, including works on co-training and feature learning (Blum and Mitchell, 1998; Sindhwani et al., 2005), and graph-based methods (Ng and Silva, 2018), metric-learning based methods (Yu et al., 2017) and sampling noisy labels (Vahdat, 2017; Ravikiran et al., 2020). In this work, we concentrate on temporal ensembling due to its simplicity and study it in the context of datasets with diverse intraclass variability.

3 Temporal Ensembling

Temporal Ensembling is an enhancement of the Π-model (Laine and Aila, 2017) and is based on the idea of self-ensembling where it ensembles earlier outputs as an unsupervised target (See Equations 1-3). This process of self-ensembling is analogous to that of pseudo-labels. More specifically, in temporal ensembling training is done with dataset under different augmentations and dropout regularization, which presents the network to acquire noise-invariant features.

\[
l_B(z, \tilde{z}, y) = \text{masked crossentropy}(z, y) + w(t) \ast \text{MSE}(z, \tilde{z})
\]

\[
\text{masked crossentropy}(z, y) = -\frac{1}{|B \cap L|} \sum_{i \in (B \cap L)} \log z_i[y_i]
\]

\[
\text{MSE}(z, \tilde{z}) = \frac{1}{C|B|} \sum_{i \in B} || z_i - \tilde{z}_i ||^2
\]

This, in turn, allows the neural network to not shift its predictions in slightly modified variants of the same inputs. After every training epoch, the ensemble outputs refresh with the current new prediction and the previous ensemble prediction through the exponential moving average (EMA) method shown in Equation 4. Also, to generate the input training target \( \tilde{z} \) as shown in equation 5, temporal ensembling corrects the startup bias of \( Z \). More details on temporal ensembling can be found in (Laine and Aila, 2017) and (Ferret, 2018).

\[
Z = \alpha Z + (1 - \alpha)z
\]

\[
\tilde{z} = \frac{Z}{1 - \alpha^t}
\]

4 Experimental Setup

To answer the research questions from section 1 in this work, we compare and contrast performance of temporal ensembling on various datasets under varying settings of labeled samples. This section provides a brief overview of the datasets (section 4.1) and parameter settings (section 4.2) used.
4.1 Dataset
For this work, we use the datasets, which are presented in Table 1. More details of datasets are presented inplace across sections 5.1-5.3 when necessary.

| Dataset       | Image Size | Train Size | Test Size | Description                                      |
|---------------|------------|------------|-----------|--------------------------------------------------|
| MNIST         | 28x28      | 60k        | 10k       | Handwritten characters of 0-9                    |
| KMNIST        |            |            |           | Handwritten Kuzushiji Characters                 |
| Fashion-MNIST |            |            |           | Clothing images from Zolando.com                 |

Table 1: Summary of dataset and its characteristics used experiments 5.1-5.3

4.2 Parameter Settings
The various settings employed to answer each of the RQ’s are as described. Besides, all the RQ’s employ standard settings for most of the parameters, as shown in Table 2 below.

- **RQ1**: For analysis of RQ1, we train and test temporal ensembling models with all the three datasets, under common parameter settings from Table 2 with 300 and 500 epochs respectively.
- **RQ2**: For RQ2, we use parameters from Table 2 and vary labeled examples in range 100-500.
- **RQ3**: For RQ3, in addition to setting mentioned in Table 2, we run experiments with 10 different randomly sampled seeds.

| Hyperparameters          | Values |
|--------------------------|--------|
| Dropout                  | 0.5    |
| Standard Deviation       | 0.15   |
| Feature maps in Conv.1   | 16     |
| Feature maps in Conv.2   | 32     |
| Weight Normalization     | True   |
| Learning Rate            | 0.002  |
| Beta                     | 0.99   |
| Batch Size               | 100    |
| Alpha                    | 0.6    |
| Data Normalization       | channel wise |

Table 2: Hyperparameter settings used for experiments in sections 5.1-5.3

5 Experiments
5.1 RQ1: Performance of Temporal Ensembling under Intraclass Variability
The first research question asks how *intraclass variability influences the accuracy using temporal ensembling*. To answer this, the accuracy score is reported on the datasets of MNIST, KMNIST, and Fashion-MNIST, respectively. The reason behind selecting these datasets is that all of them are grayscale with the same image size and proportion of images, but KMNIST and Fashion-MNIST offer wide intraclass variability. None of these datasets require any size normalization, thus making results comparable across the datasets.

Besides, as mentioned in section 4.2, here we trained temporal ensembling for 100, 300, and 500 epochs, respectively. The consolidated results so obtained are as shown in Table 3. Also, we ran five episodes of training with varying seed sampling. Hence the results presented in Table 3 are averaged with a net standard deviation of accuracy. Detailed episode level results are in section 7.1. Comparing the results of different datasets, we can see that MNIST gives the highest accuracy, followed by Fashion-MNIST.
| Epochs | 100       | 300       | 500       |
|--------|-----------|-----------|-----------|
| MNIST  | 93.27±2.104 | 97.21±0.808 | 97.43±1.069 |
| KMNIST | 50.05±20.07   | 49.90±20.20   | 60.66±3.700  |
| Fashion-MNIST | 74.14±1.580 | 73.95±0.975  | 71.31±2.443  |

Table 3: Accuracy of Temporal Ensembling on MNIST, KMNIST and Fashion-MNIST. All the results are with seed size of 100 and five episodes of training.

Figure 1: Training loss behavior of MNIST using Temporal Ensembling and KMNIST, respectively. Besides, MNIST has faster convergence and stagnation in the loss (See Figure 1), while KMNIST (Figure 2) and Fashion-MNIST (Figure 3) show delay in convergence time and increase in loss values with higher epochs for both supervised and unsupervised components. We believe such behavior is due to two reasons, namely (i) similarity (low variance) in train-test samples and samples within same classes for MNIST compared to KMNIST and Fashion-MNIST and (ii) unsupervised labels getting biased towards one of the classes, and no updates in the subsequent iterations of temporal ensembling. However, more experiments are warranted to validate both (i) and (ii).

Further, from Figure 1 and 2, we can see that such behavior is consistent across episodes with different seed samples for both KMNIST and Fashion-MNIST. Besides, the results with KMNIST and Fashion-MNIST with some of the seeds differ by a large margin (See Tables 8-10 in Appendix). Overall, analysis of temporal ensembling here shows that it is less tuned for datasets that have higher intraclass variances and might, therefore, have higher performances across varying datasets that do not match such criteria. Overall to summarize, our findings are as follows.

- Accuracy score is highest for MNIST (97.21%), and among datasets that offer intraclass variance, there is a sheer drop in performance with KMNIST producing 60.66% and Fashion-MNIST showing 71.31%. We believe such a drop is due to intraclass variances and variance between train-test samples, both of which require further analysis.

- Unlike MNIST, both KMNIST and Fashion-MNIST present with a unique behavior of lack of
Figure 2: Training loss behavior of KMNIST using Temporal Ensembling

Figure 3: Training loss behavior of Fashion-MNIST using Temporal Ensembling
convergence for unsupervised loss and instead show an increase at higher epochs. This raises a question that the unsupervised labels may be biased towards a specific set of class(es) and aren’t changing across iterations. At the same time, it can also be conjectured that this could be due to the selection of learning rate or any other components of the temporal ensembling algorithm. Validating these requires a more detailed analysis of temporal ensembling and its components.

• Temporal ensembling is less suitable for datasets that offer higher intraclass variances, as evident from results. As such, it is not directly usable despite the similarity in size and proportion of datasets.

5.2 RQ2: Effect of Seed Size on Temporal Ensembling under Intraclass Variability

Previously in section 5.1, we analyzed the impact of intraclass variance on temporal ensembling by studying performance on the MNIST, KMNIST, and Fashion-MNIST datasets. However, for all the experiments in section 5.1 we maintained a uniform seed size of 100 labeled samples. Besides, in Figures 2 and 3 we saw that seeds indeed have an impact on the overall results. As such, in this section, we shall investigate the effect of seed size, i.e., the number of labeled samples. More specifically, for RQ2, we vary seed size in the range of 100-500 and analyze the performance of models trained for both 300 and 500 epochs, respectively.

| Seed Size | 100       | 200       | 300       | 400       | 500       |
|-----------|-----------|-----------|-----------|-----------|-----------|
| MNIST     | 97.21±0.8082 | 96.93±1.62 | 97.82±0.2128 | 97.41±3.285 | 97.37±0.2679 |
| KMNIST    | 49.90±20.20  | 66.79±3.501 | 70.02±1.9912 | 72.24±2.3667 | 75.35±1.5323 |
| Fashion-MNIST | 73.95±0.9755 | 76.71±1.524 | 78.7±1.1386   | 80.40±0.6780 | 81.53±1.036   |

Table 4: Accuracy of Temporal Ensembling on MNIST, KMNIST and Fashion-MNIST with training for 300 epochs and 5 episodes.

| Seed Size | 100       | 200       | 300       | 400       | 500       |
|-----------|-----------|-----------|-----------|-----------|-----------|
| MNIST     | 97.43±1.0698 | 97.53±0.7756 | 97.71±0.2838 | 97.93±0.1158 | 97.35±0.5745 |
| KMNIST    | 60.66±3.700  | 69.99±1.4461 | 70.11±3.0019 | 74.27±1.2281 | 75.36±1.6777 |
| Fashion-MNIST | 71.31±2.4430 | 76.81±0.8309 | 79.92±0.9292   | 80.68±0.9838 | 80.40±0.8048   |

Table 5: Accuracy of Temporal Ensembling on MNIST, KMNIST and Fashion-MNIST with training for 500 epochs and 5 episodes.

Accuracy score with standard deviation over five episodes of training each is shown in Tables 4 and 5. Detailed experimental results are in section 7.2. From Tables 4 and 5, we can see that for MNIST datasets, the number of seeds indeed has no impact on the overall results with minor changes across varying sizes when tested at both 300 and 500 epochs, respectively. In fact, majority of the results are around 97% with a maximum of 97.93% when trained for 500 epochs, 400 labeled samples (0.5% higher than highest results from Table 1) and a minimum of 96.93% with 300 epochs, 200 labeled samples (0.5% lower than best results from Table 1).

Again this could be argued due to the nature of the dataset, where there is less non-uniformity (lower intraclass variability) within the classes, and train-test splits, and adding more seeds has minimal impact on the network learning behavior. Figure 4 and 5 also show loss curves for seed sizes of 300 and 500, trained for 500 epochs. As we can see, even with higher seeds, the convergence behavior seems very similar to RQ1 irrespective of training epochs. Further, this moderately supports the hypothesis that the number of seeds is a vital hyperparameter, and the performance of temporal ensembling indeed could be controlled by selecting an optimal amount of seeds. This also begs the question of rather than selecting the number of seeds, is it possible to select seeds that favor the test data distribution, which we shall revisit in section 5.3.

From Tables 4 and 5 for KMNIST and Fashion-MNIST, we can see that results improve drastically with the addition of more seeds. To start with for KMNIST, the improvement in the number of seeds
Figure 4: Training loss behavior of MNIST using Temporal Ensembling with 300 Seeds

Figure 5: Training loss behavior of MNIST using Temporal Ensembling with 500 Seeds
Figure 6: Training loss behavior of KMNIST using Temporal Ensembling with 300 Seeds

Figure 7: Training loss behavior of KMNIST using Temporal Ensembling with 500 Seeds
from 100 to 500 improves the result by 25% with 300 epochs and 14.7% with 500 epochs to result of 75%. Meanwhile, Figure 6 and 7 show the loss curves with 300 and 500 seeds respectively across the five episodes of training under each epoch settings. As we can see with higher seeds, the network training behavior is indeed different, where we see lower values of both supervised and unsupervised loss. Further,
Figure 9: Training loss behavior of Fashion-MNIST using Temporal Ensembling with 300 Seeds

Figure 10: Training loss behavior of Fashion-MNIST using Temporal Ensembling with 500 Seeds
such behavior is constant across various episodes of training. Examples from 300 and 500 seeds used during the preparation of KMNIST is as shown in Figure 8.

Fashion-MNIST again shows similar behavior like KMNIST except with net improvement of 8%, with a maximum 81.53% with 500 seeds. Yet the loss curves (Figures 9 & 10) and seeds (Figure 11) show similar behavior like KMNIST. What is very interesting to see is that performance of Fashion-MNIST is

Figure 11: Examples Fashion-MNIST seeds used in experiments.
very close to the two-layer convolution baseline (TLCB) of 87% (Table 6), which is trained with complete 60k images. So even though temporal ensembling uses 150% lesser images than TLCB, the performance difference is only 7% lower. Similarly, for KMNIST, we can see that performance 17% lower than the simplest K-Nearest Neighbor baseline of 92.1% without any feature engineering (Table 6).

| Dataset      | Approach                           | Seed Size | Results  |
|--------------|------------------------------------|-----------|----------|
|              | K-Nearest Neighbors Baseline       | 60k       | 92%      |
| KMNIST       | Temporal Ensembling (500 epochs)   | 500       | 75.35%   |
|              | Two-Layer Convolution Neural Network | 60k       | 87%      |
|              | Temporal Ensembling (500 epochs)   | 500       | 81.53%   |

Table 6: Comparison of Results from Temporal Ensembling with Varying Seed Size (this work) against respective baselines for KMNIST and Fashion-MNIST with 60k images.

These two results confirm that there is some correlation between test data and seed size. Besides, this also shows that KMNIST to be a competitive baseline for temporal ensembling compared to MNIST and Fashion-MNIST in the context of intraclass variances. Also, these results strengthen our initial argument from section 1, that there is a need for experimental studies, such as those involving intraclass variability in semi-supervised learning approaches to find out what, apart from the number of labeled samples, what other aspects impacts performance. To summarise, our findings are:

- Seed size drastically impacts results under the conditions of intraclass variability, wherewith larger seeds size produces better results.
- KMNIST and Fashion-MNIST show an improvement of 14.7% and 8% respectively with an increase in seed size from 100 to 500, producing competitive results again baselines trained on complete 60k images. While the net improvement obtained vary across both the datasets, the trend is consistent.
- MNIST is a simple dataset, and from results, we can see that its not a strong baseline for temporal ensembling under the context of intraclass variability, which in turn could be fulfilled by KMNIST.
- Besides, from results, we argue that there is a need for more detailed studies on the effect of intraclass variability in the context of general semi-supervised learning.

5.3 RQ3: Effect of Seed Type on Temporal Ensembling under Intraclass Variability

Previously in section 5.2, we saw that with increasing seed size, the accuracy improved for all the datasets, and in the case of KMNIST and Fashion-MNIST, we observed substantial gains. Here we will examine how the type of seed will impact the results. This is similar to the argument of generalization in neural networks, where the trained model should at least correctly classify images identical to the one it was prepared for. However, in the context of temporal ensembling, the labeled examples serve two purposes, namely improving the quality of self-ensembled labels and, at the same time, also improve accuracy.

As such, one can hypothesize that starting with seeds with distributions closer to test set samples may achieve higher results and vice versa. Previous works have shown that all the datasets from section 4.1 have examples of those similar to that of train distribution, and some are profoundly different. As such, to answer this RQ, we repeat experiments in line with RQ2, except we perform training with ten episodes with 300 seeds and 500 epochs. Also, to account for training randomness on the accuracy, each of the models was tested at every epoch. Finally, the model from the 500th epoch was chosen due to its highest accuracy. Consolidated results with standard deviation over ten episodes are as shown in Table 7.

Comparing Table 7 with results from RQ2 (Table 5), with different samples of seeds, we can see that we have both higher and lower results than those from RQ2. Similar to findings from RQ1 and RQ2, MNIST produces results closer to 97% despite random sampling of seeds across episodes. Further, the lowest effect of 96.95% and the highest of 98.03% could be observed. Figure 14 shows, convergence behavior of both the models. As we can see, both the models converge similarly and inline with RQ1 and RQ2.

\[1\] Please note that, the results for RQ3 is still preliminary and requires further analysis & validation.
| Training Episode | MNIST   | KMNIST  | Fashion-MNIST |
|------------------|---------|---------|---------------|
| #1               | 96.95%  | 72.28%  | 77.94%        |
| #2               | 97.03%  | 77.40%  | 80.05%        |
| #3               | 98.24%  | 60.79%  | 78.13%        |
| #4               | 98.58%  | 74.69%  | 78.15%        |
| #5               | 98.09%  | 74.06%  | 79.62%        |
| #6               | 98.14%  | 74.58%  | 79.25%        |
| #7               | 98.30%  | 68.70%  | 80.01%        |
| #8               | 97.64%  | 67.30%  | 76.87%        |
| #9               | 97.96%  | 66.86%  | 79.98%        |

Table 7: Results of MNIST, KMNIST and Fashion-MNIST trained for 500 epochs, 300 seeds across 10 episodes of training with different seed samples.

Figure 12: Training loss behavior of MNIST trained for 10 episodes with varying seeds. Highest and Lowest models corresponds to seeds seed_14 & seed_102 respectively.

Meanwhile, KMNIST and Fashion-MNIST show widely varying results (See Table[7]). To begin with, KMNIST produces the lowest accuracy of 66.86% and the highest of 77.40%. The lowest accuracy of KMNIST is closer to results obtained with 100 seeds, which in turn suggests that some of the seeds selected were practically useless. Similar behavior could be observed with the Fashion-MNIST dataset with the highest accuracy of 80.05 and lowest of 77.94%, where again, the lowest result is close to that of results with 200 seeds in Table[5].

The fact that the performance of some experiments with varying seeds is significantly lower than the other would raise a question that the models may have to overfit, and results obtained could be lower. However, as mentioned earlier, this was avoided by checking for model accuracy at each epoch. Also, if this is indeed the cases, then it, in turn, raises a new question of if there are samples that would allow faster and better convergence. Overall from these preliminary experiments
Figure 13: Training loss behavior of KMNIST trained for 10 episodes with varying seeds. Highest and Lowest models corresponds to seeds seed\_179 & seed\_92 respectively. See Figures 15 & 16 in appendix.

- We find that seed selection indeed impacts the results across all the datasets. For the cases of KMNIST, we can see that the highest and lowest results obtained with different seeds differ by 2% and 9%, respectively. Similarly, for Fashion-MNIST, the numbers are 0.5% and 3%, respectively.
- Performance with some of the seeds is significantly lower, which suggests some of the seeds offer more to training than others, which points to the identification of optimal seed images for temporal ensembling.
- We further emphasize that the results shown here require further in-depth experimentation, analysis of results, and theoretical grounding\(^1\).

6 Conclusion and Future Work

This paper investigated the effect of intraclass variability on temporal ensembling. Firstly by analyzing different datasets, we showed that the dataset differs widely in terms of their intraclass variations. KMNIST offers the highest intraclass variability, followed by Fashion-MNIST and MNIST. From our study on RQ1, we find that given constant parameter settings, intraclass variability indeed affects the overall performance with KMNIST producing 60.66% and Fashion-MNIST showing 71.31%. Further, we also found that KMNIST and Fashion-MNIST present with a problem of lack of convergence of unsupervised loss and instead show an increase at higher epochs. Overall, our study of RQ1 suggests that temporal ensembling is not directly usable for datasets with high intraclass variability.

From our review of the effect of seed size in RQ2, we see that higher seed size results in better accuracy across all the datasets KMNIST and Fashion-MNIST show an improvement of 14.7% and 8% respectively with an increase in seed size from 100 to 500, producing competitive results again baselines trained on complete 60k images. Further, we could see that KMNIST serves as a competitive baseline for temporal ensembling as it accounts for intraclass variability. Finally, in RQ3, we demonstrate that with different seed images, we get different results.
Overall across a broad range of datasets, we examined how intraclass variability temporal ensembling performs. While there is considerable variation within the classes across the datasets, it is consistent that the class with more intraclass variability is harder to do classify. This connection with intraclass variability occurs to such an extent that, in fact, temporal ensembling with 1/3 of the selected seeds offers similar accuracy (See section 5.3). However, at the same time, this result serves as a decent baseline. We believe seeds are critical, and the temporal ensembling with different seeds is not sufficiently effective at generalizing beyond the image contexts found in training data. To close this gap and advance temporal ensembling for practical applications various aspects need more in-depth exploration including (a) effect of varying other hyperparameters (Table 4.2) in temporal ensembling, (b) relationship between seed types and data distribution and (c) reason for the rise in the unsupervised loss at higher epochs.

References

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In COLT’ 98.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep learning for classical japanese literature. ArXiv, abs/1812.01718.

Johan Ferret. 2018. Blog on Semi-supervised image classification via Temporal Ensembling.

Jonathan Gordon and José Miguel Hernández-Lobato. 2017. Bayesian semisupervised learning with deep generative models. arXiv: Machine Learning.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. CoRR, abs/1312.6114.

Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In NIPS.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. ArXiv, abs/1610.02242.
Acknowledgements
We would like to thank Johan Ferret (Ferret, 2018), for his pytorch 0.4 implementation of temporal ensembling.

Author Contributions
MR designed the study. SV carried out the primary experiments. Both the authors carried out analysis, secondary experiments, contributed to developing the outline and editing the manuscript. MR wrote the manuscript and SV reviewed the manuscript.

7 Appendix
We now present various experimental details and results for each of the RQs.

7.1 Detailed Experimental Results of RQ1
Tables 8-10 present detailed results for RQ1.

7.2 Detailed Experimental Results of RQ2
Tables 11-14 shows experimental results obtained with varying seeds from 200 to 500 respectively.

7.3 Seed Images used in experiments of RQ3
Figures 15-16 shows seeds that produce highest and lowest results for MNIST datasets in RQ3.
| Dataset | Epochs | Experiment | Accuracy | Accuracy (Best Model) |
|---------|--------|------------|----------|----------------------|
| MNIST   | 100    | 1          | 95.78%   | 95.78%               |
|         |        | 2          | 93.86%   | 91.37%               |
|         |        | 3          | 95.23%   | 90.59%               |
|         |        | 4          | 95.04%   | 93.11%               |
|         |        | 5          | 95.66%   | 95.51%               |
|         |        | **Average**| 95.11%   | 93.27%               |
|         | 300    | 1          | 95.89%   | 95.70%               |
|         |        | 2          | 96.98%   | 97.13%               |
|         |        | 3          | 97.86%   | 98.05%               |
|         |        | 4          | 97.47%   | 97.56%               |
|         |        | 5          | 97.45%   | 97.60%               |
|         |        | **Average**| 97.13%   | 97.21%               |
|         | 500    | 1          | 98.15%   | 98.16%               |
|         |        | 2          | 95.58%   | 95.33%               |
|         |        | 3          | 97.59%   | 97.61%               |
|         |        | 4          | 98.02%   | 98.18%               |
|         |        | 5          | 97.78%   | 97.86%               |
|         |        | **Average**| 97.42%   | 97.43%               |

Table 8: Detailed Experimental results for RQ1 on MNIST

| Dataset | Epochs | Experiment | Accuracy | Accuracy (Best Model) |
|---------|--------|------------|----------|----------------------|
| KMNIST  | 100    | 1          | 59.20%   | 58.64%               |
|         |        | 2          | 58.27%   | 58.55%               |
|         |        | 3          | 59.41%   | 61.77%               |
|         |        | 4          | 10.00%   | 9.99%                |
|         |        | 5          | 62.41%   | 61.31%               |
|         |        | **Average**| 49.86%   | 50.05%               |
|         | 300    | 1          | 53.45%   | 55.98%               |
|         |        | 2          | 10.00%   | 10.00%               |
|         |        | 3          | 57.84%   | 64.69%               |
|         |        | 4          | 57.09%   | 61.97%               |
|         |        | 5          | 47.84%   | 56.85%               |
|         |        | **Average**| 45.24%   | 49.90%               |
|         | 500    | 1          | 50.90%   | 56.49%               |
|         |        | 2          | 50.18%   | 60.35%               |
|         |        | 3          | 49.94%   | 63.86%               |
|         |        | 4          | 50.39%   | 55.26%               |
|         |        | 5          | 53.52%   | 64.32%               |
|         |        | **Average**| 50.99%   | 60.06%               |

Table 9: Detailed Experimental results for RQ1 on KMNIST
| Dataset      | Epochs | Experiment | Accuracy | Accuracy (Best Model) |
|-------------|--------|------------|----------|-----------------------|
| Fashion-MNIST | 100    | 1          | 75.94%   | 75.28%                |
|             |        | 2          | 72.82%   | 73.00%                |
|             |        | 3          | 72.31%   | 73.28%                |
|             |        | 4          | 73.74%   | 72.48%                |
|             |        | 5          | 75.55%   | 76.67%                |
|             |        | Average    | 74.07%   | 74.14%                |
|             | 300    | 1          | 68.61%   | 73.14%                |
|             |        | 2          | 73.02%   | 74.47%                |
|             |        | 3          | 69.86%   | 73.43%                |
|             |        | 4          | 74.49%   | 75.63%                |
|             |        | 5          | 71.31%   | 73.09%                |
|             |        | Average    | 71.46%   | 73.95%                |
|             | 500    | 1          | 70.69%   | 72.46%                |
|             |        | 2          | 66.58%   | 69.86%                |
|             |        | 3          | 66.99%   | 73.41%                |
|             |        | 4          | 65.74%   | 67.21%                |
|             |        | 5          | 70.08%   | 73.59%                |
|             |        | Average    | 68.02%   | 71.31%                |

Table 10: Detailed Experimental results for RQ1 on Fashion-MNIST
| Dataset     | Epochs | Experiment | Accuracy   | Accuracy (Best Model) |
|-------------|--------|------------|------------|-----------------------|
| MNIST       | 100    | 1          | 97.57%     | 97.48%                |
|             |        | 2          | 95.30%     | 95.67%                |
|             |        | 3          | 96.77%     | 96.77%                |
|             |        | 4          | 96.64%     | 96.51%                |
|             |        | 5          | 95.40%     | 94.85%                |
|             |        | Average    | 96.38%     | 96.26%                |
|             | 300    | 1          | 97.59%     | 97.23%                |
|             |        | 2          | 98.09%     | 98.14%                |
|             |        | 3          | 97.93%     | 98.00%                |
|             |        | 4          | 93.86%     | 93.75%                |
|             |        | 5          | 97.16%     | 97.53%                |
|             |        | Average    | 96.93%     | 96.93%                |
|             | 500    | 1          | 97.63%     | 97.63%                |
|             |        | 2          | 97.12%     | 96.90%                |
|             |        | 3          | 97.74%     | 98.24%                |
|             |        | 4          | 95.70%     | 96.42%                |
|             |        | 5          | 97.75%     | 98.46%                |
|             |        | Average    | 97.19%     | 97.53%                |
| KMNIST      | 100    | 1          | 63.97%     | 64.32%                |
|             |        | 2          | 69.24%     | 69.44%                |
|             |        | 3          | 66.85%     | 64.85%                |
|             |        | 4          | 61.36%     | 58.81%                |
|             |        | 5          | 68.41%     | 68.21%                |
|             |        | Average    | 66.01%     | 65.13%                |
|             | 300    | 1          | 62.87%     | 62.73%                |
|             |        | 2          | 64.74%     | 62.75%                |
|             |        | 3          | 68.03%     | 68.99%                |
|             |        | 4          | 67.29%     | 68.00%                |
|             |        | 5          | 70.54%     | 71.50%                |
|             |        | Average    | 66.69%     | 66.79%                |
|             | 500    | 1          | 65.17%     | 70.11%                |
|             |        | 2          | 62.11%     | 70.78%                |
|             |        | 3          | 65.23%     | 67.17%                |
|             |        | 4          | 66.75%     | 70.91%                |
|             |        | 5          | 70.57%     | 71.00%                |
|             |        | Average    | 65.97%     | 69.99%                |
| Fashion-MNIST | 100  | 1          | 77.14%     | 77.19%                |
|             |        | 2          | 78.71%     | 77.80%                |
|             |        | 3          | 74.22%     | 72.40%                |
|             |        | 4          | 77.11%     | 76.55%                |
|             |        | 5          | 77.39%     | 76.21%                |
|             |        | Average    | 76.91%     | 76.03%                |
|             | 300    | 1          | 75.01%     | 74.93%                |
|             |        | 2          | 77.02%     | 77.57%                |
|             |        | 3          | 74.26%     | 77.62%                |
|             |        | 4          | 74.88%     | 75.22%                |
|             |        | 5          | 78.14%     | 78.89%                |
|             |        | Average    | 75.43%     | 76.71%                |
|             | 500    | 1          | 76.72%     | 78.19%                |
|             |        | 2          | 75.57%     | 76.05%                |
|             |        | 3          | 74.82%     | 75.87%                |
|             |        | 4          | 76.60%     | 77.10%                |
|             |        | 5          | 76.10%     | 76.86%                |
|             |        | Average    | 75.96%     | 76.81%                |

Table 11: Results of MNIST, KMNIST and Fashion-MNIST with 200 seeds.
| Dataset   | Epochs | Experiment | Accuracy | Accuracy (Best Model) |
|-----------|--------|------------|----------|----------------------|
| MNIST     | 100    | 1          | 96.42%   | 96.71%               |
|           |        | 2          | 95.65%   | 95.63%               |
|           |        | 3          | 96.67%   | 96.45%               |
|           |        | 4          | 96.35%   | 96.67%               |
|           |        | 5          | 96.12%   | 95.95%               |
|           |        | Average    | 96.24%   | 96.28%               |
|           | 300    | 1          | 97.85%   | 97.82%               |
|           |        | 2          | 97.96%   | 98.04%               |
|           |        | 3          | 97.91%   | 97.91%               |
|           |        | 4          | 97.45%   | 97.42%               |
|           |        | 5          | 98.06%   | 97.92%               |
|           |        | Average    | 97.85%   | 97.82%               |
|           | 500    | 1          | 97.78%   | 97.85%               |
|           |        | 2          | 97.11%   | 97.70%               |
|           |        | 3          | 98.12%   | 98.12%               |
|           |        | 4          | 97.57%   | 97.65%               |
|           |        | 5          | 97.67%   | 97.25%               |
|           |        | Average    | 97.65%   | 97.71%               |
| KMNIST    | 100    | 1          | 73.91%   | 70.67%               |
|           |        | 2          | 68.88%   | 68.74%               |
|           |        | 3          | 69.00%   | 65.49%               |
|           |        | 4          | 67.33%   | 68.54%               |
|           |        | 5          | 71.60%   | 70.24%               |
|           |        | Average    | 70.18%   | 68.74%               |
|           | 300    | 1          | 70.34%   | 71.18%               |
|           |        | 2          | 65.85%   | 66.05%               |
|           |        | 3          | 70.64%   | 71.04%               |
|           |        | 4          | 71.83%   | 71.99%               |
|           |        | 5          | 70.74%   | 70.75%               |
|           |        | Average    | 69.88%   | 70.02%               |
|           | 500    | 1          | 68.80%   | 71.64%               |
|           |        | 2          | 67.87%   | 69.48%               |
|           |        | 3          | 63.09%   | 64.51%               |
|           |        | 4          | 69.39%   | 72.52%               |
|           |        | 5          | 70.17%   | 72.38%               |
|           |        | Average    | 67.86%   | 70.11%               |
| Fashion-MNIST | 100     | 1          | 78.36%   | 78.32%               |
|           |        | 2          | 78.76%   | 78.47%               |
|           |        | 3          | 79.43%   | 79.20%               |
|           |        | 4          | 77.05%   | 77.77%               |
|           |        | 5          | 78.44%   | 78.94%               |
|           |        | Average    | 78.41%   | 78.54%               |
|           | 300    | 1          | 77.29%   | 77.55%               |
|           |        | 2          | 77.89%   | 78.14%               |
|           |        | 3          | 78.23%   | 77.69%               |
|           |        | 4          | 80.26%   | 80.26%               |
|           |        | 5          | 79.37%   | 79.88%               |
|           |        | Average    | 78.61%   | 78.70%               |
|           | 500    | 1          | 79.92%   | 79.69%               |
|           |        | 2          | 78.99%   | 81.14%               |
|           |        | 3          | 78.62%   | 79.99%               |
|           |        | 4          | 80.32%   | 80.65%               |
|           |        | 5          | 79.61%   | 78.44%               |
|           |        | Average    | 79.49%   | 79.92%               |

Table 12: Results of MNIST, KMNIST and Fashion-MNIST with 300 seeds.
| Dataset     | Epochs | Experiment | Accuracy | Accuracy (Best Model) |
|-------------|--------|------------|----------|-----------------------|
| MNIST       | 100    | 1          | 97.18%   | 96.62%                |
|             |        | 2          | 95.30%   | 96.19%                |
|             |        | 3          | 97.02%   | 95.18%                |
|             |        | 4          | 96.29%   | 96.24%                |
|             |        | 5          | 96.03%   | 95.90%                |
|             |        | **Average**| **96.36%** | **96.03%**           |
|             | 300    | 1          | 97.23%   | 97.39%                |
|             |        | 2          | 97.22%   | 96.78%                |
|             |        | 3          | 97.45%   | 97.56%                |
|             |        | 4          | 97.77%   | 97.66%                |
|             |        | 5          | 97.46%   | 97.65%                |
|             |        | **Average**| **97.43%** | **97.41%**           |
|             | 500    | 1          | 97.95%   | 98.86%                |
|             |        | 2          | 97.46%   | 97.68%                |
|             |        | 3          | 97.82%   | 97.70%                |
|             |        | 4          | 97.41%   | 97.55%                |
|             |        | 5          | 98.05%   | 97.85%                |
|             |        | **Average**| **97.74%** | **97.73%**           |
| KMNIST      | 100    | 1          | 71.05%   | 71.82%                |
|             |        | 2          | 71.67%   | 71.67%                |
|             |        | 3          | 73.39%   | 73.11%                |
|             |        | 4          | 72.38%   | 71.15%                |
|             |        | 5          | 72.31%   | 72.00%                |
|             |        | **Average**| **72.16%** | **71.95%**           |
|             | 300    | 1          | 75.77%   | 73.93%                |
|             |        | 2          | 75.00%   | 71.47%                |
|             |        | 3          | 70.23%   | 69.73%                |
|             |        | 4          | 75.67%   | 75.95%                |
|             |        | 5          | 71.35%   | 70.12%                |
|             |        | **Average**| **73.60%** | **72.24%**           |
|             | 500    | 1          | 73.90%   | 74.24%                |
|             |        | 2          | 76.33%   | 76.49%                |
|             |        | 3          | 73.96%   | 73.48%                |
|             |        | 4          | 73.03%   | 74.30%                |
|             |        | 5          | 71.35%   | 72.86%                |
|             |        | **Average**| **73.71%** | **74.27%**           |
| Fashion-MNIST | 100    | 1          | 79.96%   | 79.96%                |
|             |        | 2          | 78.77%   | 77.63%                |
|             |        | 3          | 78.76%   | 78.82%                |
|             |        | 4          | 80.48%   | 79.60%                |
|             |        | 5          | 78.77%   | 79.92%                |
|             |        | **Average**| **79.35%** | **79.19%**           |
|             | 300    | 1          | 79.93%   | 80.07%                |
|             |        | 2          | 80.56%   | 80.92%                |
|             |        | 3          | 80.49%   | 81.10%                |
|             |        | 4          | 79.72%   | 80.69%                |
|             |        | 5          | 78.84%   | 79.24%                |
|             |        | **Average**| **79.75%** | **80.40%**           |
|             | 500    | 1          | 80.67%   | 81.64%                |
|             |        | 2          | 79.55%   | 80.06%                |
|             |        | 3          | 77.83%   | 81.73%                |
|             |        | 4          | 79.47%   | 79.13%                |
|             |        | 5          | 80.05%   | 80.85%                |
|             |        | **Average**| **79.51%** | **80.68%**           |

Table 13: Results of MNIST, KMNIST and Fashion-MNIST with 400 seeds.
| Dataset     | Epochs | Experiment | Accuracy  | Accuracy (Best Model) |
|-------------|--------|------------|-----------|-----------------------|
| MNIST       | 100    | 1          | 96.73%    | 96.33%                |
|             |        | 2          | 96.79%    | 96.60%                |
|             |        | 3          | 96.73%    | 96.10%                |
|             |        | 4          | 95.97%    | 96.09%                |
|             |        | 5          | 95.11%    | 95.40%                |
|             |        | **Average**| **96.27%**| **96.10%**            |
|             | 300    | 1          | 98.01%    | 97.83%                |
|             |        | 2          | 97.25%    | 97.34%                |
|             |        | 3          | 96.80%    | 96.99%                |
|             |        | 4          | 97.05%    | 97.33%                |
|             |        | 5          | 97.10%    | 97.35%                |
|             |        | **Average**| **97.24%**| **97.37%**            |
|             | 500    | 1          | 98.00%    | 98.03%                |
|             |        | 2          | 96.68%    | 96.66%                |
|             |        | 3          | 96.68%    | 96.66%                |
|             |        | 4          | 97.64%    | 97.64%                |
|             |        | 5          | 97.84%    | 97.74%                |
|             |        | **Average**| **97.31%**| **97.35%**            |
| KMNIST      | 100    | 1          | 73.53%    | 72.52%                |
|             |        | 2          | 72.78%    | 69.95%                |
|             |        | 3          | 75.81%    | 75.73%                |
|             |        | 4          | 70.35%    | 71.73%                |
|             |        | 5          | 75.72%    | 73.38%                |
|             |        | **Average**| **73.64%**| **72.66%**            |
|             | 300    | 1          | 73.52%    | 73.57%                |
|             |        | 2          | 73.71%    | 74.00%                |
|             |        | 3          | 75.72%    | 76.61%                |
|             |        | 4          | 74.02%    | 74.97%                |
|             |        | 5          | 77.03%    | 77.59%                |
|             |        | **Average**| **74.80%**| **75.35%**            |
|             | 500    | 1          | 74.88%    | 75.43%                |
|             |        | 2          | 75.27%    | 74.51%                |
|             |        | 3          | 73.70%    | 72.99%                |
|             |        | 4          | 76.76%    | 78.11%                |
|             |        | 5          | 75.25%    | 75.77%                |
|             |        | **Average**| **75.17%**| **75.36%**            |
| Fashion-MNIST| 100    | 1          | 80.55%    | 80.21%                |
|             |        | 2          | 81.58%    | 82.40%                |
|             |        | 3          | 78.93%    | 79.84%                |
|             |        | 4          | 82.08%    | 81.66%                |
|             |        | 5          | 80.53%    | 80.40%                |
|             |        | **Average**| **80.73%**| **80.90%**            |
|             | 300    | 1          | 79.87%    | 81.03%                |
|             |        | 2          | 80.09%    | 81.41%                |
|             |        | 3          | 81.81%    | 82.45%                |
|             |        | 4          | 81.75%    | 82.82%                |
|             |        | 5          | 80.19%    | 79.92%                |
|             |        | **Average**| **80.74%**| **81.53%**            |
|             | 500    | 1          | 80.69%    | 81.46%                |
|             |        | 2          | 79.65%    | 79.84%                |
|             |        | 3          | 80.20%    | 81.13%                |
|             |        | 4          | 80.27%    | 80.28%                |
|             |        | 5          | 79.79%    | 79.28%                |
|             |        | **Average**| **80.12%**| **80.40%**            |

Table 14: Results of MNIST, KMNIST and Fashion-MNIST with 500 seeds.
Figure 15: Seed \textit{seed.14} used in MNIST experiment as part of RQ3
Figure 16: Seed seed_102 used in MNIST experiment as part of RQ3