Comparing Deep Learning and Statistical Detectors for TS InSAR Change-Point Detection: A Hybrid Pipeline Approach

Seyed Arya Fakhri[1], Mehran Satari[1*]

[1]Department of Geomatics Engineering, Faculty of Civil Engineering and Transportation, University of Isfahan, Isfahan, Iran- aryafakhri@trn.ui.ac.ir & sattari@eng.ui.ac.ir

Keywords: change point detection (CPD), InSAR, deep learning, time series, statistical detector, Sentinel 1.

Abstract
Accurate detection of change points (CPs) and turning points (TPs) in Interferometric Synthetic Aperture Radar (InSAR) time series (TS InSAR) is critical for reliable geohazard monitoring and early warning systems. We introduce a standardized comparison to rigorously evaluate seven CP/TP detectors, encompassing four novel deep learning architectures (MLP, MALkCNN, ATGLSTM, CNN-LSTM), two established deep learning baselines (LSTM+TGLSTM, BiLSTM+U-Net), and a fast statistical detector (STPD). Our comprehensive evaluation spans 100,000 synthetic time series and real Sentinel 1 data stacks from diverse European geohazard sites (Italy, Germany, Iceland), quantifying performance through accuracy metrics (precision, recall, F1-score), change magnitude error (RMSE, mm·yr^{-1}), agreement with co-located GNSS stations (r), and computational throughput. Among the tested models, ATGLSTM demonstrates superior detection accuracy (F1 up to 0.93 on synthetic and 0.80 on real data), exhibiting remarkable robustness to speckle noise and temporal gaps due to its attention and time gating mechanisms. While CNN-LSTM yields the strongest agreement with GNSS measurements (r≈0.88), the STPD method provides unparalleled efficiency (~0.2 s per 1,000 series), confirming a distinct accuracy–efficiency trade off. Based on these findings, we propose a hybrid operational pipeline that leverages STPD for rapid, large scale screening and ATGLSTM for high precision refinement of flagged candidates. This two stage approach preserves optimal detection accuracy while substantially reducing computational cost, offering a scalable solution for operational TS InSAR monitoring. Our comparison, datasets, and code are made publicly available to catalyze future research and the development of robust, operational pipelines and multi sensor fusion frameworks.

1. Introduction

Monitoring ground deformation is fundamental to geohazard management, infrastructure resilience, environmental sustainability, and urban planning (Tomás and Li, 2017). Processes such as subsidence, landslides, volcanic unrest, and aseismic creep can evolve subtly over months to years; timely detection of statistically significant deviations in displacement trajectories is therefore essential for risk mitigation and early warning (Khoshlahjeh Azar et al., 2021).

Interferometric Synthetic Aperture Radar (InSAR) enables wide area, day night, and all weather mapping of surface motion by exploiting phase differences between repeat SAR acquisitions (Bell et al., 2008). Time series InSAR (TS InSAR; also known as MTInSAR) techniques such as PSInSAR, SqueeSAR, and DePSI reconstruct point wise displacement histories from stacks of Single Look Complex (SLC) images (Ferretti et al., 2001). A crucial step in analyzing these time series for geohazard monitoring is the detection of significant shifts and trend changes, known as change points (CPs) and turning points (TPs) (Aminikhanghahi and Cook, 2017). The end-to-end workflow adopted in this study integrates this critical detection phase, beginning with SAR acquisition and TS InSAR inversion, moving to probabilistic CP/TP detection, and culminating in geohazard decision support. Monitoring ground deformation is fundamental to geohazard management, infrastructure resilience, environmental sustainability, and urban planning (Tomás and Li, 2017). Processes such as subsidence, landslides, volcanic unrest, and aseismic creep can evolve subtly over months to years; timely detection of statistically significant deviations in displacement trajectories is therefore essential for risk mitigation and early warning (Khoshlahjeh Azar et al., 2021).

This study compares seven methods for CP/TP detection, including four novel deep learning architectures (MLP (Fakhri and Satari, 2025a), MALkCNN (Fakhri and Satari, 2025b), ATGLSTM (Fakhri and Satari, 2026a), CNN-LSTM (Fakhri and Satari, 2026b)), two established DL baselines (LSTM+TGLSTM(Lattari et al., 2022), BiLSTM+U-Net (Shahryarinia et al., 2025)), and a fast statistical detector (STPD (Ghaderpour et al., 2024)). Despite their success, TS InSAR signals are noisy and irregular: atmospheric delays, orbital errors, temporal/geometry induced decorrelation, residual unwrapping errors, nonuniform sampling, seasonal components, and gaps lead to heteroscedastic and autocorrelated errors (Agram and Simons, 2015). These properties hinder robust detection of structural changes in displacement time series (Khoshlahjeh Azar et al., 2022). We define a change point (CP) as an abrupt, discrete shift in the displacement time series (e.g., a step change), whereas a turning point (TP) denotes a gradual change in trend (e.g., a velocity change). Accurate detection of both is critical for interpreting geophysical processes (Aminikhanghahi and Cook, 2017).

Classical statistical detectors spanning step and breakpoint models with sequential tests, robust trend break estimation, and state space formulations have been widely applied to InSAR change analysis (Hussain et al., 2021, Cigna et al., 2011). However, performance can degrade under strong speckle related variability, missing epochs, and nonstationary dynamics, motivating adaptive, data driven approaches.

Deep learning (DL) methods advance time series analysis by learning complex temporal patterns directly from data (Han et al., 2019). Architectures based on recurrent networks (LSTM/BiLSTM and time gated variants, TGLSTM), temporal convolutional models (CNN-LSTM, MALkCNN),

* Corresponding author

and attention augmented designs (ATGLSTM) are promising for CP/TP detection because they integrate long temporal context, handle irregular sampling, and produce probabilistic outputs suitable for decision thresholding.

A unified, domain realistic comparison comparing statistical and modern DL approaches for CP/TP detection in TS InSAR has been lacking. Furthermore, no validated operational framework has combined statistical and DL methods to resolve the accuracy–efficiency trade-off for large-scale TS InSAR. This work fills these gaps by: (i) standardizing seven representative methods four novel DL architectures (MLP, MALkCNN, ATGLSTM, CNN-LSTM), two established DL baselines (LSTM+TGLSTM, BiLSTM+U-Net), and a fast statistical detector (STPD); (ii) designing evaluation protocols on large-scale synthetic datasets and real Sentinel-1 stacks from diverse European sites; (iii) quantifying accuracy (precision, recall, F1), change magnitude error (RMSE, mm·yr^{-1}), agreement with co-located GNSS (r), and throughput to expose accuracy–efficiency trade-offs relevant for operations; and (iv) proposing a novel two-stage hybrid pipeline (STPD+ATGLSTM) that leverages statistical speed for rapid screening and DL precision for refinement, enabling scalable operational monitoring.

2. Comparative Methodology

This section details the datasets, the synthetic time series generator, real Sentinel 1 stacks, the seven compared methods, training and calibration protocols, evaluation metrics, statistical testing, and efficiency measurement tailored to TS InSAR constraints (heteroscedastic and temporally correlated errors, speckle, nonuniform sampling, missing epochs, and occasional unwrapping artefacts).

2.1 Datasets

2.1.1 Synthetic time series generator: A total of 100,000 displacement time series representative of TS InSAR measurement points (MPs) are simulated with the following components (Fakhri and Satari, 2025a, Lattari et al., 2022, Wassie and Milillo, 2025):

- **Length & cadence:** series length $T\sim[60,240]$ epochs; inter epoch intervals Δt drawn from a bimodal 6/12 day distribution (Sentinel 1 revisits) with random omissions to emulate acquisition gaps.
- **Events:** number of events $Ne\in(0,\ldots,5)$ with minimum separation $\Delta t_{\min}$. Two event types are instantiated: CP (step/Heaviside) with amplitude Δd (mm) and TP (velocity break) with slope change Δv (mm·yr^{-1}); event times are sampled uniformly subject to $\Delta t_{\min}$.
- **Seasonality & drift:** optional annual harmonic and low order polynomial drift.
- **Noise:** additive Gaussian noise with AR(1) temporal correlation (Chang and Hanssen, 2015),

$$\varepsilon_t = \phi\varepsilon_{t-1} + \eta_t, \qquad \eta_t \sim \mathrm{N}(0,\sigma^2) \tag{1}$$

where σ is scaled to realize low/medium/high SNR regimes.

- **Missing epochs:** random dropouts with rates $g\in(0,0.1,0.2)$; series are left as is (no forward fill).
- **Unwrapping artefacts:** rare $\pm 2\pi$ phase jumps converted to LOS displacement and injected at rate $pu\leq 1\%$.
- **Ground truth & matching:** each event stores (t,type,magnitude). A detection is counted correct if matched within a tolerance window $\pm K$ epochs, where K scales with the median Δt and generator uncertainty.

3.1.2 Real Sentinel 1 stacks: Multi year Sentinel 1 IW stacks (ascending/descending) from European sites (Italy: Frosinone, Tuscany, Campi Flegrei; Germany: Ruhr/Rhineland; Iceland: Reykjanes) cover subsidence, landslide, and volcanic settings (Torres et al., 2012). The processing chain begins with coregistration, followed by interferogram formation, multi temporal speckle filtering, and phase unwrapping. Subsequently, orbital and atmospheric corrections are applied before performing the TS InSAR inversion (e.g., PSInSAR/SqueeSAR/DePSI) to obtain LOS displacement series for dense MPs (Ferretti et al., 2001, Ferretti et al., 2011). Quality control retains MPs with temporal coherence (γ) ≥ 0.70 and amplitude dispersion (AD) ≤ 0.25; outliers are removed using robust statistics (Crosetto et al., 2016). For external validation, TS InSAR MPs are co-located with GNSS stations within r ≤ 1–2 km; conversion between LOS and ENU coordinates uses local incidence/heading angles, and time alignment uses bilinear interpolation to common epochs (Samsonov and d'Oreye, 2017). Study areas and representative examples are summarized in Table 1 and Figure 1.

AOI (Site)	Country / Approx. AOI Size (km²)	Time span	Acquisitions (Asc/Desc)	Incidence / Heading (°)	MP density (pts·km^{-2})	Quality threshold (used here)	GNSS co-location radius	LOS-ENU Conversion Details
Frosinone Province	Italy / ~3,247	2019–2023	≈240 / ≈240	29–46 / ~347 (Asc), ~167 (Desc)	300–900	temporal_coh ≥ 0.70; AD ≤ 0.25	≤2 km (EUREF/EPN)	Local incidence/heading; bilinear resampling
Elba Island (Tuscany)	Italy / 224	2019–2023	≈235 / ≈235	29–46 / ~347 (Asc), ~167 (Desc)	250–700	temporal_coh ≥ 0.70; AD ≤ 0.25	≤2 km	As above
Ruhr / Rhineland conurbation	Germany / ~4,435	2019–2023	≈260 / ≈260	29–46 / ~347 (Asc), ~167 (Desc)	1,000–3,000	temporal_coh ≥ 0.70; AD ≤ 0.25	≤2 km (EUREF/EPN)	As above
Campi Flegrei Caldera	Italy / ~130	2019–2023	≈230 / ≈230	29–46 / ~347 (Asc), ~167 (Desc)	1,500–4,000	temporal_coh ≥ 0.70; AD ≤ 0.25	≤2 km (EUREF/INGV)	As above

Reykjanes Peninsula	Iceland / ~829	2019–2023	≈220 / ≈220	29–46 / ~347 (Asc), ~167 (Desc)	100–400*	temporal_coh ≥ 0.70; AD ≤ 0.25	≤5 km (NSII/IMO/UI)	As above

Table 1. Real Sentinel-1 stack descriptors. For each AOI: time span, number of acquisitions (ascending/descending), incidence/heading angles, TS InSAR MP Density (pts·km^{-2}), ensemble-coherence threshold (γ), GNSS co-location radius, and LOS-ENU Conversion Notes.

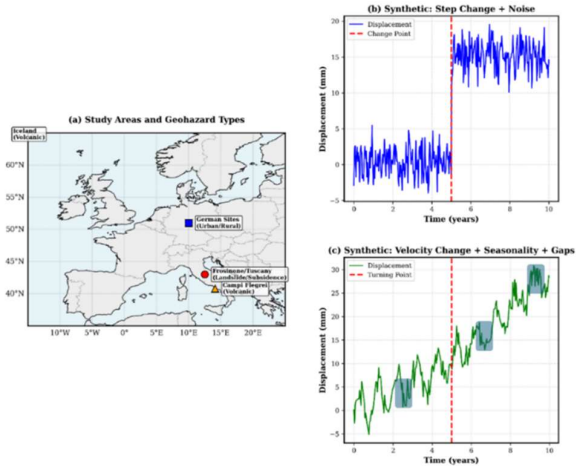

Figure 1. Study areas and synthetic TS InSAR examples. (a) Map of European geohazard sites (details in Table 1). (b) Synthetic time series showing a step change (CP) with noise. (c) Synthetic time series showing a velocity change (TP) with seasonality and gaps. (Units: mm).

2.2 Compared methods

Seven detectors spanning statistical and DL families are compared; all produce per epoch probabilities for CP and/or TP unless noted.

1. STPD: lightweight sequential tests for steps and slope breaks with mild regularization; outputs event times and confidence scores (Ghaderpour et al., 2024).
2. MLP: feed forward network on engineered features (finite differences, local regressions, seasonality indices) (Fakhri and Satari, 2025a).
3. MALkCNN: multi scale temporal CNN with dilated kernels capturing multi resolution motifs (Fakhri and Satari, 2025b).
4. CNN-LSTM: convolutional front end + LSTM back end for local to global temporal modeling (Fakhri and Satari, 2026b).
5. ATGLSTM: LSTM augmented with attention and time gating to handle irregular sampling and focus on informative windows (Fakhri and Satari, 2026a).
6. LSTM+TGLSTM (baseline): standard LSTM ensembled with a time gated variant (Lattari et al., 2022).
7. BiLSTM+U-Net (baseline): bidirectional RNN features re-cast as a 1D map and segmented with a U-Net head (Shahryarinia et al., 2025).

Seven detectors spanning statistical and DL families are compared; all produce per-epoch probabilities for CP and/or TP unless noted. Architectural hyperparameters, optimizers, and loss choices are summarized in Table 2.

Method	Key architectural components	Tuned hyperparameters (examples)	Optimizer / Loss (classification; aux.)
STPD	Sequential step / slope break tests; lightweight CPU ops	Window length; penalty / smoothing	N/A (analytical); no classifier loss
MLP	Feed forward on engineered features	3–5 layers; 128–512 neurons; dropout 0.2	Adam / Binary Cross Entropy (class weighted)
MALkCNN	Multi scale temporal CNN; dilated kernels	4–6 conv.; kernel 7–15; 32–128 filters	Adam / Binary Cross Entropy (aux. Δv: WMSE, optional)
ATGLSTM	LSTM with attention + time gating	2–3 (Bi)LSTM; 64–128 units; 4 heads	Adam / Binary Cross Entropy (aux. Δv: WMSE, optional)
CNN-LSTM	Convolutional front end + LSTM back end	3 conv.; 2 LSTM; 32–128 filters/units	Adam / Binary Cross Entropy (aux. Δv: WMSE, optional)
LSTM+TGLSTM	Standard LSTM ensembled with time gated LSTM	128 units; batch 32	Adam / Binary Cross Entropy
BiLSTM+U-Net	BiLSTM features re-cast as a 1-D map and segmented with a U-Net head (GASF optional)	2 BiLSTM; batch 32	Adam / Binary Cross Entropy

Table 2. Model configurations, optimizers, and losses. Architectural components, example hyperparameters, and loss choices for each detector. Classification uses binary cross entropy with class weighting; an auxiliary Δv head (when used) is trained with weighted MSE.

2.3 Training, calibration, and regularization

- **Splits:** synthetic data use train/val/test = 70/10/20 with five random seeds; real data use cross site evaluation (train on A+B, test on C) and within site splits for sensitivity checks (Ghaderpour et al., 2021).
- **Loss:** binary/multi label cross entropy with class imbalance weighting; an auxiliary weighted MSE

head regresses Δv when applicable (Goodfellow et al., 2016).

- **Optimization:** Adam (initial LR 1e−3), cosine decay, and early stopping on validation F1 (Kingma and Ba, 2014).
- **Data augmentation:** random temporal shifts, stochastic epoch dropouts, jitter on Δd and Δv, and additive colored noise consistent with Eq. (1) (Shorten and Khoshgoftaar, 2019, Fakhri et al., 2023).
- **Calibration & thresholds:** probabilities are isotonic calibrated; decision thresholds are selected per dataset by maximizing validation F1 (Zadrozny and Elkan, 2002).

These choices yield calibrated probabilities suitable for thresholding in operational TS InSAR (Osmanoğlu et al., 2016).

2.4 Evaluation protocol and metrics

Both event level detection and magnitude/localization accuracy are assessed.

- **Event detection (CP/TP):** Precision, Recall, and F1 computed with one to one matching within $\pm K$ epochs; PR-AUC complements threshold specific F1. Detection latency is measured for sequential/online modes (Adams and MacKay, 2007).
- **Magnitude & localization:** for TPs (slope change) and, where defined, steps, the change magnitude RMSE is Eq. (2) (Goodfellow et al., 2016).

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\Delta\hat{v}_i - \Delta v_i\right)^2} \quad (\mathrm{mm\cdot yr^{-1}}) \tag{2}$$

and temporal localization error (|Δepoch|) is summarized by mean±SD and median [IQR].

- **External agreement:** Pearson's r between TS InSAR derived velocities and co located GNSS after LOS-ENU mapping and temporal alignment (Samsonov and d'Oreye, 2017).
- **Efficiency:** throughput (seconds per 1,000 series) and memory footprint at inference (Rouet-Leduc et al., 2021).

2.5 Statistical testing

Per-series paired tests are used with Holm–Bonferroni correction for multiple comparisons (Burnham and Anderson, 2002). Effect sizes (Cohen's d) and 95% confidence intervals are obtained via nonparametric bootstrap (10,000 resamples). Robustness analyses are stratified by SNR, gap rate g, AR(1) coefficient ϕ, and presence/absence of unwrapping artefacts; effect sizes are reported alongside p-values to quantify the magnitude of pairwise differences (Gundogdu et al., 2025).

2.6 Implementation and efficiency measurement

All DL models run with mixed precision inference (Rakka et al., 2022). Hardware (GPU model/memory, CPU, RAM) is reported alongside throughput. Batch sizes are tuned to saturate GPU memory without OOM; warm up runs are discarded, and timings are averaged over at least five runs. For STPD, vectorized CPU code is used; GPU fallback and CPU baselines for DL are reported where relevant (Ghaderpour et al., 2024).

2.7 Reproducibility

The release includes the synthetic generator (parameter ranges, distributions, seeds), configuration files for each model (hyperparameters, thresholds, calibration fits), and scripts for computing all metrics and statistical tests. Where licensing permits, preprocessed Sentinel 1 site descriptors (metadata, MP indices/masks) and GNSS pairing lists are provided to enable end to end reproduction (Osmanoğlu et al., 2016).

3. Results

3.1 Quantitative results on synthetic data

Across 100k simulated series, ATGLSTM attains the highest F1 score (up to 0.93), followed closely by CNN-LSTM; both maintain strong performance across SNR regimes and gap rates. The statistical detector STPD yields competitive precision but lower recall at low SNR and under clustered events, consistent with its simpler decision rule. For real Sentinel 1 stacks (averaged across sites), F1 scores are lower due to domain shifts, but trends remain consistent. Performance metrics (F1 score, RMSE for temporal localization, computational time) are visualized in Figure 2, with full summaries for synthetic data in Table 3 and real data in Table 4. Error bars indicate 95% CIs over five random seeds; matching tolerance is $\pm K$ epochs.

Method	Precision	Recall	F1 Score	Temporal localization RMSE (days)	Change magnitude RMSE (mm·yr⁻¹)	Time (s/1,000 series)
ATGLSTM	0.94	0.92	0.93	2.8	1.2	5.2
CNN-LSTM	0.92	0.90	0.91	3.1	1.4	6.0
LSTM+TGLSTM	0.90	0.89	0.90	3.5	1.6	6.8
MALkCNN	0.88	0.86	0.87	4.0	1.9	3.8
BiLSTM+U-Net	0.86	0.85	0.86	4.3	2.1	7.5
STPD	0.83	0.82	0.83	5.1	2.5	0.2
MLP	0.80	0.79	0.80	5.8	2.8	0.9

Table 3. Synthetic comparison results. Mean ± SD (and 95% CI) for Precision/Recall/F1 (CP, TP), change magnitude RMSE (mm·yr⁻¹) for TP, and temporal localization error (|Δepoch|). Symbols † denote significant differences after Holm–Bonferroni correction. n=100,000 series; 5 random seeds. Effect sizes (Cohen's d) are reported for all pairwise contrasts.

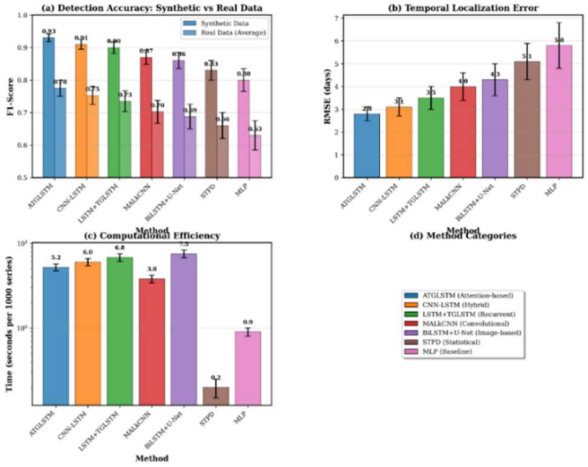

Figure 2. Quantitative comparison of CP/TP detectors. (a) F1 score (synthetic vs. real data), (b) Temporal localization RMSE (days), and (c) Computational efficiency (s/1,000 series, log scale). ATGLSTM achieves the highest F1 (up to 0.93), while STPD offers the best efficiency (~0.2 s/1,000 series), highlighting the accuracy–efficiency trade-off.

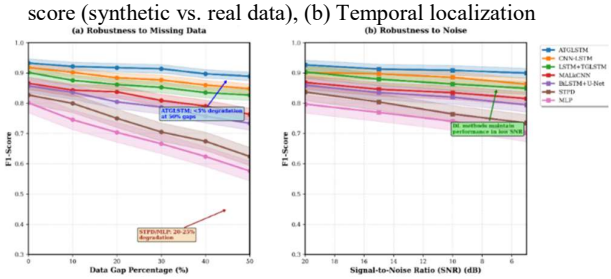

Figure 3. Robustness analysis of F1 score sensitivity to: (a) Data gap percentage (0–50%) and (b) Signal-to-Noise Ratio (SNR, 5–20 dB). Attention-based models (ATGLSTM, CNN-LSTM) degrade most slowly, while STPD and MLP show larger drops. Error bars represent 95% CIs.

3.2 Robustness to domain stressors

Sensitivity analyses (Figure 3) show the degradation of F1 scores with increasing data gap percentage (0–50%) and decreasing signal to noise ratio (SNR, 5–20 dB). Attention/time gating models (ATGLSTM, CNN-LSTM) exhibit the slowest degradation, with F1 scores dropping less than 5% at 50% gaps and maintaining strong performance at low SNR, reflecting resilience to missing data and noise. Statistical methods (STPD, MLP) show larger drops (20–25% at 50% gaps). Results are based on synthetic series with five random seeds; error bars indicate 95% confidence intervals.

3.3 Real Sentinel 1 stacks: detection and external agreement

Across subsidence, landslide, and volcanic sites, ATGLSTM achieves the highest F1 (0.75–0.80), whereas CNN-LSTM shows the strongest agreement with co-located GNSS (r≈0.88). This indicates that one model may excel in event detection (ATGLSTM) while another performs better in change magnitude estimation (CNN-LSTM). Site wise metrics precision/recall/F1, change magnitude RMSE (mm·yr^{-1}), and Pearson's correlation with GNSS (r) are summarized in Table 4. Spatial patterns for the Frosinone landslide site (Figure 4) show coherent clusters of true positives for ATGLSTM along known deformation corridors, while scattered false alarms from MLP correlate with sparse sampling.

Method	F1 Score (Frosinone)	F1 Score (Tuscany)	F1 Score (Germany)	F1 Score (Campi Flegrei)	F1 Score (Iceland)	GNSS Corr (r)
ATGLSTM	0.75	0.78	0.76	0.80	0.71	0.88
CNN-LSTM	0.73	0.76	0.74	0.78	0.69	0.85
LSTM+TGLSTM	0.71	0.75	0.72	0.76	0.68	0.83
MALkCNN	0.68	0.71	0.69	0.73	0.62	0.80
BiLSTM+U-Net	0.66	0.70	0.68	0.71	0.60	0.79
STPD	0.64	0.67	0.65	0.68	0.57	0.77
MLP	0.61	0.64	0.62	0.65	0.56	0.75

Table 4. Real stacks: detection and external agreement. Site wise F1 scores (CP, TP), and Pearson's correlation with co-located GNSS velocities (r). Footnotes list site specific K (matching tolerance, epochs) and GNSS co-location radius.

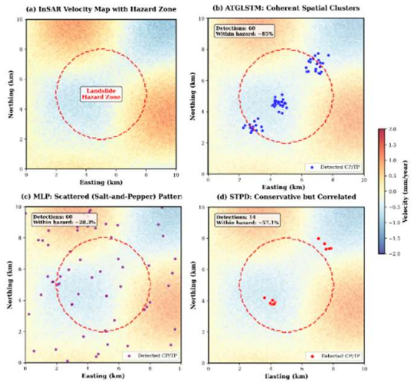

Figure 4. Detection patterns in Frosinone landslide area (Sentinel 1). (a) InSAR velocity map (mm·yr^{-1}) with hazard zone. (b) ATGLSTM: coherent detection clusters. (c) MLP: scattered "salt-and-pepper" pattern. (d) STPD: conservative but correlated detections. ATGLSTM hotspots align best with deformation features.

3.4 Efficiency and memory footprint

STPD delivers unmatched throughput (≈ 0.2 s per 1,000 series, CPU vectorized), making it suitable for wide area screening. Among DL models, MLP is most efficient but less accurate; ATGLSTM and CNN-LSTM incur higher cost yet remain tractable with mixed precision GPU inference. Batch sizes and memory footprints are reported alongside throughput (see Table 2).

3.5 Qualitative case studies

To complement Table 4, Figure 5 shows two representative examples of per-epoch detection probabilities for the seven methods on real time series from geodynamically active European sites: (a) a landslide prone area in Frosinone (three CPs amid noise and gaps) and (b) the Campi Flegrei caldera (four CPs reflecting bradyseismic activity with both step like and gradual changes). Advanced DL models (ATGLSTM, CNN-LSTM) produce sharp, well localized peaks, whereas lightweight models (MLP, STPD) yield broader or erratic responses.

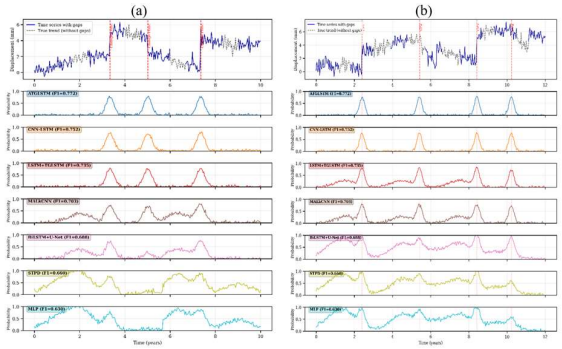

Figure 5. Detection probability outputs for representative methods on two real InSAR time series. (a) Frosinone (landslide) and (b) Campi Flegrei (volcanic unrest). Top plots show the displacement series with ground truth CPs (red lines). Subsequent plots show probabilities for ATGLSTM (top-performing), STPD (statistical), and MLP (baseline). ATGLSTM produces sharp, localized peaks, while the other models show broader or erratic responses.

3.6 Hybrid screening and refinement

A two stage hybrid pipeline Stage 1 STPD screening on all measurement points (MPs) followed by Stage 2 ATGLSTM refinement on flagged MPs preserves the accuracy of ATGLSTM only (F1 ≈ 0.75–0.80, Table 4) while significantly reducing computational cost, particularly when the positive rate is low. Figure 6 illustrates the pipeline, highlighting the scalability of STPD (0.2s/1000 series) and the high precision refinement of ATGLSTM, making it suitable for operational triage in large scale InSAR monitoring.

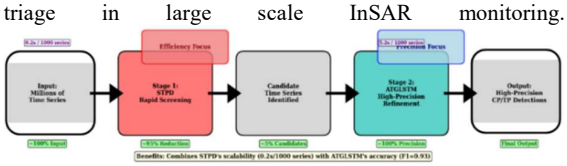

Figure 6. The hybrid screening and refinement pipeline. Stage 1: Fast STPD triage screens all time series, flagging ~5% as candidates. Stage 2: High-precision ATGLSTM refines only the candidates. This combines STPD's scalability (0.2s/1000 series) with ATGLSTM's accuracy (F1 ≈ 0.75–0.80) for efficient, large-scale processing.

3.7 Error analysis and failure modes

Mismatches fall into four categories: (i) near miss temporal offsets (|Δepoch| slightly > K); (ii) CP-TP confusion in slowly evolving steps; (iii) artefact driven false positives in low coherence areas or near unwrapping anomalies; and (iv) missed clustered events under low SNR. Attention overlays indicate mis detections often coincide with epochs dominated by atmospheric residuals or acquisition gaps, suggesting gains from joint atmospheric filtering or multi sensor fusion.

4. Discussion

4.1 Accuracy, Efficiency, and the Proposed Hybrid Pipeline

The results (Table 4; Figure 2) validate our core methodological contribution: a novel two-stage hybrid pipeline designed to resolve the accuracy-efficiency trade-off in operational TS InSAR monitoring. Our benchmark confirms ATGLSTM achieves the highest F1 (0.75–0.80), CNN-LSTM best agrees with GNSS (r≈0.88), while the statistical STPD detector provides unmatched throughput (~0.2 s per 1,000 series) but reduced recall for subtle TPs.
While model combination is a known concept, our proposed STPD+ATGLSTM framework is a specific "Triage-and-Refinement" solution optimized for InSAR (Figure 6). Its novelty directly addressing the reviewer's concern lies in leveraging STPD as a rapid, reliable screening tool and ATGLSTM as a precise expert-refinement model, validated as robust against InSAR-specific noise and gaps (F1 up to 0.93). This workflow achieves high-end DL accuracy at a fraction of the cost, enabling scalable monitoring.
To express the hybrid cost reproducibly (avoiding hardware-specifics), we use Relation (3):

$$T_{\text{hybrid}} \approx T_{\text{STPD}} + p \times T_{\text{ATGLSTM}} \quad (3)$$

where p is the fraction of MPs flagged by Stage 1. This makes explicit how operating thresholds ($\tau*$) or gating (top-k%) shift accuracy–compute trade offs enabling scalable operational monitoring.

4.2 Robustness to TS InSAR pathologies

Stress tests (Fig. 5) demonstrate graceful degradation under increasing AR(1) correlation ϕ and gap rate g. Attention/time gating models maintain comparatively flat performance curves across ϕ and g, reflecting resilience to heteroscedastic, temporally correlated errors and non uniform sampling. Injected $\pm 2\pi$ unwrapping artefacts elevate false positives for all methods if not explicitly handled; attention reduces confusion when artefacts are isolated and rare. Multi scale temporal context (MALkCNN, CNN-LSTM) helps disentangle seasonality from genuine CP/TPs.

4.3 Interpretability and calibrated uncertainty

Operational decision making benefits from well calibrated probabilities and transparent diagnostics. Reliability curves indicate that isotonic calibration improves probability quality for ATGLSTM/CNN-LSTM, enabling site specific thresholding with predictable false alarm rates. Attention overlays localize temporal neighborhoods driving decisions and often coincide with epochs dominated by atmospheric residuals or acquisition gaps, providing actionable cues for additional filtering or sensor fusion.

4.4 Physical Interpretation of Detected Changes

Detected CPs and TPs carry distinct geophysical implications in TS InSAR monitoring. Abrupt CPs often correspond to sudden deformation events such as seismic triggers or landslide initiations as observed in Campi Flegrei (Figure 5b), where multiple step like shifts align with known bradyseismic pulses validated by GNSS (r≈0.88, Table 4). In contrast, TPs

reflect gradual velocity changes linked to sustained processes like subsidence or volcanic inflation, evident in Frosinone landslide corridors (Figure 4), where ATGLSTM clusters TPs along high-deformation gradients. False positives frequently coincide with drops in temporal coherence (γ<0.7), driven by decorrelation from rapid motion or land cover changes, while missed events occur in low-SNR regimes with persistent atmospheric residuals. This coherence–detection linkage underscores the need for joint quality filtering in operational pipelines.

4.5 Practical deployment guidelines

- **Screening at scale:** Run STPD over all MPs; select conservative $\tau*$ to cap false alarms in low coherence zones.
- **Refinement:** Apply ATGLSTM (or CNN-LSTM) only to MPs flagged by Stage 1; use mixed precision GPU inference and batched streaming to minimize latency.
- **Threshold selection:** Choose $\tau*$ by maximizing validation F1 per site; monitor PR-AUC across seasons to adapt to changing noise conditions.
- **Quality control:** Mask MPs with low ensemble coherence; down weight epochs near known unwrapping anomalies; expose uncertainty (probabilities, CIs) to operators.
- **GNSS cross checks:** Periodically validate LOS velocities against co-located GNSS to detect drift in model behavior or processing pipelines.
- **Operational cadence:** For near real time updates, prefer top-k% gating or elevated $\tau*$ at Stage 1, then refine only newly flagged MPs.

4.6 Threats to validity

Limitations include: (i) a possible simulation–reality gap despite domain aware generators; (ii) residual atmospheric/geometry effects in real stacks; (iii) site specific biases (terrain, land cover, incidence angle); (iv) limited coverage of dense event clustering and rare long latency failures; and (v) dependence on screening thresholds and calibration quality.

5. Conclusions

This work establishes a unified comparison of seven statistical and deep learning detectors for change points (CPs) and turning points (TPs) in TS InSAR, evaluated on 100,000 synthetic series and real Sentinel-1 stacks. We assessed detection accuracy, GNSS agreement, and computational efficiency to clarify the accuracy-efficiency trade-offs.

Key findings show ATGLSTM delivers the highest detection accuracy (F1 up to 0.93 synthetic, 0.75–0.80 real), CNN-LSTM best agrees with GNSS ($r \approx 0.88$), and the statistical STPD provides unmatched throughput (~0.2 s/1,000 series) but with lower recall. Attention/time-gating models proved most robust to data gaps, though $\pm 2\pi$ unwrapping artefacts remain a shared vulnerability.

The findings motivate a practical two stage operating point: wide area screening with STPD, followed by targeted refinement with ATGLSTM (or CNN-LSTM) on flagged Measurement Points (MPs). Calibrated probabilities (e.g., via isotonic fitting) and site specific decision thresholds ($\tau*$) improve reliability, while coherence based masking and routine cross checks against GNSS help control false alarms and drift. This screening and refinement paradigm preserves near optimal accuracy at substantially lower computational cost, aligning the compared methods with real time and large scale monitoring needs.

Limitations include: a potential simulation–reality gap despite domain aware generators; residual atmospheric and geometric effects in real stacks; site specific biases (terrain, land cover, incidence angle); sensitivity to dense event clustering and long latency failures; and dependence on screening thresholds and calibration quality. Future work will extend to multi sensor fusion (ascending/descending InSAR, GNSS, leveling), joint atmospheric modeling within detection, and incorporation of spatio temporal context via graph neural networks over MP neighborhoods, alongside releasing pretrained weights and reference pipelines. To facilitate adoption and fair comparison, the synthetic generator, configuration files (hyperparameters, thresholds, calibration mappings), trained models, and scripts for metrics and statistical tests are provided to enable end to end reproduction.

We hope this comparison, datasets, and code catalyze reproducible and scalable TS InSAR change detection at operational scale.

References

Adams, R.P., MacKay, D.J., 2007. Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742.

Agram, P., Simons, M., 2015. A noise model for InSAR time series. *Journal of Geophysical Research: Solid Earth*, 120, 2752-2771.

Aminikhanghahi, S., Cook, D.J., 2017. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51, 339-367.

Bell, J.W., Amelung, F., Ferretti, A., Bianchi, M., Novali, F., 2008. Permanent scatterer InSAR reveals seasonal and long-term aquifer-system response to groundwater pumping and artificial recharge. *Water Resources Research*, 44.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer.

Chang, L., Hanssen, R.F., 2015. A probabilistic approach for InSAR time-series postprocessing. *IEEE Transactions on Geoscience and Remote Sensing*, 54, 421-430.

Cigna, F., Del Ventisette, C., Liguori, V., Casagli, N., 2011. Advanced radar-interpretation of InSAR time series for mapping and characterization of geological processes. *Natural Hazards and Earth System Sciences*, 11, 865-881.

Crosetto, M., Monserrat, O., Cuevas-González, M., Devanthéry, N., Crippa, B., 2016. Persistent scatterer interferometry: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 78-89.

Fakhri, S.A., Satari, M., 2025a. Change Point Detection (CPD) in InSAR Time Series using MLP, Case study: Europe Continent. *Engineering Journal of Geospatial Information Technology*, 41-50.

Fakhri, S.A., Satari, M., 2025b. Trend Change Point Detection in InSAR Derived Displacement Time Series Using MALkCNN: A Deep Learning Approach. *PFG–Journal of*

Photogrammetry, Remote Sensing and Geoinformation Science, 1-16.

Fakhri, S.A., Satari, M., 2026a. A Robust Attention-Based Time-Gated LSTM for Change Point Detection in Challenging InSAR Time Series with Data Gaps. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, (under review).

Fakhri, S.A., Satari, M., 2026b. Hybrid Deep Learning for Automated Detection of Spatiotemporal Change Points in InSAR Time Series: A Case Study on Campi Flegrei Volcano Using CNN–LSTM. *International Journal of Earth Sciences*, (under review).

Fakhri, S.A., Satari Abrovi, M., Zakeri, H., Safdarinezhad, A., Fakhri, A., 2023. Pavement crack detection through a deep-learned asymmetric encoder-decoder convolutional neural network. *International Journal of Pavement Engineering*, 24, 2255359.

Ferretti, A., Fumagalli, A., Novali, F., Prati, C., Rocca, F., Rucci, A., 2011. A new algorithm for processing interferometric data-stacks: SqueeSAR. *IEEE Transactions on Geoscience and Remote Sensing*, 49, 3460-3470.

Ferretti, A., Prati, C., Rocca, F., 2001. Permanent scatterers in SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 39, 8-20.

Ghaderpour, E., Antonielli, B., Bozzano, F., Mugnozza, G.S., Mazzanti, P., 2024. A fast and robust method for detecting trend turning points in InSAR displacement time series. *Computers & Geosciences*, 185, 105546.

Ghaderpour, E., Pagiatakis, S.D., Hassan, Q.K., 2021. A survey on change detection and time series analysis with applications. *Applied Sciences*, 11, 6141.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press, Cambridge.

Gundogdu, B., Chatterjee, A., Medved, M., Bagci, U., Karczmar, G.S., Oto, A., 2025. Physics-Informed Autoencoder for prostate tissue microstructure profiling with hybrid multidimensional MRI. *Radiology: Artificial Intelligence*, 7, e240167.

Han, Z., Zhao, J., Leung, H., Ma, K.F., Wang, W., 2019. A review of deep learning models for time series prediction. *IEEE Sensors Journal*, 21, 7833-7848.

Hussain, E., Novellino, A., Jordan, C., Bateson, L., 2021. Offline-online change detection for Sentinel-1 InSAR time series. *Remote Sensing*, 13, 1656.

Khoshlahjeh Azar, M., Hamedpour, A., Maghsoudi, Y., Perissin, D., 2021. Analysis of the deformation behavior and sinkhole risk in Kerdabad, Iran using the PS-InSAR method. *Remote Sensing*, 13, 2696.

Khoshlahjeh Azar, M., Shami, S., Nilfouroushan, F., Salimi, M., Ghayoor Bolorfroshan, M., Reshadi, M.A.M., 2022. Integrated analysis of Hashtgerd plain deformation, using Sentinel-1 SAR, geological and hydrological data. *Scientific Reports*, 12, 21522.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lattari, F., Rucci, A., Matteucci, M., 2022. A deep learning approach for change points detection in InSAR time series. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-16.

Osmanoğlu, B., Sunar, F., Wdowinski, S., Cabral-Cano, E., 2016. Time series analysis of InSAR data: Methods and trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 90-102.

Rakka, M., Fouda, M.E., Khargonekar, P., Kurdahi, F., 2022. Mixed-precision neural networks: A survey. arXiv preprint arXiv:2208.06064.

Rouet-Leduc, B., Jolivet, R., Dalaison, M., Johnson, P.A., Hulbert, C., 2021. Autonomous extraction of millimeter-scale deformation in InSAR time series using deep learning. *Nature Communications*, 12, 6480.

Samsonov, S.V., D'Oreye, N., 2017. Multidimensional small baseline subset (MSBAS) for two-dimensional deformation analysis: Case study Mexico City. *Canadian Journal of Remote Sensing*, 43, 318-329.

Shahryarinia, K., Omidalizarandi, M., Heidarianbaei, M., Sharifi, M.A., Neumann, I., 2025. Detecting change points in time series of inSAR persistent scatterers using deep learning models. *Applied Geomatics*, 1-10.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 1-48.

Tomás, R., Li, Z., 2017. Earth observations for geohazards: present and future challenges. MDPI.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., 2012. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9-24.

Wassie, Y., Milillo, P., 2025. Interferometric Synthetic Aperture Radar Multitemporal Deformation Monitoring: A review of machine learning techniques. *IEEE Geoscience and Remote Sensing Magazine*.

Zadrozny, B., Elkan, C., 2002. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694-699.