

## Title

A Siamese U-Transformer for change detection on MRI brain for multiple sclerosis, a model development and external validation study

## Authors

**Brendan S Kelly**<sup>1-4</sup> (Corresponding Author), Prateek Mathur<sup>2</sup>, Ronan P Killeen<sup>1</sup>, Aonghus Lawlor<sup>2</sup>

## Affiliations

1. St Vincent's University Hospital, Dublin, Ireland
2. Insight Centre for Data Analytics, UCD, Dublin, Ireland
3. Wellcome Trust – HRB, Irish Clinical Academic Training, Dublin, Ireland
4. School of Medicine, University College Dublin, Dublin, Ireland

## Summary

### Background

Multiple Sclerosis (MS), is a chronic idiopathic demyelinating disorder of the CNS. Imaging plays a central role in diagnosis and monitoring. Monitoring for progression however, can be repetitive for neuroradiologists, and this has led to interest in automated lesion detection. Simultaneously, in the computer science field of Remote Sensing, Change Detection (CD), the identification of change between co-registered images at different times, has been disrupted by the emergence of Vision Transformers. CD offers an alternative to semantic segmentation leveraging the temporal information in the data.

### Methods

In this retrospective study with external validation we reframe the clinical radiology task of new lesion identification as a CD problem. Consecutive patients who had MRI studies for MS at our institution between 2019 and 2022 were reviewed and those with new lesion(s) were included. External data was obtained from the MSSEG2 challenge and OpenMS. Multiple CD models, and a novel model (NeUFormer), were trained and tested. Results were analysed on both paired slices and at the patient level. Expected Cost (EC) and F2 were independently and prospectively chosen as our primary evaluation metrics. For external data we report DICE and F1 to allow for comparison with existing data. For each test set 1000 bootstrapping simulations were performed by sampling 10 patient samples with replacement giving a non parametric estimate of the confidence interval. Wilcoxon statistics were calculated to test for significance.

### Findings

43,440 MR images were included for analysis (21,720 pairs). The internal set comprised of 170 patients (110 for training, 30 for tuning, 30 testing) with 120 females and 50 males, average age of 42 (range 21 – 74). 60 (40 + 20) patients were included for external validation.

In the CD experiments (2D) our proposed NeuFormer model achieved the best (lowest) Expected Cost (EC) ( $p=0.0095$ ), the best F2 and second best DICE ( $p<0.0001$ ). At the patient level our NeuFormer model had the joint highest number of True Positive lesions, and lowest number of False negatives ( $p<0.002$ ). For CD on external data, NeuFormer achieved the highest DICE on both datasets ( $p<0.0001$ ). NeuFormer had the lowest or joint lowest number of False Positives on external data ( $p<0.0001$  in all cases).

### Interpretation

Reformulating new lesion identification as a CD problem allows the use of new techniques and methods of evaluation. We introduce a novel Siamese U-Transformer, NeUFormer, which combines concepts from U-Net, Siamese Networks, and vision transformers to create a model with improved small lesion detection and the consistently best EC. Its ability to increase detection of small lesions, balanced with relatively few false positives, and superior generalisability has the potential to greatly impact the field of the identification of radiologic progression of MS with AI.

## **Research in context**

### **Evidence before this study**

Multiple Sclerosis (MS), a chronic and idiopathic demyelinating disorder of the CNS, is diagnosed using the McDonald criteria based on MRI interpretation. Without a definitive MS biomarker, AI holds promise is for uncovering unique features indicative of MS, improving diagnostics and identifying progression. Research in the field typically centres on segmentation and classification, leaving a gap in evaluating temporal imaging changes. The MSSEG2 challenge has now enabled more research into new lesion identification in MS. Even so, most solutions are based on semantic segmentation architectures and rely on limited metrics for evaluation. The identification of small lesions also remains a challenge.

Remote Sensing (RS) is the science of obtaining information about objects or areas from a distance, typically from aircraft or satellites. In the RS literature, Change Detection (CD) refers to the identification of significant alterations in co-registered images captured at different times. In this way CD offers an alternative to semantic segmentation leveraging the temporal information in the data. This field was dominated by convolutional neural networks but has recently been disrupted by transformer-based architectures. Transformers, fuelled by their success in NLP, are gaining popularity across all computer vision tasks due to their larger effective receptive field and enhanced context modelling between image pixels. Inspired by these developments, we incorporate some of these ideas into our NeUFormer model.

### **Added value of this study**

This study redefines the task of identifying progression on MRI brain in MS as a CD problem, borrowing concepts from RS. This approach allows for both pixel- and patient-level evaluation and rethinks standard metrics to suit specific clinical needs. This acknowledges the distinction between trivial variation in segmentation and clinically significant change. State-of-the-art CD models are assessed at this task, and a novel model, NeuFormer, is introduced. NeuFormer synergistically combines concepts from the classical U-Net (which was originally intended for brain segmentation), Siamese architecture adaptations specifically for CD, Swin-UNETR (a U-Transformer developed by MONAI to integrate the shifting window structure of the Swin transformer into medical imaging) and ChangeFormer which also uses attention at scale specifically for CD, leveraging improved spatiotemporal reasoning to create a model which is better for small lesion identification and with the consistently lowest EC associated with its decisions.

### **Implications of all the available evidence**

Reframing lesion identification as CD enables an alternative to semantic segmentation leveraging the temporal information in the data, enhancing the model's relevance and customization for specific medical tasks. We also propose the flexible Expected Cost metric, as it facilitates varying action thresholds and helps to customise tools to stakeholder preferences.

Siamese vision transformers show promise for CD on MRI in MS including for smaller lesions which are traditionally difficult for computer vision models to identify. This may be to the intrinsic spatiotemporal advantages of vision transformers, with positional embedding, over patch based convolutional methods. NeUFormer's ability to increase detection of small lesions, balanced with relatively few false positives and excellent generalisability has the potential to greatly impact the field of the identification of radiologic progression of MS with AI.

## **Main Text**

### **Introduction**

Multiple Sclerosis (MS), a chronic and idiopathic demyelinating disorder of the CNS, is diagnosed using the McDonald criteria based on MRI interpretation<sup>1,2</sup>. MS differs from many chronic diseases as its imaging features may precede clinical symptoms. As such, imaging plays a central role in diagnosis, tracking progression, and evaluating treatments<sup>1,2</sup>. New T2/FLAIR MS lesions are the primary biomarker for assessing both disease progression and medication response<sup>3,4</sup>. Indeed, lack of new lesions in the CNS is a key indicator of medication effectiveness<sup>5</sup>. Monitoring these lesions, however, is often monotonous and repetitive for neuroradiologists<sup>6</sup>, and this combined with radiology's supply-demand challenges<sup>7</sup> has led to increased

interest in methods for automating lesion detection<sup>8</sup>. Over the past two decades, research has heavily focused on computer-assisted segmentation methods<sup>8</sup>, with a recent surge in AI methodologies<sup>9</sup>. Current research trends are evolving from simple identification of MS lesions on T2/FLAIR to analysing images over different times<sup>8</sup>. The MSSEG2 challenge by The Medical Image Computing and Computer Assisted Intervention Society (MICCAI), targeting new lesion detection, has significantly boosted interest in this research domain<sup>3</sup>. Significant issues remain in the literature with only modest performance reported for new lesion identification<sup>10</sup> and there are well known difficulties to correctly evaluate stable cases (which outnumber change cases significantly)<sup>11</sup>. Furthermore, identification of smaller lesions remains problematic and is an outstanding issue<sup>10</sup>. The limited kernel size of convolution layers in Fully Convolutional Neural Networks (FCNNs)<sup>12</sup> can result in sub-optimal performance in modelling long-range spatial information<sup>13</sup>. This can adversely affect the segmentation of lesions of varying sizes<sup>12</sup>.

Remote Sensing (RS) is the science of obtaining information about objects or areas from a distance, typically from aircraft or satellites<sup>14</sup>. In the RS literature, Change Detection (CD) identifies significant alterations in co-registered images captured at different times<sup>15</sup>. CD offers an alternative to semantic segmentation leveraging the temporal information in the data. The type of change varies by application, including alterations in man-made structures, vegetation, and environmental shifts like polar ice cap melting or deforestation. Effective CD models distinguish these changes while filtering out irrelevant variations due to seasonal shifts, shadows, atmospheric changes, and lighting differences<sup>14</sup>. Current leading CD methods primarily use deep convolutional networks (ConvNets) for their strong feature extraction capabilities<sup>13</sup>. Recently, the success of Transformers in Natural Language Processing (NLP) has inspired their application in multiple other domains, and especially in computer vision tasks<sup>14</sup>. Attention-based models such as the original Vision Transformer<sup>16</sup> and more recently SAM/SAM-Med<sup>17</sup> and SWIN<sup>18</sup> continue to improve performance on standard computer vision challenges. This is in part due to their larger effective receptive field than deep ConvNets, enhancing context modelling between image pixels<sup>13</sup>. Indeed, a Siamese vision transformer (ChangeFormer) has reached state of the art performance in CD challenges<sup>14</sup>. There have been some recent medical applications of vision transformers in the medical imaging literature, especially in the domain of semantic segmentation<sup>19,20</sup>. However there are very few works which use these models to leverage the temporal information in these data to enhance the identification and classification of change over time<sup>21,22</sup>. For example the nnU-Net, a leading segmentation model often considered state of the art for MS lesion segmentation, does not use the longitudinal nature of the scans and treats each image individually during training and inference<sup>23</sup>.

Emerging research demonstrates that “task set-up” or “problem formulation” can have significant impact on the result of medical AI experiments<sup>24</sup>. In this study we reframe the clinical radiology issue of new MS lesion classification on MRI as a CD problem inspired by the RS literature. Here we consider only new or enlarging lesions (as defined by the MAGNIMS criteria<sup>25</sup>) as relevant change. This enables us to consider different evaluation metrics that consider the cost of decisions made by the models rather than just segmentation performance (e.g. DICE score) or lesion identification (e.g. accuracy or F1 score), which are important, but not task specific<sup>26</sup>. The “Expected Cost” (EC) allows individually chosen weights for the error rates (such that missing a lesion can be penalized more) and can be made prevalence-independent. We can evaluate performance on both co-registered slices and the whole 3D stack allowing for scrutiny at both the pixel and patient level. We also describe a novel model for **New** lesion identification, a Siamese **U-Transformer**, abbreviated as NeUFormer.

## Methods

### Study design and data sources

This retrospective study with external validation was designed according to both RSNA and ESR published principles<sup>27,28</sup> with patient expert involvement<sup>29</sup>. The EC was independently and prospectively chosen as a suitable primary evaluation metric<sup>26</sup>. The manuscript was prepared using the CLAIM checklist<sup>30</sup>. We received full IRB approval and the requirement for prospective consent was waived. This research constitutes Level 5A evidence (Data quality and AI model development with external testing) as it represents one retrospective study with internal and external data used for final reporting<sup>31</sup>. Subsections of this cohort have been published previously but under different experimental conditions with different research questions. The external cohort is publicly available and has been previously described<sup>3</sup>. All internal data are available from the authors upon reasonable request and the code to recreate these experiments is available on github.

Consecutive patients who had at least two MRI brain studies for MS at our institution between January 2019 and December 2022 were reviewed (Figure 1). Those with a new lesion on follow up imaging were included in this study. Images were acquired on a 1.5 T system (SIEMENS MAGNETOM Avanto syngo MR B19, SIEMENS, Munich, Germany). Imaging sequences included a three-dimensional T2 fluid-attenuated inversion-recovery (FLAIR) sequence using the following parameters: acquired voxel size, 1.1 x 1.1 x 1.1 mm; TR 6000 ms; TE 413 ms; TI 2030ms; acquisition time 6 mins 44 s; orientation, sagittal. We used all publicly available data from MSSEG-2<sup>3</sup> and OpenMS<sup>32</sup> to externally validate our findings.

#### Ground truth labelling

A baseline automated segmentation of MS lesions was generated using DeepMedic<sup>33</sup>. These baseline segmentations were then manually corrected by one of two certified radiologists in their first year post board examination using ITK Snap V3.8.0<sup>34</sup>. Cases with progression were initially identified from the radiologic report, and confirmed at a dedicated research re-read. Radiologic progression (new or enlarging lesions) was defined according to the MAGNIMS consensus guidelines<sup>25</sup>. Specifically for new lesions the largest linear measurement for lesion definition had to be 3 mm or more in at least one plane. In the case of enlarging lesions, subtractions of co-registered intensity normalized images were used to confirm that the lesion had unequivocally enlarged. Cases with progression were first segmented and manually corrected as above and then additionally verified by a third radiologist who is a subspecialist neuroradiologist with over 10 years post fellowship experience.

#### Image processing

The raw MRI sequences acquired from the scanner were in DICOM format which were anonymized by removing any identifiable information pertaining to the patient or the practitioner. The anonymized DICOM slices were then converted to the NiftI format using the dcm2niix utility (v1.0.20220720). The produced images were then rigidly registered to the first T1 sequence using the FMRIB's Linear Registration Tool (FLIRT). Rigid body registration was applied with 6 degrees of freedom, no angular search and spline interpolation, rest of the properties were set to their default values. After registration, the FSL Brain Extraction Tool (BET) was applied with a fractional intensity threshold of 0.4 to the first T1 of all patients. This mask was then applied to the remaining scans for each patient respectively. Bias Field correction using the FSL FAST utility was not performed after no contrast improvement was empirically observed. Since the MRI scans in our dataset were acquired from the same Scanner, the need for standardisation was diminished.

Patients with more than two FLAIR images were chosen. Each extracted slice is cropped to the largest brain cross-section while maintaining the image aspect ratio. The cropped slice is then intensity normalised and rescaled to 256x256. A Contrast Limited Histogram Equalization is applied to the cropped slices to enhance the tissue contrast. Slices with no brain volume are discarded. The lesion masks follow a similar pre-processing pipeline except for the Histogram Equalization step. In addition, partial lesions which are too small to be considered progression are removed before the final labels are produced. This is because lesions less than 3mm are not considered to represent progression in the MAGNIMS criteria<sup>25</sup> and is standard across similar tasks<sup>3,11</sup>. The image slices and the difference maps are generated to conform to the usual data structures in CD challenges. Data were partitioned at the patient level into training, validation/tuning and test sets in a ratio of approximately 65:17.5:17.5.

#### Model development and training

A novel Siamese U-Transformer (NeUFormer, NeU) was developed (Figure 2). The network is inspired by the "Swin UNETR"<sup>12</sup> semantic segmentation model, and modified to create a Siamese architecture via additional skip connections at each subsampling scale, with a classification head for change detection. The aim here is to enhance the more abstract and less localized information from the later-stage encoded data with spatial details which are learned in the network's earlier layers. Uniquely for NeUFormer, these skip connections also integrate information from different resolutions. Through this we aim to create a model with improved spatiotemporal awareness, primed for superior change detection.

The input to the network is two 256x256 images and the output is a change map. Each arm consists of a "Swin UNETR" encoder (separated into two streams of equal structure with shared weights as in a traditional

Siamese network) and decoder. The encoder has 4 stages with 2 transformer blocks at each stage. A modified 2D version of the Swin UNETR 3D shifting window self-attention mechanism is employed, allowing for feature extraction at five different resolutions (Figure 2). For the decoder, NeUFormer also has a U-Shaped FCNN design, with the encoded feature representations used by the decoder at each resolution via skip connections. Extending this concept, the feature representations are concatenated together before and after the encoding part of the architecture via additional “Siamese” skip connections<sup>35</sup>. This is done for each arm and at each resolution, with the goal of producing precise class predictions. We leverage synergy of the Swin transformer’s shifting windows giving hierarchical attention at different scales<sup>18</sup>, the U-Net’s proven ability at biomedical imaging semantic segmentation<sup>36</sup>, and the discriminative power of the Siamese architecture for change detection<sup>35</sup>.

For comparison, baseline models included a fully connected Bitemporal U-Net (BUN) (that concatenates bi-temporal images and passes to a ConvNet to detect changes), two different Siamese U-Nets (which extracts multi-level features of bi-temporal images from a Siamese ConvNet and either concatenates (SUC) them or finds their difference (SUD))<sup>35</sup>, a bitemporal vision transformer (ViT) (which uses a transformer encoder-decoder network to enhance the context-information of ConvNet features via semantic tokens followed by feature differencing to obtain the change map)<sup>13</sup>, Swin UNETR (UTR) (which replaces the encoder of a vision transformer with a U-Net)<sup>12</sup> and ChangeFormer (CFR) (which utilises a transformer encoder with Multi-Layer Perception decoder in a Siamese network architecture to efficiently render multi-scale long-range details required for accurate CD)<sup>14</sup>. All models were implemented in PyTorch and trained from scratch (without pretraining) using an NVIDIA GeForce GTX 1080 GPU. Data augmentation was performed with random flip, random re-scale, random crop and Gaussian blur. Models were trained using a combined weighted Cross-Entropy and DICE loss using AdamW optimizer and a batch size of 8. Due to biased classes (only a small % of image was made up by the target (change) class) images with change were used for training along with a matched number of stable controls in a ratio of 1:1.

## Analysis

For the change detection task we completed problem fingerprints for our research questions<sup>26</sup>. Due to the unequal severity of class confusions (with false negatives rated as more important) and an existence based penalization of special outliers the EC (Multi-Class) and F2 score (Per-Class) were chosen prospectively and independently as the primary evaluation metrics for our study<sup>26,37</sup>. EC is a measure of accuracy that can incorporate different disease prevalence and account for differences in impact between false positives and false negative predictions. The closer to 0 the better the expected cost; while scores of >1 are possible, the indicate a futile model. Following a stakeholder focus group a cost matrix of  $[[0.0, 10.0], [1.0, 0.0]]$  was used to calculate the EC<sup>29</sup>. This penalises FNs 10 times higher than FPs. We also report the precision, recall, F0.5, F1 and DICE scores to allow for comparison with other studies. For the DICE, a score of 1.0 was recorded both where a model correctly predicted the positive change map or a blank change mask as many cases have “no change”. We also evaluate on small lesions (those with segmentation masks with only connected components less than 100 pixels) separately as this is a consistently a problem in the literature. The 100 pixel connected component limit was chosen as this is the limit at which segmentation models have been shown to deteriorate for 256x256 images<sup>17</sup>. For each test set 1000 bootstrapping simulations were created by sampling 10 patient samples with replacement giving a non parametric estimate of the confidence interval. Wilcoxon statistics were calculated to test for significance. To minimise hypothesis testing only the best and second best model were compared for primary evaluation metrics (EC, F2, F1, DICE).

We also evaluate our models on external data (MSSEG-2<sup>3</sup> and OpenMS<sup>32</sup>) to test their generalisability and potential clinical applicability. For both cases only the training data could be accessed, and it was then processed according to the same pipeline as the internal data. The mix of cases that are positive and negative for progression in the external data allows us to report accuracy, sensitivity and specificity at the patient level (3D). Total True Positive, True Negative and False Positive lesions are reported for all datasets.

## Role of the funding source

The funding source for this study had no role in the experimental design of the study, data collection, data analysis, data interpretation, or writing of this report.

## Results

In total 43,440 MR images were included for analysis (21,720 pairs including train, tuning, test and external data). The internal set comprised of 170 patients (110 used only for training, 30 for tuning and 30 testing) and the external set comprised of 60 patients for external validation (40 from MSSEG2 and 20 from OpenMS) (Figure 1). There were 120 females and 50 males in the internal data with an average age of 42 (range 21 – 74). Demographics are summarised in Table 1. Due to the nature of the condition, the target (change) class was underrepresented in the images with the target class representing only 0.0001% of the total pixels (see Table 2).

For the two primary evaluation metrics, NeuFormer achieved the best (lowest) EC of 0.467 ( $p=0.0095$ ) and had the highest F2 score 0.329 (Figure 3). NeuFormer also achieved the lowest EC and F2 for small lesions (lesions with connected components less than 100 pixels). Results of the CD experiments are summarised in Table 3A for all lesions and Table 3B for small lesions only. Additionally Figure 4 shows the qualitative difference in output for the different models. For patient level evaluation, NeuFormer had the joint highest number of True Positive lesions ( $p=0.0011$ ) and lowest number of False negatives ( $p<0.0001$ ) (Table 4).

For external validation we focus on the widely used metrics of DICE and F1 to allow for comparison. Two external datasets MSSEG2 and OpenMS were analysed. MSSEG2 has both progression and stable cases allowing for analysis of progression identification at the patient level. NeuFormer consistently performed well across both datasets, achieving the highest DICE on both ( $p<0.0001$ ), the second highest F1 for OpenMS, and third highest F1 for MSSEG2 (Table 5 A and B). For patient level evaluation on external data the highest progression accuracy was achieved by UTR for the MSSEG2 dataset but the model had 723 False Positives (compared to 22 by the best performing model NeuFormer) with a specificity of 0.0 for progression Table 6.

## Discussion

In this study we reformulate the problem of diagnosis progression on MRI brain in MS into a CD problem inspired by the RS literature. This allows us to evaluate model performance at both the pixel and patient level, and rethink the existing evaluation methods, choosing metrics relevant to the specific problem space. In this way, we can consider which model has the lowest clinical cost associated with its decisions rather than just an evaluation of its segmentation performance. This acknowledges the distinction between trivial variation in segmentation and clinically significant change. We evaluate state of the art CD models and we introduce a novel model, NeuFormer, which synergistically combines concepts from the classical U-Net, Siamese architectures, and Vision Transformers with Shifting windows to create a model with the consistently lowest EC associated with its decisions. In particular, it also has the lowest cost and highest recall and F2 performance when examining small lesions, and demonstrates robust generalisability to external data. Its ability to increase detection of small lesions, balanced with relatively few false positives, has the potential to greatly impact the field of the identification of radiologic progression of MS with AI. Our two primary evaluation metrics were expected cost and F2 score. These metrics emphasise the importance of minimising False Negatives as in our use case this could mean missing an opportunity to change or initiate a treatment. We also consider F1 and DICE, especially for the external data, as these metrics are more commonly used in the literature related to the those datasets.

In our internal dataset, NeuFormer performs best in terms of EC and F2, for both all lesions and small lesions only. Its F1, F0.5 and precision all showed a relative increase in performance for small lesions showing that in addition to its ability for find small lesions (high recall), it also had relatively fewer false positives (better precision). The generalisability of NeuFormer is clear from its performance on OpenMS and on MSSEG2 achieving the best DICE score for both datasets. It has only the third best F1 on MSSEG2 (after UTR and BUN) and second best on OpenMS (after SUD). This shows good performance across the datasets, not dropping out of the top 3 despite the heterogenous data. UTR achieved the best F1 on MSSEG2, however it predicted change on every slice, rendering it useless clinically. When we consider the results at a patient level, our model had both the highest number of TPs and the lowest FPs on internal data. For MSSEG-2 NeuFormer again had the lowest number of False positives, and the third highest TPs (after UTR and SUD). However again due to UTRs high number of FPs (723) it is not usable functionally as a clinical decision model. The other models with

comparable performance SUC and SUD also had higher FPs (88 and 192 respectively). This was also borne out in the second external dataset OpenMS where again our model had the fewest FPs.

We believe our models performance and generalisability are in part due to the intrinsic advantages of vision transformers in terms of spatiotemporal awareness, especially when synergistically combined with a U-Net inspired encoder module tried and tested in the medical image domain. Then our novel model design which combines the sensitive UTR encoder with the intrinsically discriminative Siamese architecture provides a balance of consistently identifying clinically relevant changes without excessive FPs. Trivial changes in semantic segmentations are often attributable to differences in inter- and intra- user variability, the CD paradigm allows us to focus on these relevant changes when evaluating the models. In this context the importance of the NeUFormer's discriminative ability becomes clear, allowing the identification of progression at the patient level without overburdening a decision system with FPs.

Our study has several limitations. The retrospective study design limits the level of evidence. Furthermore, as our internal experiments only involved those patients with progression, there is a selection bias. While this is a common issue in clinical radiology research<sup>38</sup> it remains a clear limitation. Another key issue was the class imbalance. The target class is very underrepresented in the problem of change detection in MS. Indeed in our datasets only 0.0001% of the pixels were in the target class. For this reason it is necessary to oversample the patients with change in the training process. Even with our oversampling method, we still need to use a Cross Entropy loss weighted 1000:1 in favour of the change class to get the model to make any predictions of change. Since the models are agnostic to the nature of the input images, the slice wise approach provides samples of both cases with and without change. A patient-level control approach will only add to an already imbalanced dataset, making the problem of under representation of the target class worse. We then include stable patients in the evaluation to ensure the model works across different groups.

Our sample size was modest, but 170 pairs with change for training compares favourably to MSSEG2 (the largest existing dataset) which comprised only 100 pairs of patients in total, of whom half were stable. Images were resized to 256x256 and primarily assessed in 2D. While there is loss of the 3D information, our model does integrate the temporal information intrinsic to the interpretation task. Due to positional embedding the transformer based methods can interrogate the whole slice rather than subsections. The alternative models are mostly patch based and thus do not take the whole 3D volume en-bloc. In this way a trade-off is necessary at one level of abstraction or another. The demographic information for internal and external data does not contain information on ethnicity, meaning subgroup analysis was not possible. While the internal data was acquired at one institution on one MRI scanner, the external data was heterogenous with different nations, hospitals, protocols, scanners and even a mix of 1.5 and 3 Tesla magnets.

Prospective evaluation would be necessary before our model could be implemented into clinical practise. The robust external validation results show that the method holds promise but practical implementation remains a challenge. Due to ethical considerations and patient preference an autonomous change detection model it is not currently feasible or desirable<sup>29</sup>. Therefore we propose that NeUFormer could be used to screen for studies likely to contain significant change and be used to triage radiologist workflow. As the output of the model is a binary change map (Figure 4), an intuitive explanation is given to the interpreting physician. Furthermore scalability and computational performance is less of an acute issue than in use cases such as stroke as the proposed use case refers mainly to outpatient imaging. Our approach holds further potential clinical utility outside of MS, in fields where monitoring for change is crucial, such as oncologic imaging, treatment response evaluation, and screening programs.

In summary, reformulating new lesion identification as a CD problem allows the use of new techniques and methods of evaluation. We expect this format to be used to drive new innovations in AI in MS imaging, and that the experimental design can be easily applied to other diseases and modalities. Our novel NeUFormer model combines concepts from U-Net, Siamese Networks, and vision transformers to create a model with the consistently lowest cost associated with its decisions, including for smaller lesions and has the potential to screen for progression of MS on MRI brain.

## Figures

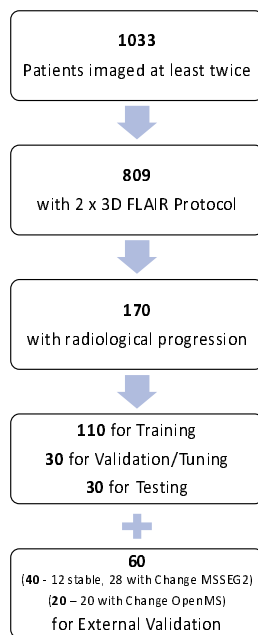


Figure 1:  
Title: Patient Flowchart  
Legend: Flowchart of patients included in our study

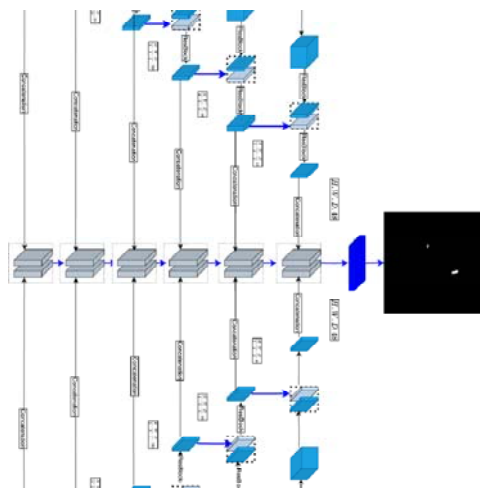


Figure 2:  
Title: Model Diagram NeUFormer  
Legend: Diagrammatic representation of the NeUFormer model. (Navy cuboid = Bottleneck Feature/Head, Light Blue cuboid = Hidden Feature, Navy arrow = Deconvolution, broken line = concatenation.)



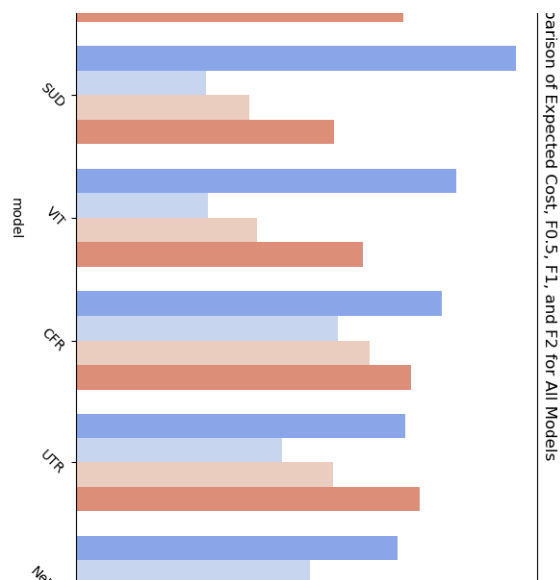


Figure 3:

Title: Quantitative Results 2D

Legend: Comparison of the performance metrics for all models on the internal data (2D). Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).



Figure 4:

Title: Qualitative Results

Legend: Qualitative comparison of different CD methods on the internal data (A), MSSEG2 (B) and OpenMS (C).

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

## Tables

Table 1 Patient Demographics

Total Participants	Average Age	Age Range	
170			
	42.3	Minimum	21
		Maximum	74
Gender	Gender Count	Change Instances	
Male	50	212	
Female	120		
MRI Studies Range	MRI Studies	Total MRI Studies	MRI Studies Average
Min	2	496	
Max	5		2.9

Table 2 Prevalence of the change class (progression) in the internal test and external data

Test Set Lesions	Patients	Images (256x256)	Number positive images	Number Positive Pixels (%)	Lesion dimension (pixels): Mean, StD (range)
Internal	30	6081	437	43823 (0.000109%)	164.71, 111.42 (14 - 407)
MSSEG2	28 (of 40)	7361	629	53408 (0.00011%)	53.95, 52.44, (6 - 360)

Table 3A CD Results internal data, all lesions

Model	Expected Cost	DICE	Precision	Recall	F0.5	F1	F2
BUN	0.546	<b>0.932*</b>	<b>0.451</b>	0.494	<b>0.459</b>	<b>0.472*</b>	0.485
SUC	<b>0.507</b>	0.886	0.244	0.624	0.277	0.350	0.476
SUD	0.638	0.848	0.161	0.559	0.187	0.249	0.374
VIT	0.551	0.873	0.161	<b>0.693</b>	0.190	0.261	0.417
CFR	0.531	<b>0.887</b>	<b>0.353</b>	0.535	<b>0.379</b>	<b>0.426</b>	<b>0.485</b>
UTR	<b>0.478</b>	0.851	0.262	<b>0.644</b>	0.297	0.372	<b>0.498</b>
NeU-C	<b>0.467*</b>	<b>0.915</b>	<b>0.302</b>	<b>0.632</b>	<b>0.337</b>	<b>0.409</b>	<b>0.519*</b>

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

\*denotes statistical significance

Table 3B CD Results internal data, all lesions

Model	Expected Cost	DICE	Precision	Recall	F0.5	F1	F2
BUN	0.666	<b>0.940*</b>	<b>0.181</b>	0.363	<b>0.201</b>	<b>0.241*</b>	<b>0.302</b>
SUC	<b>0.621</b>	<b>0.894</b>	0.081	0.475	0.097	0.139	0.241
SUD	0.670	0.855	0.049	0.503	0.060	0.089	0.176
VIT	0.663	0.881	0.048	<b>0.519</b>	0.059	0.088	0.176
CFR	0.665	0.893	<b>0.118</b>	0.386	<b>0.137</b>	0.181	0.265
UTR	<b>0.527</b>	0.858	0.114	<b>0.549</b>	0.135	<b>0.189</b>	<b>0.311</b>
NeU-C	<b>0.516*</b>	<b>0.922</b>	<b>0.125</b>	<b>0.552</b>	<b>0.148</b>	<b>0.205</b>	<b>0.329*</b>

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

\*denotes statistical significance

Table 4 Patient level results internal data

Model	TP	FP	FN	Precision	Recall	F0.5	F1	F2
BUN	46	<b>75</b>	48	<b>0.380</b>	0.489	<b>0.398</b>	<b>0.428*</b>	<b>0.463</b>
SUC	<b>60</b>	493	36	0.108	0.625	0.130	0.185	0.320
SUD	<b>60</b>	1241	38	0.046	0.612	0.057	0.086	0.177
VIT	56	520	41	0.097	0.577	0.117	0.166	0.290
CFR	<b>65</b>	<b>485</b>	<b>32</b>	<b>0.118</b>	<b>0.670</b>	<b>0.141</b>	<b>0.201</b>	<b>0.346</b>
UTR	<b>67</b>	1058	<b>31</b>	0.060	<b>0.684</b>	0.073	0.110	0.221
NeU	<b>67</b>	<b>195</b>	<b>30</b>	<b>0.256</b>	<b>0.691</b>	<b>0.293</b>	<b>0.373</b>	<b>0.515*</b>

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

\*denotes statistical significance

Table 5A External CD Results (MSSEG-2)

Model	Expected Cost	DICE	Precision	Recall	F0.5	F1	F2
BUN	0.527	<b>0.932</b>	<b>0.711</b>	0.486	<b>0.650</b>	<b>0.577</b>	0.518
SUC	<b>0.413</b>	0.928	0.520	<b>0.624</b>	0.538	0.567	<b>0.600</b>
SUD	<b>0.375*</b>	0.914	0.452	<b>0.678</b>	0.485	0.543	<b>0.617</b>
VIT	0.499	0.924	0.408	0.554	0.431	0.470	0.517
CFR	0.615	<b>0.928</b>	<b>0.810</b>	0.391	<b>0.667</b>	0.527	0.436
UTR	<b>0.388</b>	0.871	0.532	<b>0.649</b>	0.552	<b>0.585*</b>	<b>0.622*</b>
NeU	0.549	<b>0.941*</b>	<b>0.747</b>	0.461	<b>0.665</b>	<b>0.570</b>	0.500

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

\*denotes statistical significance

Table 5B External CD Results (OpenMS)

Model	Expected Cost	DICE	Precision	Recall	F0.5	F1	F2
BUN	1.893	<b>0.626</b>	<b>0.756</b>	0.313	<b>0.778</b>	0.442	0.341
SUC	0.878	<b>0.599</b>	0.554	0.391	0.654	<b>0.458</b>	<b>0.481</b>
SUD	<b>0.794</b>	0.591	0.536	<b>0.409</b>	0.636	<b>0.464*</b>	<b>0.494*</b>
VIT	0.886	0.582	0.385	<b>0.420</b>	0.398	0.402	0.406
CFR	<b>0.826</b>	0.590	<b>0.694</b>	0.275	<b>0.666</b>	0.394	0.341
UTR	0.939	0.560	0.522	<b>0.401</b>	0.615	0.454	<b>0.480</b>
NeU	<b>0.786*</b>	<b>0.632*</b>	<b>0.659</b>	0.353	<b>0.665</b>	<b>0.459</b>	0.370

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

\*denotes statistical significance

Table 6 Patient level results MSSEG-2

Model	TP	FP	FN	Precision Progression	Recall Progression	Specificity Progression	Accuracy Progression
BUN	44	<b>39</b>	93	0.451	<b>0.793</b>	0.500	0.821
SUC	<b>56</b>	88	<b>84</b>	<b>0.491</b>	0.771	<b>0.667</b>	<b>0.964</b>
SUD	<b>56</b>	192	<b>81</b>	<b>0.481</b>	<b>0.839</b>	0.417	<b>0.929</b>
VIT	40	79	93	0.462	0.727	<b>0.750</b>	0.857
CFR	43	<b>66</b>	96	0.440	0.688	<b>0.833</b>	0.786
UTR	<b>74</b>	723	<b>71</b>	<b>0.500</b>	<b>1.000</b>	0.000	<b>1.000</b>
NeU-C	<b>48</b>	<b>22</b>	92	<b>0.481</b>	0.743	<b>0.750</b>	<b>0.929</b>

Red is best, blue second best and bold third best

Bitemporal U-Net (BUN), Siamese U-Net with Concatenation (SUC), with Difference (SUD), bitemporal Vision Transformer (VIT), ChangeFormer (CFR), Swin UNeTR (UTR), NeUformer (NeU).

## References:

- 1 McNamara C, Sugrue G, Murray B, MacMahon PJ. Current and Emerging Therapies in Multiple Sclerosis: Implications for the Radiologist, Part 1—Mechanisms, Efficacy, and Safety. *Am J Neuroradiol* 2017; **38**: 1664–71.
- 2 McNamara C, Sugrue G, Murray B, MacMahon PJ. Current and Emerging Therapies in Multiple Sclerosis: Implications for the Radiologist, Part 2—Surveillance for Treatment Complications and Disease Progression. *Am J Neuroradiol* 2017; **38**: 1672–80.
- 3 O C, F C, F C, M. D. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure - Archive ouverte HAL. International Conference on Medical Image Computing and Computer Assisted Intervention, 2021. <https://hal.inria.fr/hal-03358968/> (accessed Feb 11, 2023).
- 4 Commowick O, Istace A, Kain M, *et al.* Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Sci Rep* 2018; **8**: 13650.
- 5 Yang J, Hamade M, Wu Q, *et al.* Current and Future Biomarkers in Multiple Sclerosis. *Int J Mol Sci* 2022; **23**: 5877.
- 6 Martin D, Tong E, Kelly B, Yeom K, Yedavalli V. Current Perspectives of Artificial Intelligence in Pediatric Neuroradiology: An Overview. *Front Radiol* 2021; **1**: 713681.
- 7 Kelly BS, Judge C, Bollard SM, *et al.* Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *Eur Radiol* 2022; : 1–10.
- 8 Diaz-Hurtado M, Martínez-Heras E, Solana E, *et al.* Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. *Neuroradiology* 2022; **64**: 2103–17.
- 9 Filippi M, Preziosa P, Arnold DL, *et al.* Present and future of the diagnostic work-up of multiple sclerosis: the imaging perspective. *J Neurol* 2022; : 1–14.
- 10 Wagner DT, Tilmans L, Peng K, *et al.* Artificial Intelligence in Neuroradiology: A Review of Current Topics and Competition Challenges. *Diagnostics* 2023; **13**: 2670.
- 11 Basaran BD, Matthews PM, Bai W. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. *Front Neurosci* 2022; **16**: 1007453.
- 12 Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. *Lect Notes Comput Sci* 2022; : 272–84.
- 13 Chen H, Qi Z, Shi Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans Geosci Remote Sens* 2022; **60**: 1–14.
- 14 Bandara WGC, Patel VM. A Transformer-Based Siamese Network for Change Detection. *IGARSS 2022 - 2022 IEEE Int Geosci Remote Sens Symp* 2022; **00**: 207–10.
- 15 Bandara WGC, Patel VM. Revisiting Consistency Regularization for Semi-supervised Change Detection in Remote Sensing Images. *arXiv* 2022. DOI:10.48550/arxiv.2204.08454.



- 16 [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. <https://arxiv.org/abs/2010.11929> (accessed Nov 6, 2023).
- 17 Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: An experimental study. *Méd Image Anal* 2023; **89**: 102918.
- 18 Liu Z, Hu H, Lin Y, *et al.* Swin Transformer V2: Scaling Up Capacity and Resolution. *2022 IEEE CVF Conf Comput Vis Pattern Recognit (CVPR)* 2022; **00**: 11999–2009.
- 19 Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Méd Image Anal* 2023; **85**: 102762.
- 20 Shamshad F, Khan S, Zamir SW, *et al.* Transformers in medical imaging: A survey. *Méd Image Anal* 2023; **88**: 102802.
- 21 Kelly B, Martinez M, Do H, *et al.* DEEP MOVEMENT: Deep learning of movie files for management of endovascular thrombectomy. *Eur Radiol* 2023. DOI:10.1007/s00330-023-09478-3.
- 22 Cho K, Kim J, Kim KD, *et al.* MuSiC-ViT: A multi-task Siamese convolutional vision transformer for differentiating change from no-change in follow-up chest radiographs. *Méd Image Anal* 2023; **89**: 102894.
- 23 Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021; **18**: 203–11.
- 24 Vrudhula A, Hughes JW, Yuan N, Ouyang D. The Impact of Task Set-up in Algorithm Design: Regression versus Classification. *NEJM AI* 2024. DOI:10.1056/aics2300176.
- 25 Rovira À, Wattjes MP, Tintoré M, *et al.* MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. *Nat Rev Neurol* 2015; **11**: 471–82.
- 26 Maier-Hein L, Reinke A, Christodoulou E, *et al.* Metrics reloaded: Pitfalls and recommendations for image analysis validation. *Arxiv* 2022.
- 27 Moskowitz CS, Welch ML, Jacobs MA, Kurland BF, Simpson AL. Radiomic Analysis: Study Design, Statistical Analysis, and Other Bias Mitigation Strategies. *Radiology* 2022; **304**: 265–73.
- 28 Kocak B, Baessler B, Bakas S, *et al.* CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMI. *Insights Imaging* 2023; **14**: 75.
- 29 Kelly B, Kirwan A, Quinn M, *et al.* The ethical matrix as a method for Involving People living with disease and the wider Public (PPI) in near-term artificial intelligence research. *Radiography* 2023.
- 30 Mongan J, Moy L, Jr CEK. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology Artif Intell* 2020; **2**: e200029.
- 31 Pham N, Hill V, Rauschecker A, *et al.* Critical Appraisal of Artificial Intelligence–Enabled Imaging Tools Using the Levels of Evidence System. *Am J Neuroradiol* 2023; **44**: E21–8.
- 32 Lesjak Ž, Galimzianova A, Koren A, *et al.* A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics* 2018; **16**: 51–63.
- 33 Kamnitsas K, Ferrante E, Parisot S, *et al.* Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016,

Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers. *Lect Notes Comput Sci* 2017; : 138–49.

34 Yushkevich PA, Gao Y, Gerig G, *et al.* ITK-SNAP: An Interactive Tool for Semi-Automatic Segmentation of Multi-Modality Biomedical Images. *2016 38th Annu Int Conf IEEE Eng Medicine Biology Soc EMBC 2016*; **2016**: 3342–5.

35 Daudt RC, Saux BL, Boulch A. Fully Convolutional Siamese Networks for Change Detection. *2018 25th IEEE Int Conf Image Process (ICIP) 2018*; **00**: 4063–7.

36 Ronneberger O, Fischer P, Brox T. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III. *Lect Notes Comput Sc* 2015; : 234–41.

37 Ferrer L. Analysis and Comparison of Classification Metrics. *arXiv* 2022. DOI:10.48550/arxiv.2209.05355.

38 Sica GT. Bias in Research Studies<sup>1</sup>. *Radiology* 2006; **238**: 780–9.