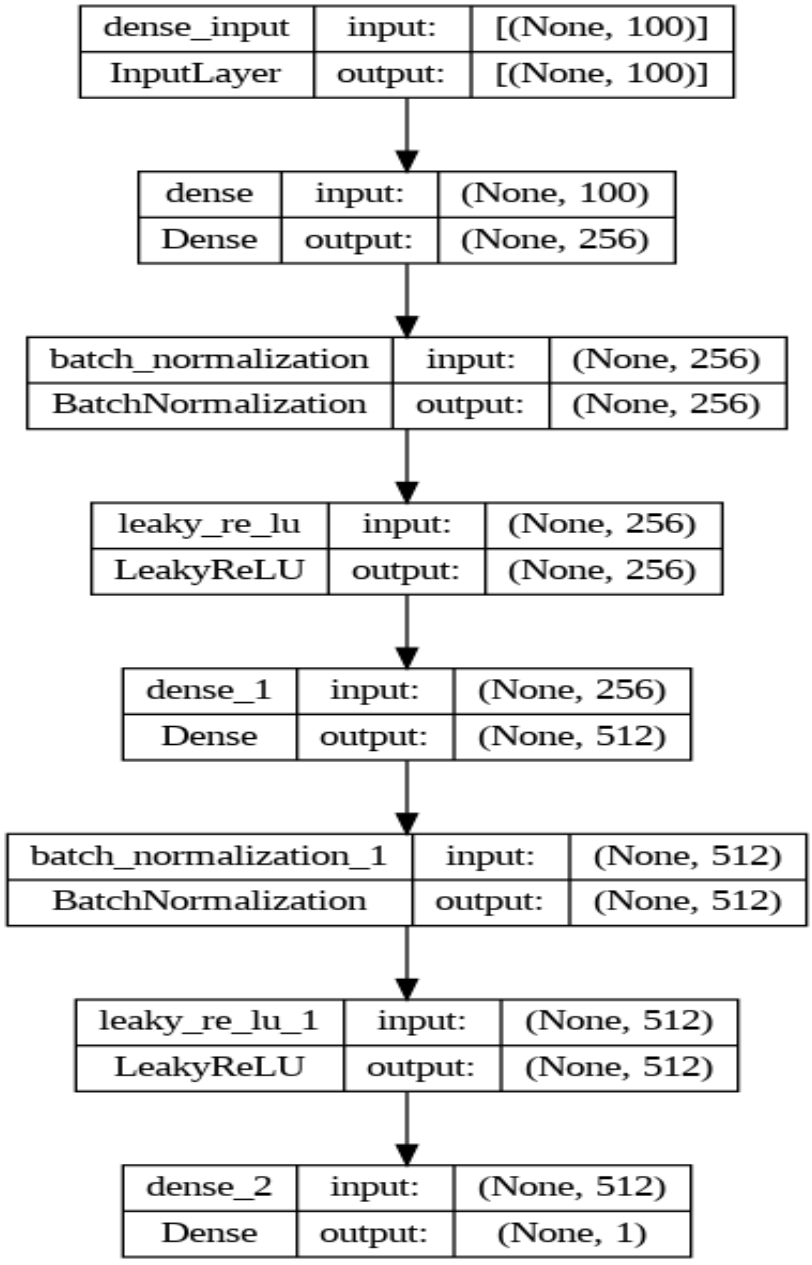
CONCEPT DRIFT DETECTION USING GAN

Dataset link:

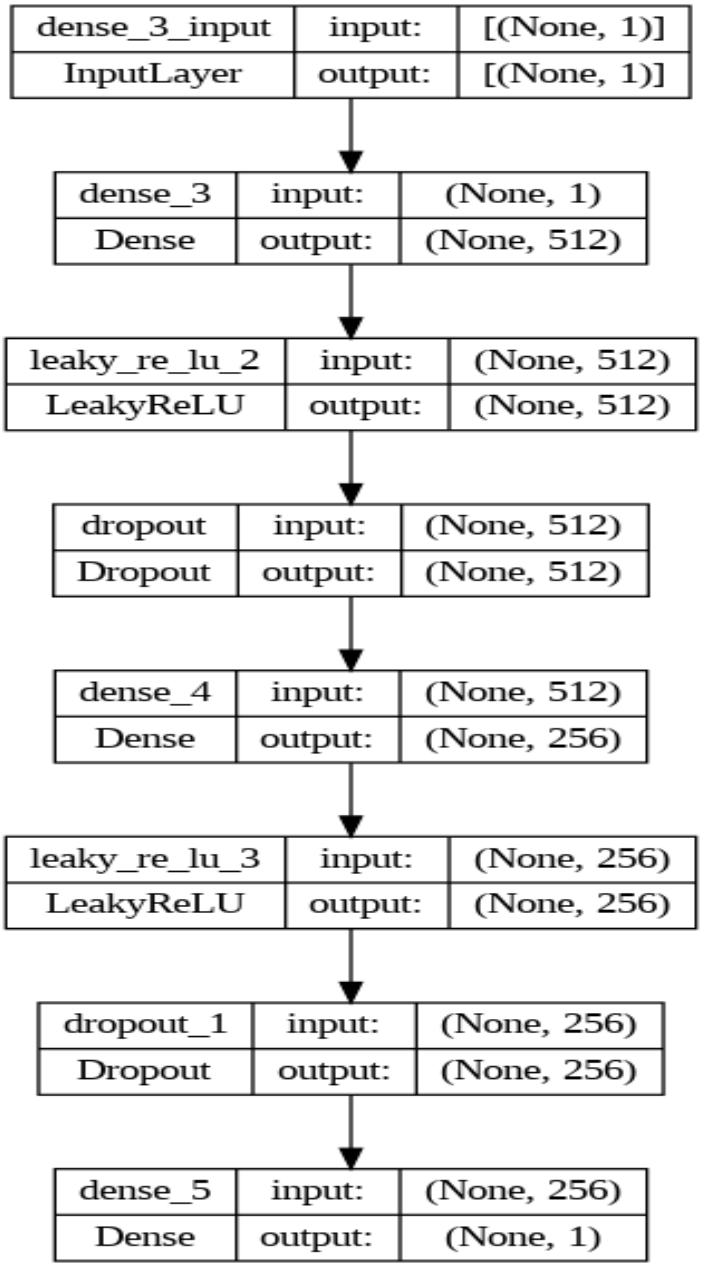
https://www.kaggle.com/datasets/victorcaelina/tuberculosis-symptoms

GAN Model Architecture:

Generator model:



Discriminator model:



Experiment details:

1. First I used the **Z-test on proportions** method to detect concept drift over the whole dataset.

2. Then I demonstrated the concept drift detection over each month of the dataset given using the same method.
3. I also used two additional methods for drift detection and analysis- the **Page-Hinkley Test** and the **KL-Divergence(Kullback-Lieber)** methods.
4. Real vs Virtual Concept drift distinction

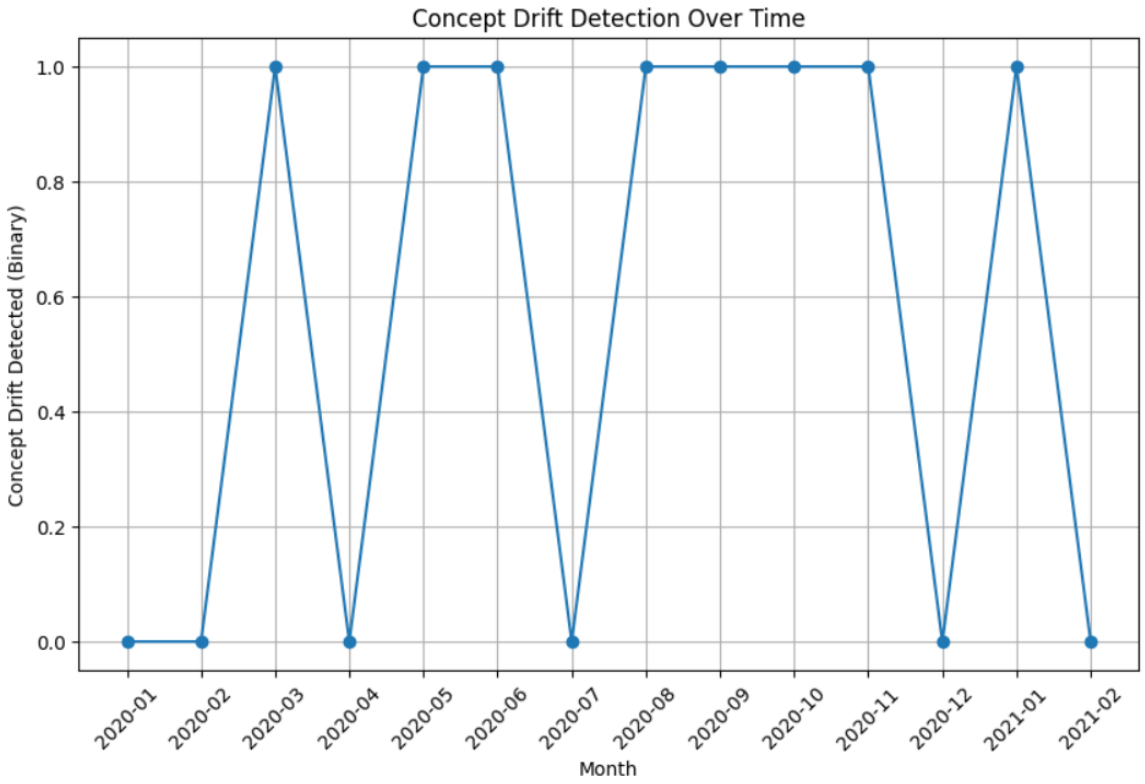
Results:

Exp-1.

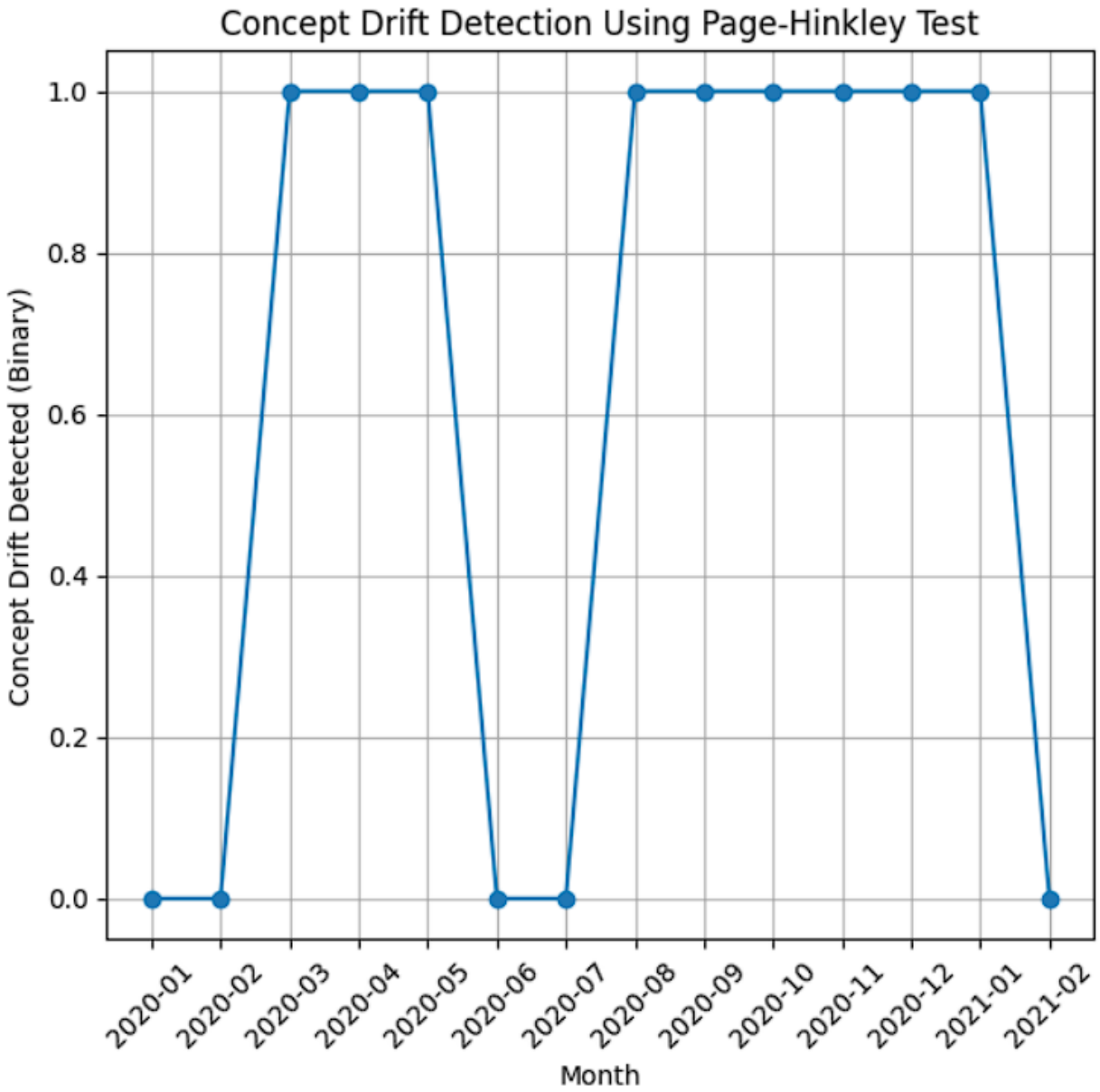
```python
from statsmodels.stats.proportion import proportions_ztest

def detect_concept_drift(real_data, synthetic_data, significance_threshold):
    # Perform a two-sample proportion Z-test
    z_stat, p_value = proportions_ztest([real_data.sum(), synthetic_data.sum()],
                                        [len(real_data), len(synthetic_data)])

    if p_value < significance_threshold:
        return True  # Concept drift detected
    return False


# Detecting concept drift using a two-sample proportion Z-test
significance_threshold = 0.4

real_binary_data = data['sputum mixed with blood'].values
synthetic_binary_data = (synthetic_data > 0.5).astype(int).flatten()

concept_drift_detected = detect_concept_drift(real_binary_data, synthetic_binary_data, significance_threshold)
print("Concept Drift Detected:", concept_drift_detected)
```

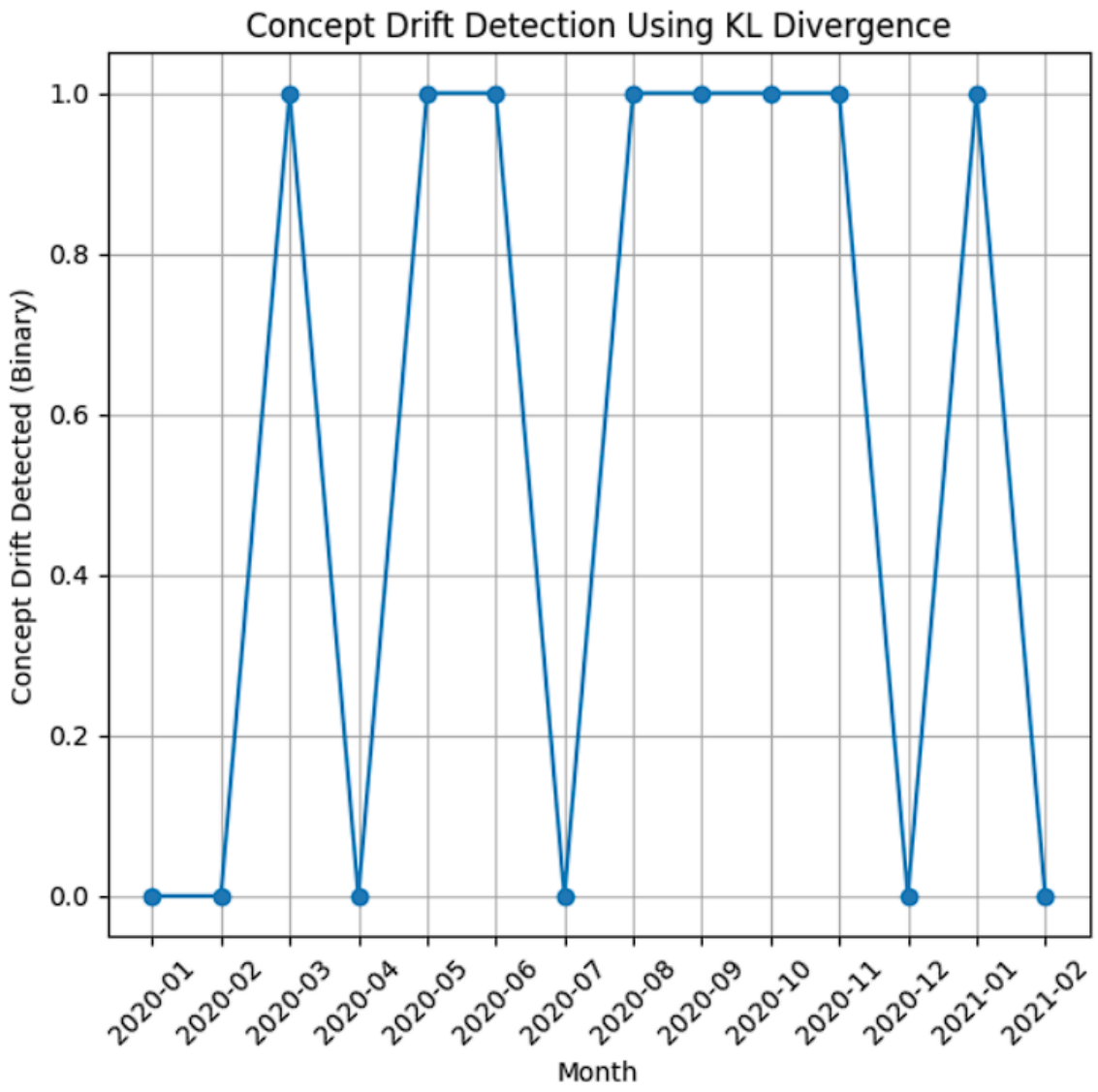
```
Concept Drift Detected: True
```

Exp-2. Same Threshold as above, i.e., 0.4

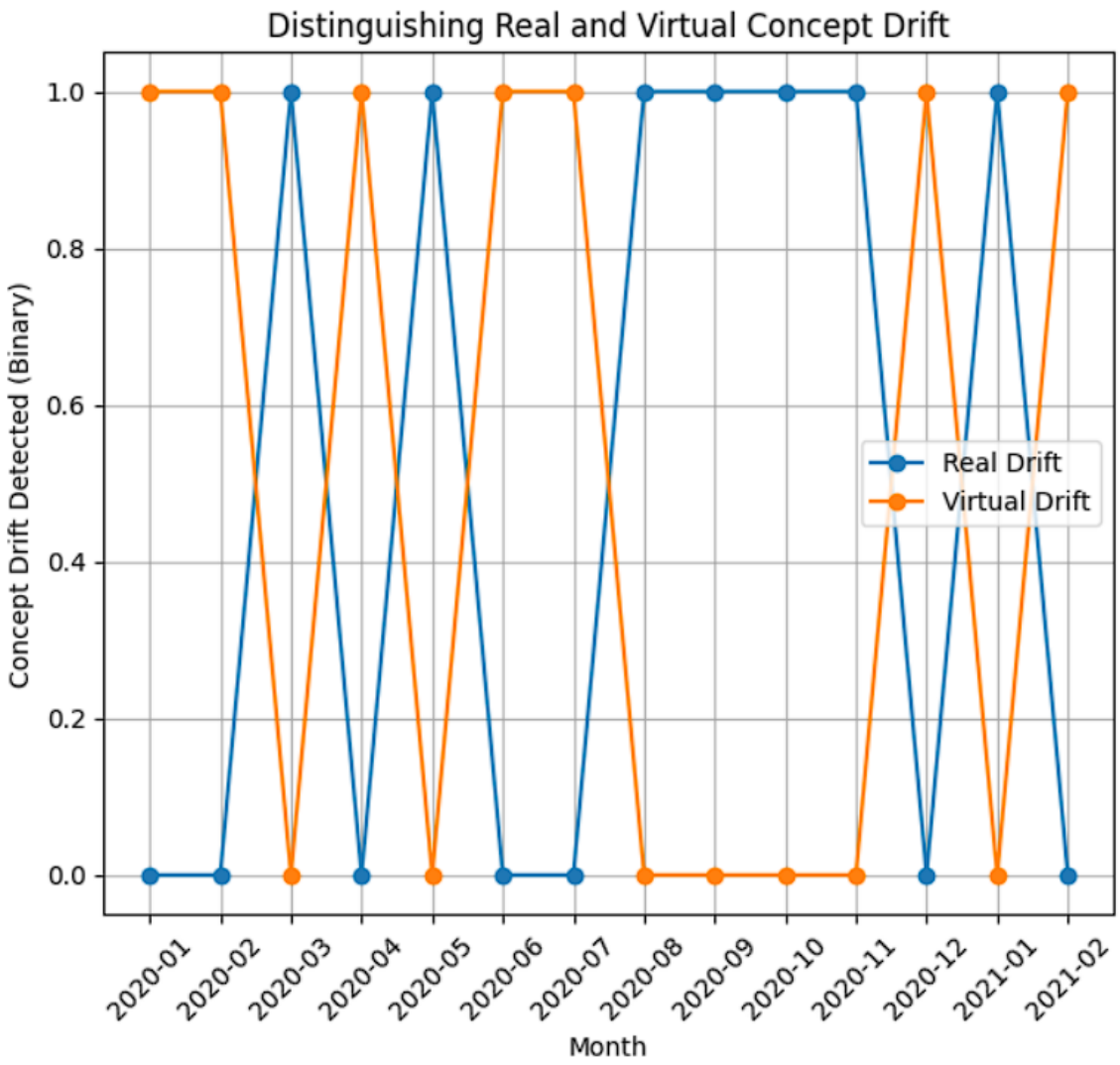


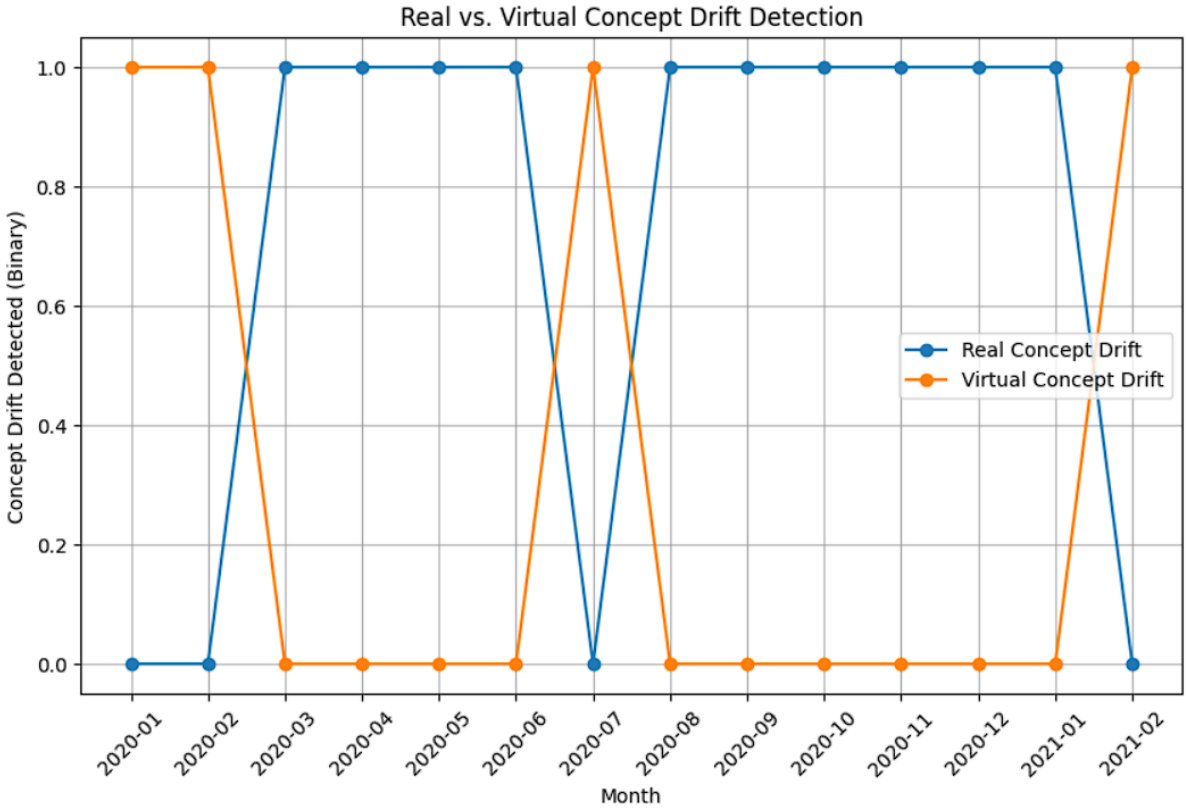
Exp-3. For PHT threshold is 30 and for KL-Divergence threshold is 0.005


Concept Drift Detection Using Page-Hinkley Test
Concept Drift Detected (Binary)
1.0
0.8
0.6
0.4
0.2
0.0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
Month



Exp-4. Same thresholds as above- In first img only PHT and KL-divergence is used; in second img all 3 methods are used


Distinguishing Real and Virtual Concept Drift
Concept Drift Detected (Binary)
1.0
0.8
0.6
0.4
0.2
0.0
Real Drift
Virtual Drift
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
Month


Real vs. Virtual Concept Drift Detection
Concept Drift Detected (Binary)
1.0
0.8
0.6
0.4
0.2
0.0
Real Concept Drift
Virtual Concept Drift
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
Month