Progress Report

Daniel Biro

January 11, 2021

Contents

1 Experiments - workflow version: 1.1.0; experiments revision: YYY

1.1 Methods

The main script for running experiments (run_experiment.py) has been updated and simplified. It now also includes shuffling the data before each model training (the initial shuffling was removed from building_dataset.py), as well as model testing. The plotting and result aggregation functionality for showcasing means and errors were moved to two separate scripts, produce_plots.py in visualization and aggregate_result.py in utils.

Plotting now happens using Pandas dataframes and Seaborn lineplots for all plots except ROC curves which still use matplotlib. Now each trained model has an ROC curve on validation and test sets, while for the other metrics there is a single plot for their progression over time on the training and validation sets. AUC values for each class are now also recorded and a separate plot shows their progression over time on the validation set.

Model testing is called similarly to model training from the model_testing.py script, this creates the new metrics files which are then aggregated to get a single mean and error on each, as well as the ROC curves for each model.

New Snakemake rules and bash scripts were added for processing the genomic regions of the 20th chromosome as the test set, which this new set containing 5264 germline variants and 2269 somatic variants along with 7533 randomly sampled healthy positions (normal class). This test set has around 600 datapoints more than the training and validations sets combined.

All Sphinx documentation has been updated and the new code uploaded to GitHub.

1.2 Results

Runtimes have been measured for each model categories and they are longer than the estimate I have given earlier. Everything had to be re-run for shuffling, new plots, extra metrics and not all of it has finished by the time of writing this report. The previously most promising models have been prioritized.

The new approximate runtimes per model trained are as follows: basic model types: 2hrs, bidirectional models: 4hrs, Transformer network (2 encoder layers): 24.5hrs. Adding dropout did not significantly increase runtime.

Below the new validation and test set results of the previously and newly best peforming models are presented. There are two caveats with the Transformer model: the previous results were of a model with 1 encoder layer, this new model has 2 layers, with the new model performing worse. This is accidental as the scripts were usually started up for overnight training, but it was planned to compare how performance changes with an additional layer for later on. The 1 layer model type will be trained and evaluated as well. The new model's results are also only 4 runs due to the long training time. Observing some of the individual model results there is room for improvement with more models being trained (and for the 1 layer model type as well).

Model	Accuracy	Precision	Recall	F1 Score
GRU	94.51% ± 0.79	90.95% ± 3.79	89.14% ± 2.88	90.01% ± 3.35
Bi-GRU + 5% Dropout	95.62% ± 0.06	95.26% ± 0.07	92.56% ± 0.1	93.89% ± 0.07
Bi-GRU + 10% Dropout	95.46% ± 0.09	95.03% ± 0.14	92.32% ± 0.2	93.65% ± 0.16
Transformer + 10% Dropout	90.74% ± 0.07	89.41% ± 0.05	85.09% ± 0.14	87.2% ± 0.09
Perceptron	**96.24% ± 0.01**	**96.3% ± 0.01**	**93.72% ± 0.03**	**94.99% ± 0.02**

Table 1: Performance of best models on the validation set

MODEL	NORMAL	GERMLINE VARIANT	SOMATIC VARIANT
GRU	0.929% ± 0.02	0.9308% ± 0.01	0.904% ± 0.05
BI-GRU + 5% DROPOUT	0.945% ± 0.0	0.9331% ± 0.0	0.9591% ± 0.0
BI-GRU + 10% DROPOUT	0.9495% ± 0.0	0.9339% ± 0.0	0.96% ± 0.0
TRANSFORMER + 10% DROPOUT	0.8288% ± 0.0	0.8208% ± 0.01	0.8674% ± 0.0
PERCEPTRON	**0.9575**% ± **0.0**	**0.9706**% ± **0.0**	**0.9591**% ± **0.0**

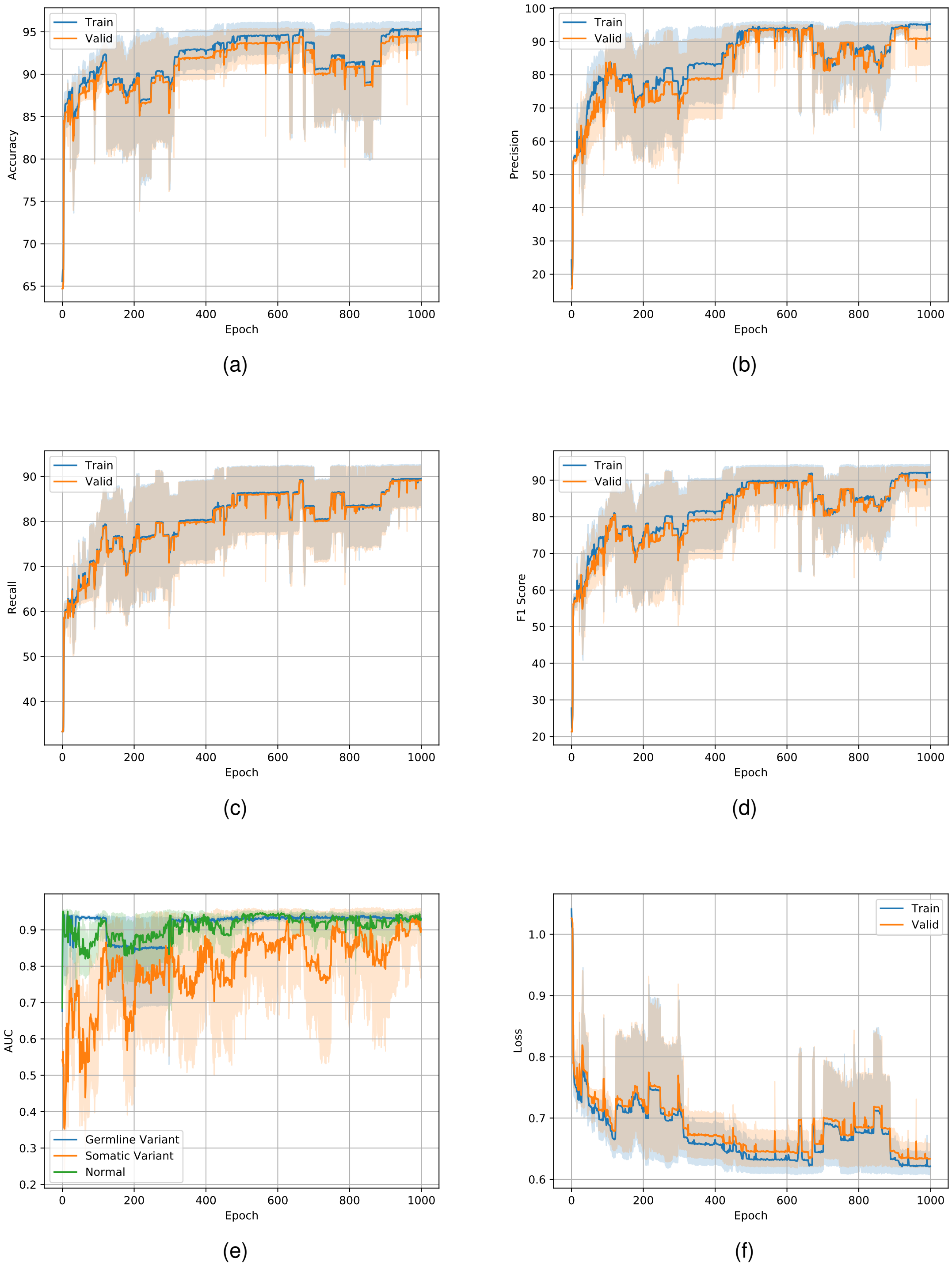
Table 2: Class AUCs of best models on the validation set

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
GRU	93.16% ± 0.78	89.88% ± 3.33	85.7% ± 2.61	87.73% ± 2.96
BI-GRU + 5% DROPOUT	**94.18**% ± **0.03**	**93.91**% ± **0.05**	**88.67**% ± **0.1**	**91.21**% ± **0.07**
BI-GRU + 10% DROPOUT	94.11% ± 0.04	93.67% ± 0.07	88.56% ± 0.12	91.04% ± 0.09
TRANSFORMER + 10% DROPOUT	79.93% ± 10.02	70.44% ± 12.51	67.33% ± 11.96	68.78% ± 12.22
PERCEPTRON	94.15% ± 0.01	94.15% ± 0.01	88.41% ± 0.01	91.19% ± 0.01

Table 3: Performance of best models on the test set

MODEL	NORMAL	GERMLINE VARIANT	SOMATIC VARIANT
GRU	0.9054% ± 0.02	0.9197% ± 0.0	0.8808% ± 0.04
BI-GRU + 5% DROPOUT	0.9172% ± 0.0	0.9235% ± 0.0	0.9266% ± 0.0
BI-GRU + 10% DROPOUT	0.9246% ± 0.0	0.9237% ± 0.0	0.9273% ± 0.0
TRANSFORMER + 10% DROPOUT	0.7371% ± 0.13	0.7176% ± 0.14	0.661% ± 0.16
PERCEPTRON	**0.928**% ± **0.0**	**0.9375**% ± **0.0**	**0.9224**% ± **0.0**

Table 4: Class AUCs of best models on the test set

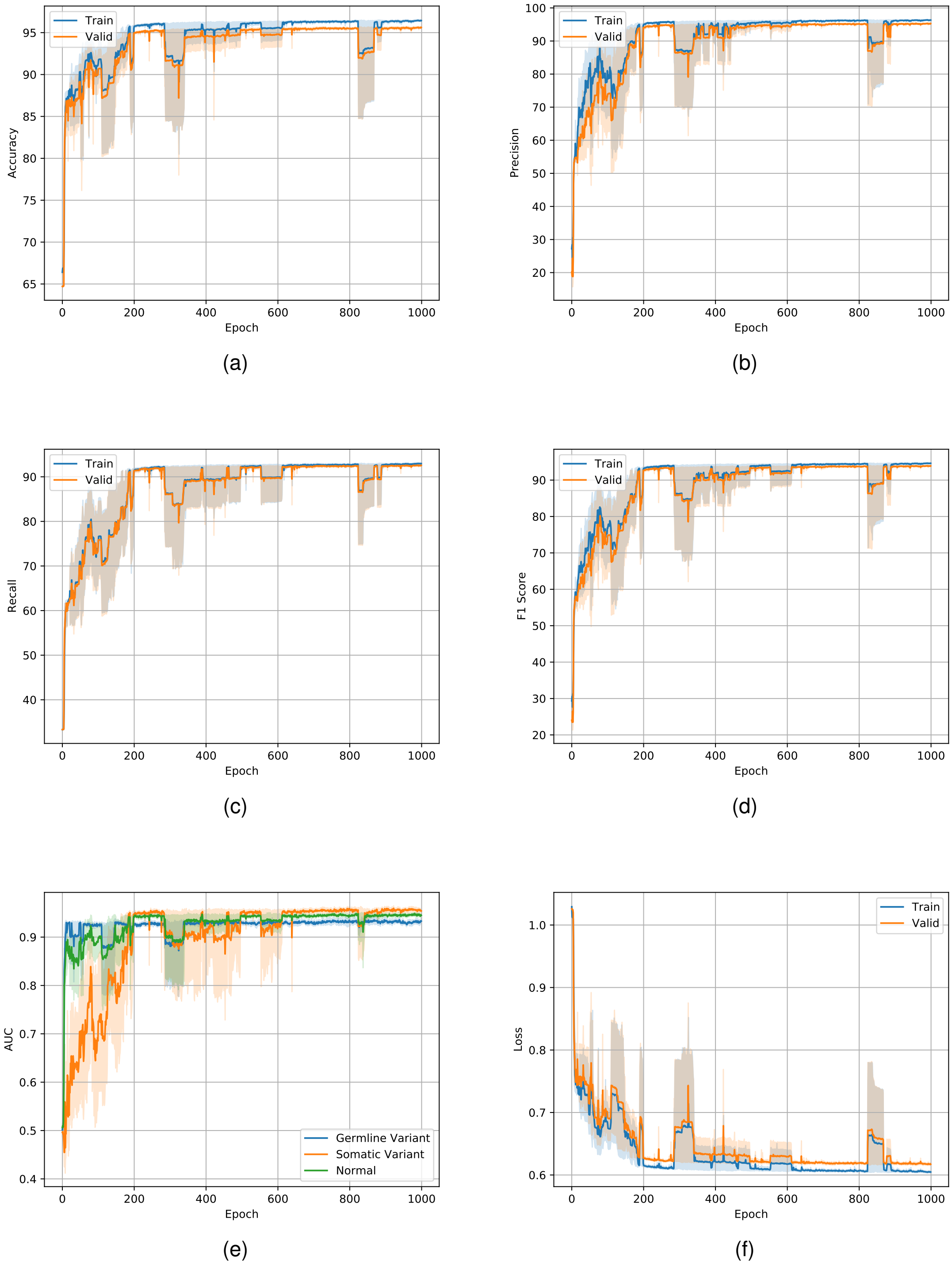


(a) (b)

(c) (d)

(e) (f)

Figure 1: Figures for GRU model



(a)

(b)

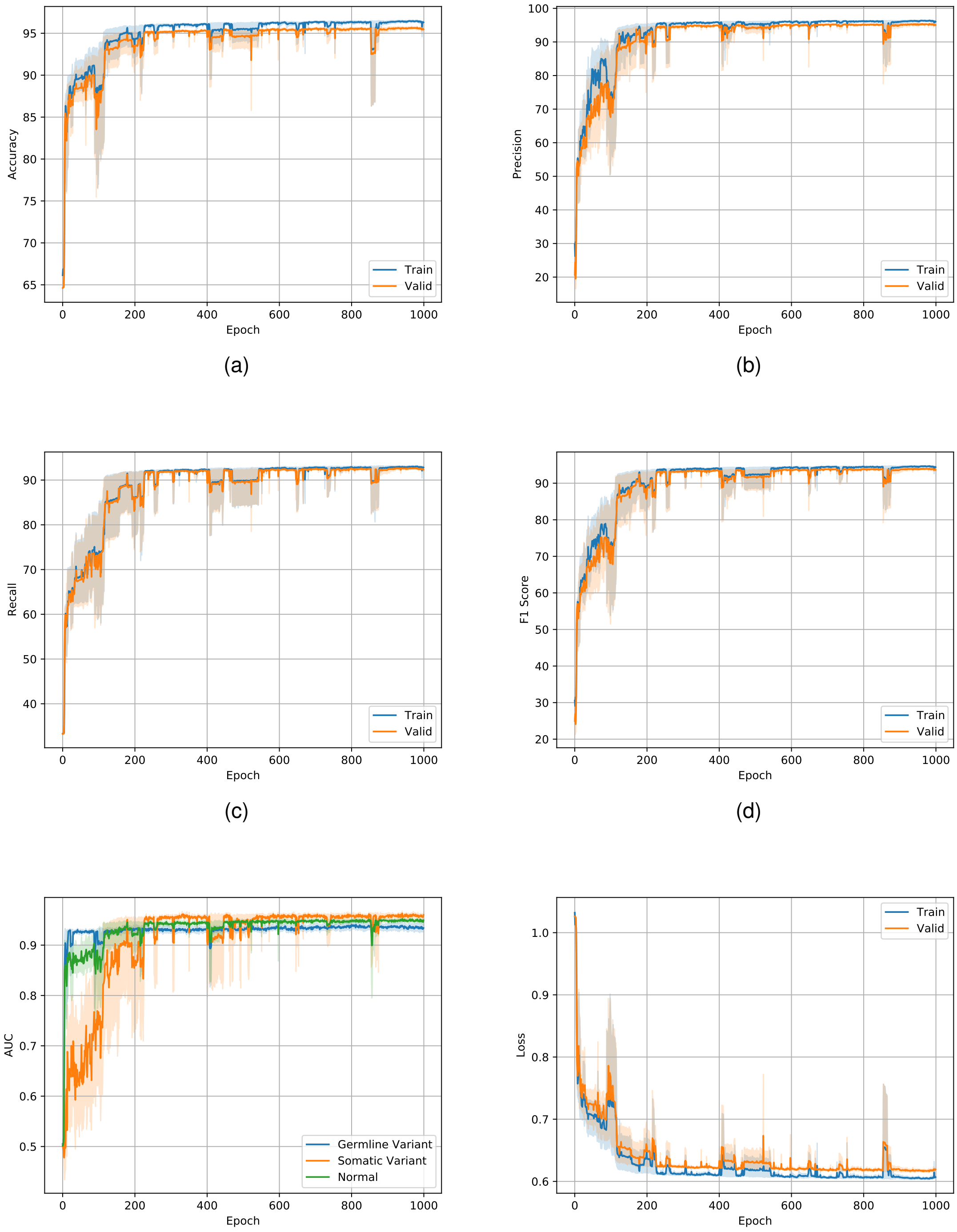
(c)

(d)

(e)

(f)

Figure 2: Figures for Bi-GRU + 5% Dropout model



(a)

(b)

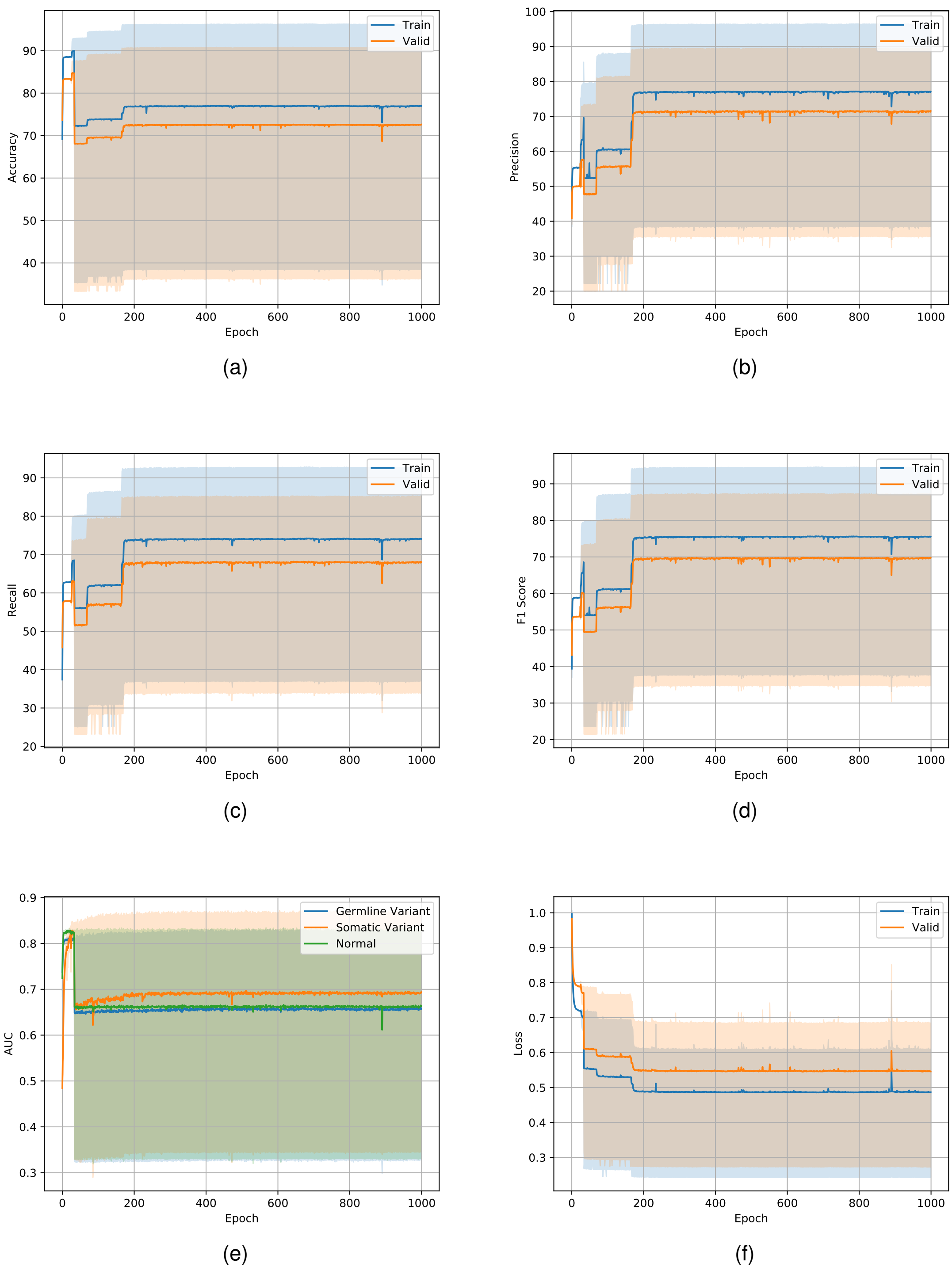
(c)

(d)

(e)

(f)

Figure 3: Figures for Bi-GRU + 10% Dropout model



(a)

(b)

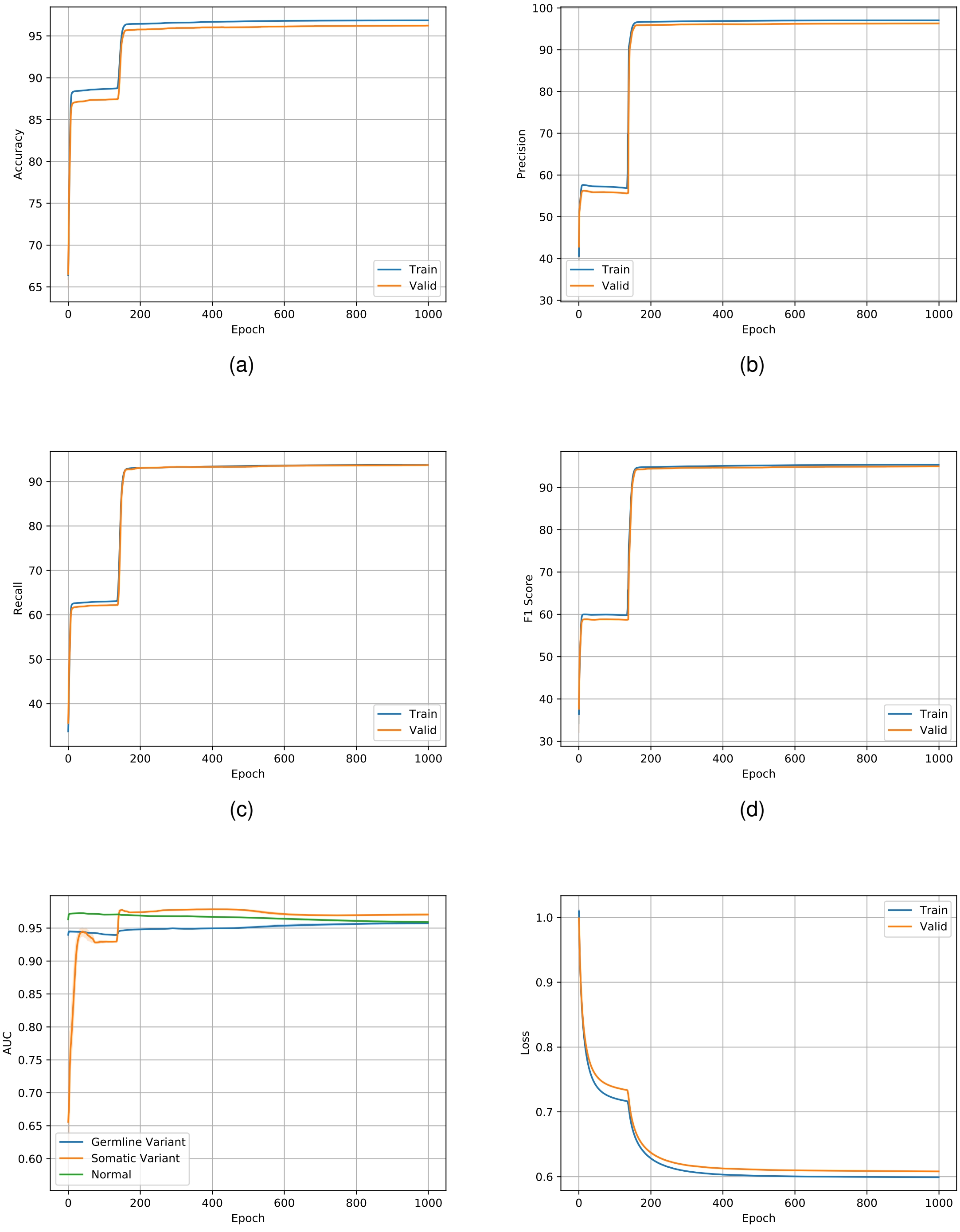
(c)

(d)

(e)

(f)

Figure 4: Figures for Transformer model



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: Figures for Perceptron Dropout model