Project Notes:

Project Title: Using Novel Phrase-Level Explanations in a Softmax-Linked Additive Explainability Model for Transformers
Name: Gupta, Neil

Knowledge Gaps:

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
Methods for reducing background bias in image-based datasets using LRP	Consulted research paper on LRP optimization in imaging	Article #5 Notes, Sections: Introduction & Discussion	10/06/2025
Evaluation metrics for interpretability techniques like LRP	Checked examples and metrics in ML interpretability papers	Article #7 Notes, Section: Results	10/08/2025
Challenges in integrating ML-based analysis with experimental microscopic data	Compared ML studies with other existing ones	Article #10 Notes, Sections: Discussion	10/09/2025

Literature Search Parameters:

These searches were performed between 8/23/25 and 12/15/2025.
List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
WPI Library	“BCI” and “Paralysis”	Lots of articles, selected one, but moved on to different topics.
WPI Library	“Biomineralization” and “Concrete”	High potential project with lots of articles. Can pursue this application.
WPI Library	“Biomineralization” and “Machine Learning”	This provided some ideas related to using ML to analyze polymer chains. However, I want to focus on AI rather than bio.
WPI Library	“Layer-wise Relevance Propagation” and “Explainable AI”	Great topic that has lots of potential. Focuses on explaining DNNS.
WPI Library	“Layer-wise Relevance Propagation” and (“Deep Neural Networks” OR “Explainable AI”)	Gave similar results to previous search. This was unproductive.
WPI Library	“Layer-wise Relevance Propagation” and “CNN”	This was a great search that focused the LRP onto CNNS. Helped me focus my project down.
WPI Library		

Tags:

BCI

Article 1	

Neural	
Article 1	

Biomineralization	
Article 2	Article 3
Article 4	

Machine Learning	
Article 2	Article 8
Article 9	

DNA	
Article 3	

Concrete	
Article 4	

DNN	
Article 5	Article 6
Article 10	

Background Bias	
Article 5	

Shortcut Learning	
Article 6	

Explainable AI (XAI)	
Article 6	Article 7
Article 9	

Bisinformation	
Article 8	

LRP	
Article 7	Article 9

Microscopy	
Article 10	

Article Template

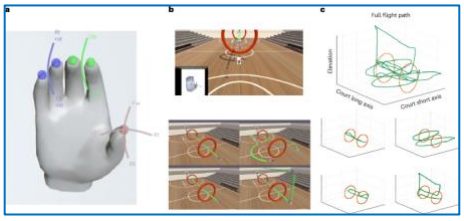
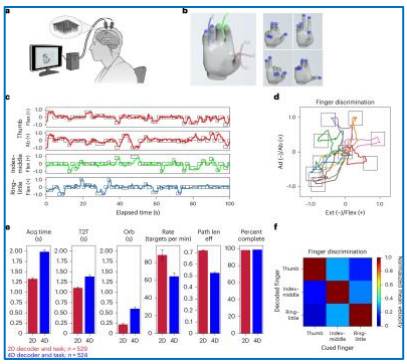
Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	
Source citation (APA Format)	
Original URL	
Source type	
Keywords	
#Tags	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

Article #1 Notes: A high-performance brain-computer interface for finger decoding and quadcopter game control in an individual with paralysis

Source Title	A high-performance brain-computer interface for finger decoding and quadcopter game control in an individual with paralysis
Source citation (APA Format)	Willsey, M. S., Shah, N. P., Avansino, D. T., & et al. (2025). A high-performance brain–computer interface for finger decoding and quadcopter game control in an individual with paralysis. *Nature Medicine*, *31*(1), 96–104. https://doi.org/10.1038/s41591-024-03341-8
Original URL	https://www.nature.com/articles/s41591-024-03341-8#Bib1
Source type	Journal Article
Keywords	BCI, paralysis, neural, motor impairments, finger tasks
#Tags	#BCI, #Neural
Summary of key points + notes (include methodology)	**Introduction:** • People with paralysis often have trouble moving. • Older BCIs could only read movements for two finger groups. This limited hand use and independence. **Methodology:** • Researchers built a BCI that recorded detailed brain signals from the motor cortex. • They added a third finger group to the system, doubling control options. • A participant with paralysis learned to control it using only thoughts. • They measured how accurate, fast, and stable the control was over several sessions. **Results:** • The participant successfully flew a virtual quadcopter through an obstacle course. • The BCI worked reliably across all trials. • This performance was better than older finger-decoding models used in humans and primates. **Implications / Conclusion:**

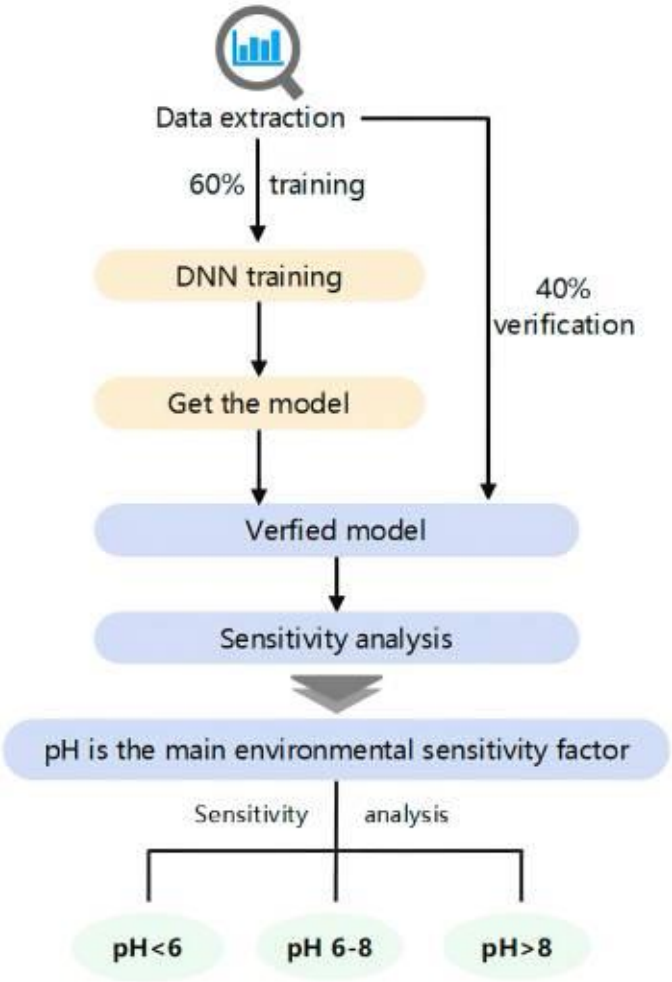
	• This study shows a way to help paralyzed people participate more in daily life. • It provides a foundation for future assistive devices, including BCIs in virtual settings and prosthetics.
Research Question/Problem/ Need	Can paralyzed people utilize a brain-computer interface to control their fingers and do activities that connect them socially?
Important Figures	 Figure 1. Finger iBCI translated to virtual quadcopter control.  Figure 2. iBCI system for dexterous finger movements. **<u>As a baseline for future figures, if the figure caption is not present in the image, then I will add detailed captions in text. If captions were included in the image themselves, no additional text captions will be added as the figure caption by the author is self-explanatory.</u>**
VOCAB: (w/definition)	BCI: a system that measures brain activity and uses it to control external devices or interact with the environment, thus replacing or supplementing the brain's natural outputs Degrees of freedom: the number of independent components that must be coordinated and controlled by the nervous system to produce a movement Intracortical BCI: a type of BCI that involves surgically implanting microelectrode arrays directly into the brain's cortex
Cited references to follow up on	• Armour, B. S., Courtney-Long, E. A., Fox, M. H., Fredine, H., & Cahill, A. (2016). Prevalence and causes of paralysis—United States, 2013. *American*

	Journal of Public Health, *106*(10), 1855–1857. https://doi.org/10.2105/AJPH.2016.303316 • Trezzini, B., Brach, M., Post, M., & Gemperli, A. (2019). Prevalence of and factors associated with expressed and unmet service needs reported by persons with spinal cord injury living in the community. *Spinal Cord*, *57*(6), 490–500. • Cairns, P., Power, C., Barlet, M., Haynes, G., & Guckelsberger, C. (2021). Enabled players: The value of accessible digital games. *Games and Culture*, *16*(3), 262–282. https://doi.org/10.1177/1555412019891291
Follow up Questions	1. How well would this brain-computer interface perform in a larger and more diverse population of individuals with varying levels? 2. How does the patient's performance change over longer periods of use? 3. Does this interface include any long-term health effects that are potentially negative?

Article #2 Notes: Prediction of the Non-Reducing Biomineralization of Nuclide–Microbial Interactions by Machine Learning: The Case of Uranium and Bacillus subtilis

Source Title	Prediction of the Non-Reducing Biomineralization of Nuclide–Microbial Interactions by Machine Learning: The Case of Uranium and Bacillus subtilis
Source citation (APA Format)	Qiang, S., Liu, L., Li, S., Wang, S., Huang, X., Yang, J., Song, J., Zhang, Y., Huang, Y., & Fan, Q. (2025). Prediction of the Non-Reducing Biomineralization of Nuclide-Microbial Interactions by Machine Learning: The Case of Uranium and Bacillus subtilis. *Toxics*, *13*(4), 305. https://doi.org/10.3390/toxics13040305
Original URL	https://pmc.ncbi.nlm.nih.gov/articles/PMC12030973/
Source type	Journal Article
Keywords	Biomineralization, Uranium, Bacillus subtilis, Machine Learning, Nuclide Immobilization
#Tags	#Biomineralization #Machine Learning
Summary of key points + notes (include methodology)	**Introduction:** • Uranium pollution is a big environmental problem because it is toxic and moves easily. • Microbial non-reducing biomineralization is a safer way to trap uranium. **General Research Need:** • Predicting biomineralization is hard because of complex environmental and microbial factors. • Computer tools are needed to improve cleanup strategies. **Methodology:** • They collected data on uranium and Bacillus subtilis, including pH, temperature, and ion levels. • Measured microbial traits related to mineral formation. • Trained machine learning models to predict uranium trapping efficiency. • Checked which environmental and microbial factors were most important. • Tested the models using cross-validation and accuracy to make sure they

	worked well. **Results:** • Machine learning models successfully predicted uranium trapping under different conditions. • Important factors included pH, temperature, ion levels, and microbial surface traits. **Implications / Conclusion:** • Machine learning can help plan bioremediation by finding the best conditions. • Non-reducing biomineralization is a strong and eco-friendly way to trap uranium.
Research Question/Problem/ Need	Can machine learning models predict the efficiency of non-reducing uranium biomineralization by Bacillus subtilis, enabling better environmental remediation strategies?

Important Figures	 Figure 1. The experimental flow chart.

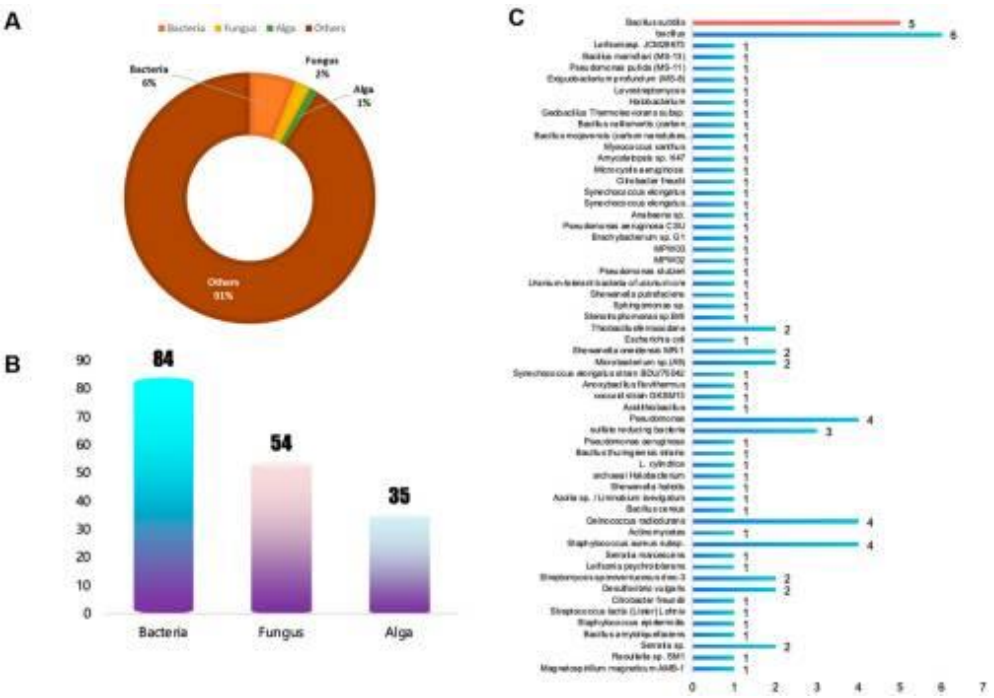


Figure 2. The characteristics of articles related to uranium adsorption: (A) percentage, (B) specific number, (C) the species of bacteria, red bar represents the Bacillus subtilis.

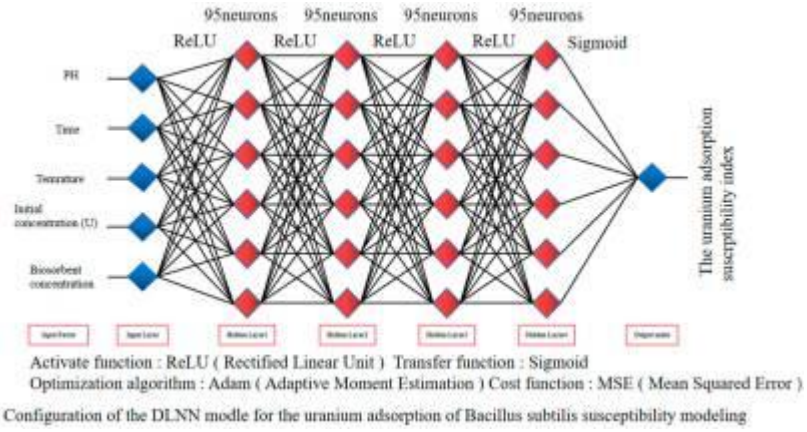


Figure 3. Configuration of the DLNN model for the uranium adsorption of Bacillus subtilis susceptibility modeling.

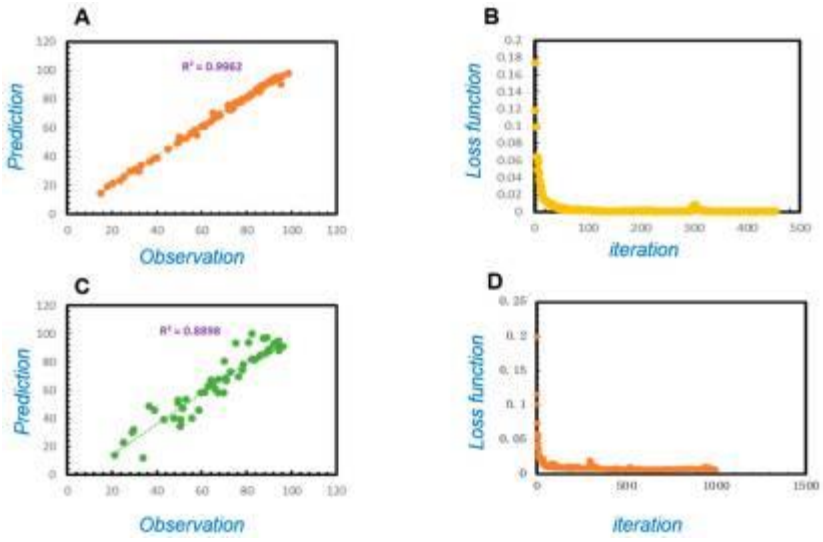


Figure 4. The tests of the model prediction performance: (A) the R2 of the training set, (B) loss function of the training set, (C) the R2 of the validating set, and (D) loss function of the validating set.

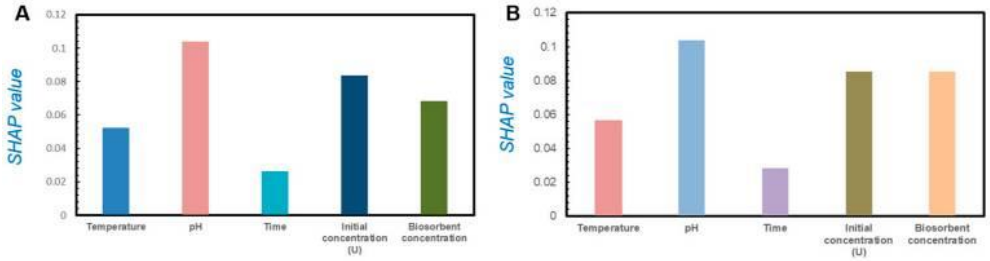


Figure 5. The contribution of variables: (A) the bar of absolute SHAP values of the training set, (B) the bar of absolute SHAP values of the validating set.

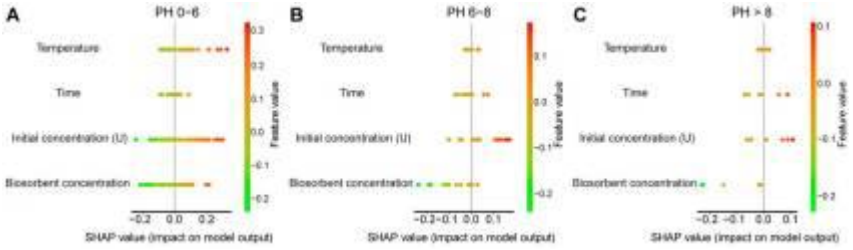


Figure 6. The specific potential increased/decreased effects of the data on the activities at (A) pH values of 0 to 6 (training), (B) pH values of 6 to 8, and (C) pH greater than 8 (training).

VOCAB: (w/definition)	Biomineralization: Biological process by which living organisms produce minerals. Non-reducing biomineralization: Mineral formation that does not involve electron transfer reactions. This relies on passive precipitation. Bacillus subtilis: A common, non-pathogenic bacterium capable of

	biomineralization. Nuclide: A type of atom defined by its number of protons and neutrons. Feature Importance: A measure of which input variables, also known as the features, most strongly influence the predictions of an ML model.
Cited references to follow up on	• Kolhe, N., Zinjarde, S., & Acharya, C. (2018). Responses exhibited by various microbial groups relevant to uranium exposure. *Biotechnology Advances*, *36*(7), 1828–1846. https://doi.org/10.1016/j.biotechadv.2018.07.002 • Wang, Z., Sun, H., Chen, Y., Song, J., He, M., Li, Q., Deng, Q., & Yang, H. (2024). Uranium resources of Europe: Development status, metallogenic provinces, and geodynamic setting. *Energy Strategy Reviews*, *54*, 101467. https://doi.org/10.1016/j.esr.2024.101467 • Banala, U. K., Das, N. P. I., Padhi, R. K., & Toleti, S. R. (2021). Alkaliphilic bacteria retrieved from uranium mining effluent: Characterization, U sequestration, and remediation potential. *Environmental Technology & Innovation*, *24*, 101893. https://doi.org/10.1016/j.eti.2021.101893
Follow up Questions	How might interactions with other microbial species influence the predictive accuracy of ML models for uranium biomineralization? Could the model be adapted to predict mineralization under fluctuating environmental conditions? Which specific microbial surface proteins are most predictive of uranium immobilization, and can these be engineered for improved remediation?

Article #3 Notes: 3D-Printed Functional Hydrogel by DNA-Induced Biomineralization for Accelerated Diabetic Wound Healing

Source Title	3D-Printed Functional Hydrogel by DNA-Induced Biomineralization for Accelerated Diabetic Wound Healing
Source citation (APA Format)	Kim, N., Lee, H., Han, G., Kang, M., Park, S., Kim, D. E., Lee, M., Kim, M. J., Na, Y., Oh, S., Bang, S. J., Jang, T. S., Kim, H. E., Park, J., Shin, S. R., & Jung, H. D. (2023). 3D-printed functional hydrogel by DNA-induced biomineralization for accelerated diabetic wound healing. *Advanced Science*, *10*(17), e2300816. https://doi.org/10.1002/advs.202300816
Original URL	https://pmc.ncbi.nlm.nih.gov/articles/PMC10265106/
Source type	Journal Article
Keywords	Biomineralization, DNA-induced mineralization, hydrogel, 3D printing, wound healing, diabetic wounds
#Tags	#Biomineralization #DNA
Summary of key points + notes (include methodology)	**Introduction:** • Chronic diabetic wounds are hard to heal because tissue repair and blood vessel growth are poor. • Biomineralized hydrogels give both support and chemical signals to help healing. **Generic Research Need:** • Regular hydrogels do not control mineral growth well and often have weak bioactivity. • Using DNA to guide biomineralization can improve mineral uniformity and scaffold performance, speeding up healing. **Methodology:** • Designed a hydrogel scaffold using DNA to guide calcium phosphate mineralization. • Made 3D-printed structures with adjustable shape and porosity for better tissue integration. • Tested in the lab for cell growth, compatibility, and mineral deposition. • Tested on diabetic wound models to check wound closure and tissue

	growth. • Analyzed scaffold composition and mineral placement using microscopy and spectroscopy. **Results:** • DNA guided consistent and aligned calcium phosphate deposition in the hydrogel. • 3D-printed biomineralized scaffolds improved cell attachment and tissue repair. • Wounds healed faster and regenerated better than with normal or non-mineralized hydrogels. **Implications / Conclusion:** • DNA-induced biomineralization makes a controllable, bioactive platform for healing. • Combining biomineralization with 3D printing improves precision and function. • This method could help not only diabetic wounds but also other tissue repairs like bone or cartilage.
Research Question/Problem/ Need	Can DNA-induced biomineralization in 3D-printed hydrogel scaffolds improve tissue regeneration and accelerate healing of chronic diabetic wounds compared to conventional or non-biomineralized hydrogels?

Important Figures

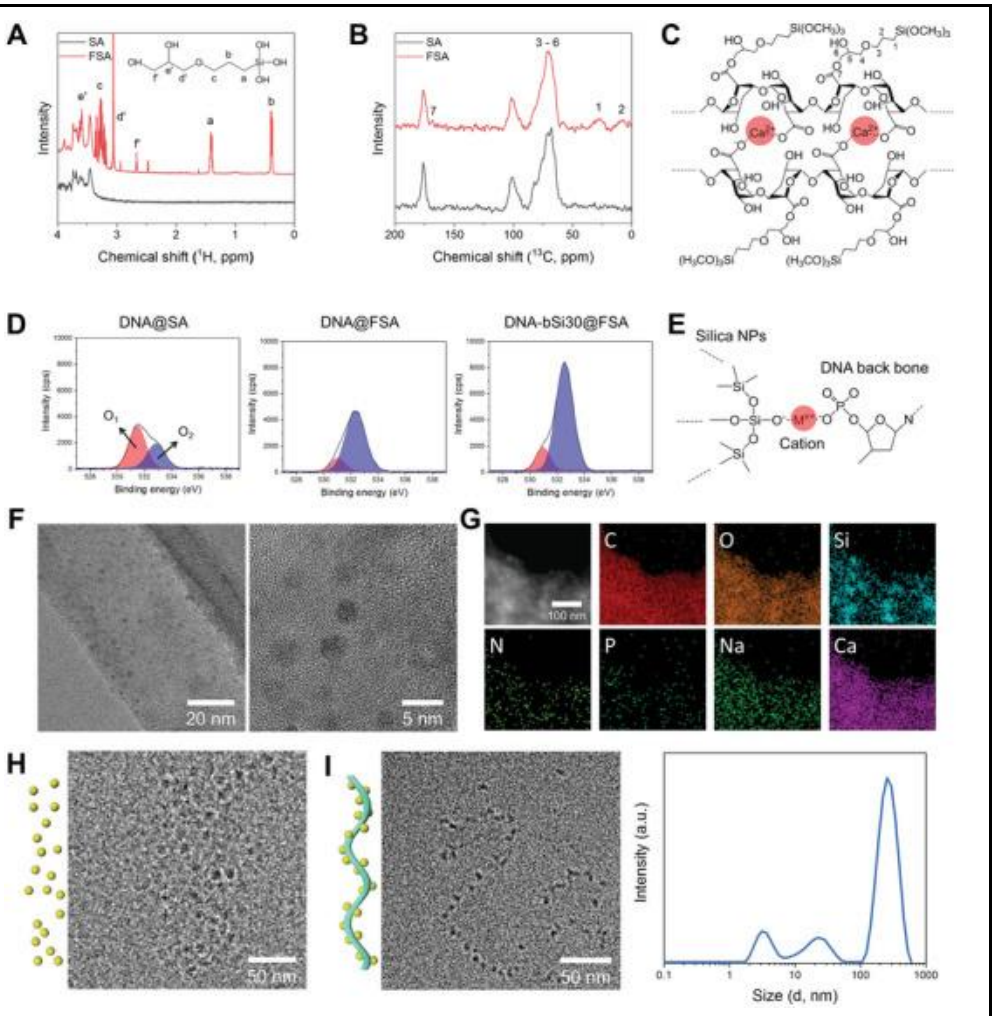


Figure 1. Results of A) liquid-state 1H NMR and B) solid-state 13C NMR of the pristine alginate and prepared functionalized alginate (FSA) hydrogel. C) Chemical structure of the prepared hydrogel inks comprising ionically crosslinked FSA based on NMR results. D) XPS results (O 1s) of DNA@SA, DNA@FSA, and DNA-bSi30@FSA, and E) schematic diagram of DNA-conjugated silica NPs. F) Representative TEM images of synthesized silica NPs in DNA@FSA ink at different magnifications. G) STEM and elemental mapping images of DNA@FSA ink. H) Silica NPs present in Si@FSA ink observed with cryo-TEM. I) Biomineralized silica NPs on DNA with a ribbon-like structure present in DNA-bSi@FSA inks and their size distribution.

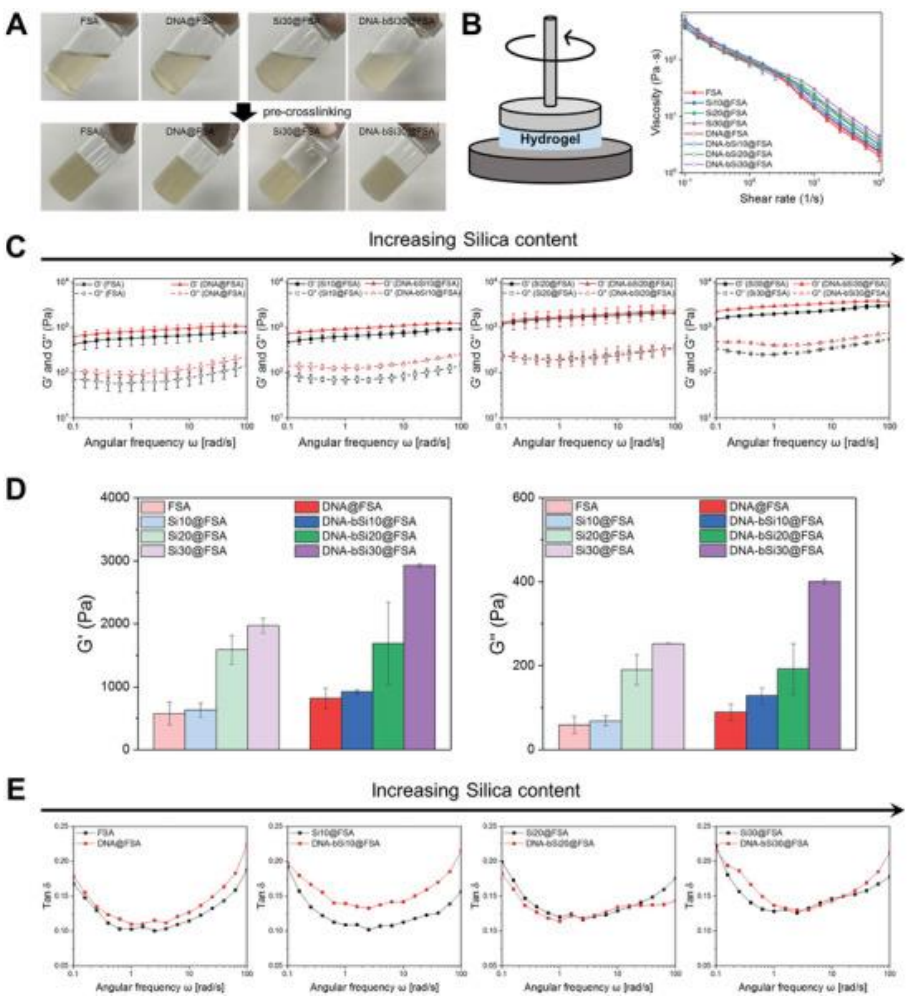


Figure 2. A) Optical images of the prepared bioactive hydrogel inks with flowability and pre-crosslinked inks exhibiting solid-like behavior. B) Schematic diagram of measured viscosity of bioactive hydrogel inks and monitored viscosity in terms of shear rate. C) Storage modulus (G′) and loss modulus (G″) of bioactive hydrogel inks with different silica concentrations based on DNA presence. D) Quantified G′ and G″ at an angular frequency of 1 Hz. E) Calculated tan δ of bioactive hydrogel inks.

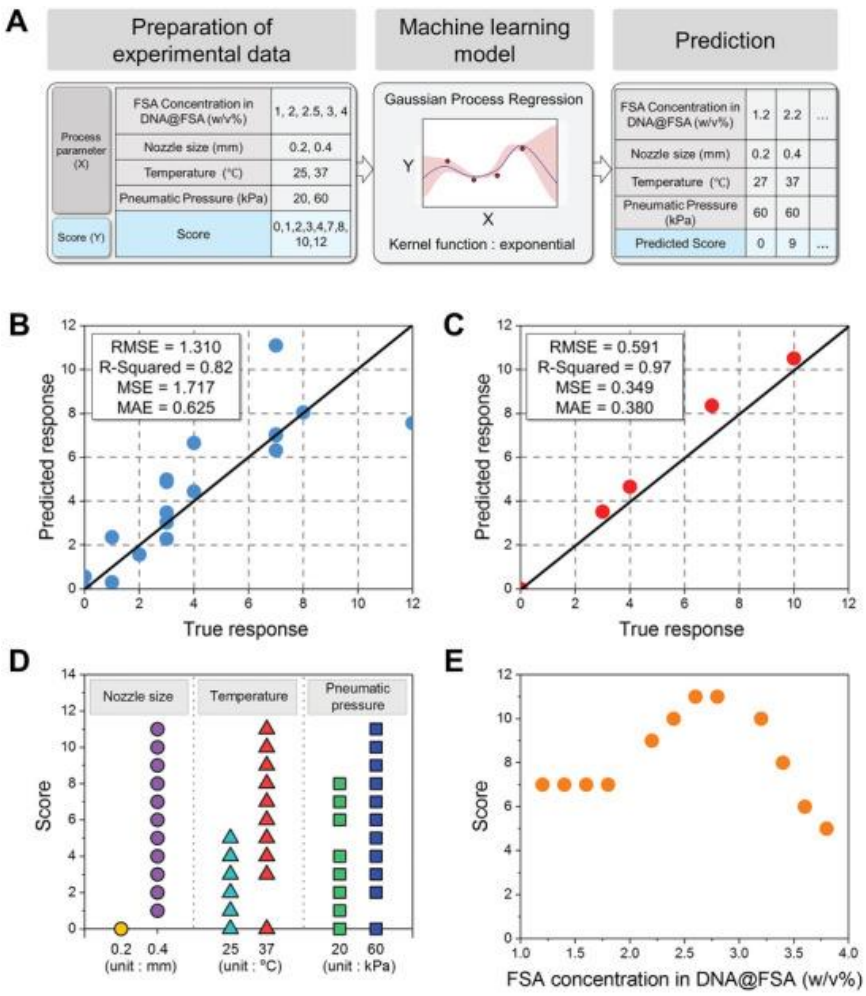


Figure 3. A) Flow chart of machine learning modeling using Gaussian process regression. B) Trained model validation (FSA concentration in DNA@FSA: 1, 2, 3, and 4 w/v%). C) Trained model test (FSA concentration in DNA@FSA: 2.5 w/v%). Score according to variable D) nozzle size, temperature, pneumatic pressure, and E) FSA concentration.

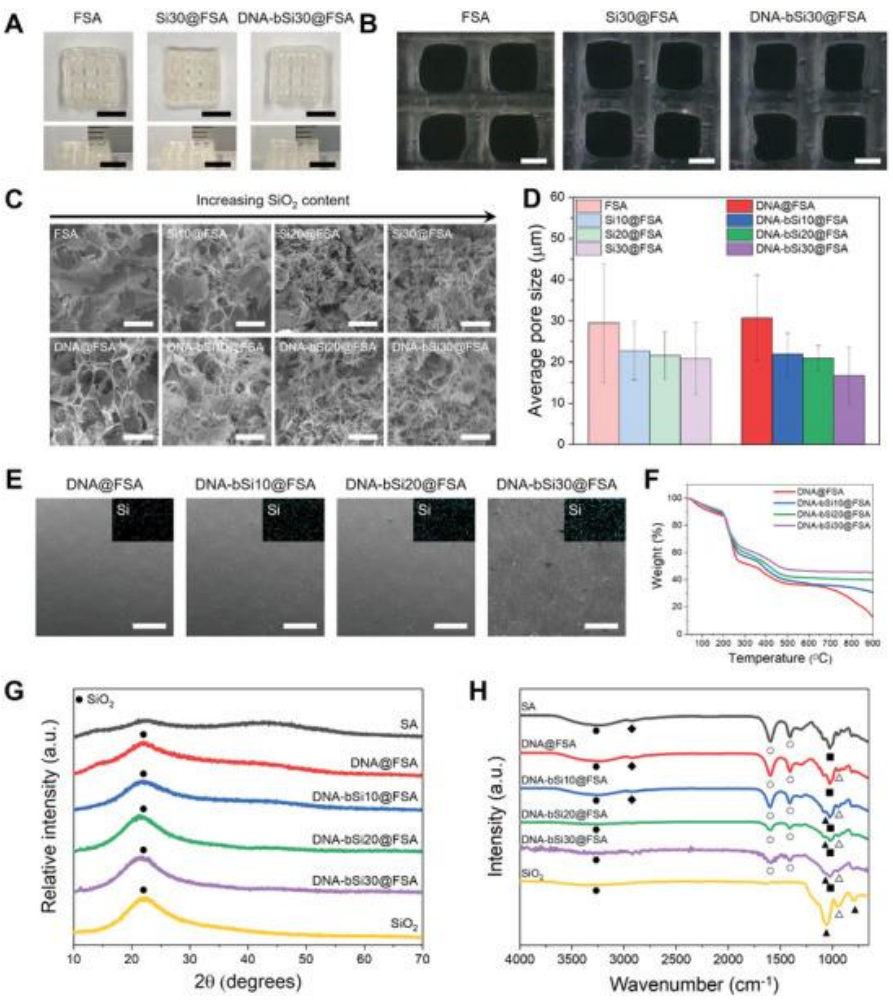


Figure 4. A) Optical images of 3D-printed hydrogel dressings captured from top and side perspectives (scale bar: 5 mm). B) Demonstration of the hydrogel dressings with a lattice structure using optical microscopy (scale bar: 1 mm). C) 3D microstructure of freeze-dried hydrogel dressings composed of micropores (scale bar: 50 μm). D) Average pore size under each condition (n = 10). E) Typical FE-SEM images and Si distributions of air-dried DNA-incorporated hydrogel scaffolds (scale bar: 500 μm). F) Monitored weight decrease from TGA. G) XRD patterns and (H) FT-IR spectra of the alginate, silica, and DNA-Si@FSA nanocomposites (●: —OH group, ♦: C—H Stretching, ○: Vibrations of asymmetric elongation of the C—O bond of the COO group, ▲: Si—O—Si bond, ■: antisymmetric elongation of C—O—C, and △: Si—OH bond).

VOCAB: (w/definition)

Hydrogel: Water-rich polymer network commonly used as a scaffold in tissue engineering.

DNA-induced biomineralization: Using DNA as a template to guide mineral

	deposition. 3D printing: Layer-by-layer fabrication method that allows precise scaffold geometry. Diabetic wound: Chronic wound associated with diabetes that is characterized by slow healing. Scaffold porosity: The presence of pores that facilitate cell infiltration and nutrient transport. Cytocompatibility: The ability of a material to support cell survival and function. Bioactive cues: Signals from the material that promote cellular responses like growth or differentiation.
Cited references to follow up on	• Jahromi, M. A. M., Zangabad, P. S., Basri, S. M. M., Zangabad, K. S., Ghamarypour, A., Aref, A. R., Karimi, M., & Hamblin, M. R. (2018). Multifunctional hydrogels for biomedical applications: A review. *Advanced Drug Delivery Reviews*, *123*(5), 33–64. https://doi.org/10.1016/j.addr.2017.09.017 • Ilhan, E., Cesur, S., Guler, E., Topal, F., Albayrak, D., Guncu, M. M., Cam, M. E., Taskin, T., Sasmazel, H. T., & Aksu, B. (2020). Development of bioactive nanocomposite hydrogels for wound healing applications. I*nternational Journal of Biological Macromolecules*, *161*(13), 1040–1054. https://doi.org/10.1016/j.ijbiomac.2020.06.153 • Liu, J., Du, P., Liu, T., Wong, B. J. C., Wang, W., Ju, H., & Lei, J. (2019). Biomineralized hydrogels for tissue regeneration and wound healing applications. *Biomaterials*, *192*(2), 179–188. https://doi.org/10.1016/j.biomaterials.2018.11.017
Follow up Questions	How does DNA sequence or structure influence the orientation and morphology of mineral deposition? Can this approach be adapted for other mineral types or tissue applications (e.g., bone or cartilage)? How do scaffold design parameters affect the efficiency of wound healing? What are the long-term effects of implanted biomineralized hydrogels in vivo?

Article #4 Notes: Performance evaluation of self-healing recycled concrete using biomineralization modified recycled aggregate as bacterial carrier

Source Title	Performance evaluation of self-healing recycled concrete using biomineralization modified recycled aggregate as bacterial carrier
Source citation (APA Format)	Luo, M., Ji, A., Li, X., & Yang, D. (2024). Performance evaluation of self-healing recycled concrete using biomineralization modified recycled aggregate as bacterial carrier. *Journal of Building Engineering, 86,* 109000. https://doi.org/10.1016/j.jobe.2024.109000
Original URL	https://www-sciencedirect-com.ezpv7-web-p-u01.wpi.edu/science/article/pii/S2352710224005680
Source type	Journal Article
Keywords	Self-healing Concrete, Biomineralization, Recycled aggregate
#Tags	#Biomineralization #Concrete
Summary of key points + notes (include methodology)	**Introduction:** • Microbes can deposit calcium carbonate to enable concrete self-repair. • Aggregates can act as carriers to protect bacteria in harsh concrete environments. • Recycled aggregates with bacterial loading and biomineralization may enhance concrete performance. **Generic Research Need:** • Ordinary recycled aggregates reduce concrete performance due to defects.

	• Integrating aggregate improvement with bacterial self-healing could improve recycled concrete properties. **Methodology:** • Recycled aggregates were loaded with bacteria • They were treated via biomineralization to enhance density and reduce porosity. • Concrete mixes were prepared with treated and untreated aggregates, then tested for the following: workability, strength, permeability, durability, and self-healing performance. • Microstructure and bacterial activity were analyzed to assess healing efficiency **Results:** • Biomineralization enhanced aggregate properties and provided a better environment for bacteria. • Concrete with bacteria-loaded, mineralized aggregates showed improved crack self-healing and reduced water permeability. • Mechanical performance and durability were better than concrete with ordinary recycled aggregates. **Implications / Conclusion:** • Recycled aggregates can serve as effective bacterial carriers for self-healing concrete. • Biomineralization enhances aggregate properties and supports bacterial activity, improving concrete performance. • This approach supports sustainable construction.
Research Question/Problem/ Need	Can biomineralization-modified recycled aggregates effectively serve as bacterial carriers to enhance the self-healing efficiency of recycled concrete?
Important Figures	

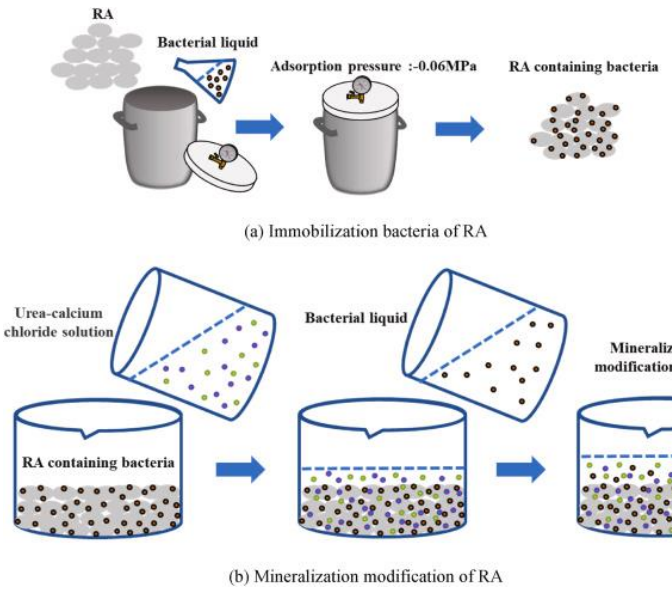


Figure 1. Immobilization bacteria and mineralization modification process of RA.

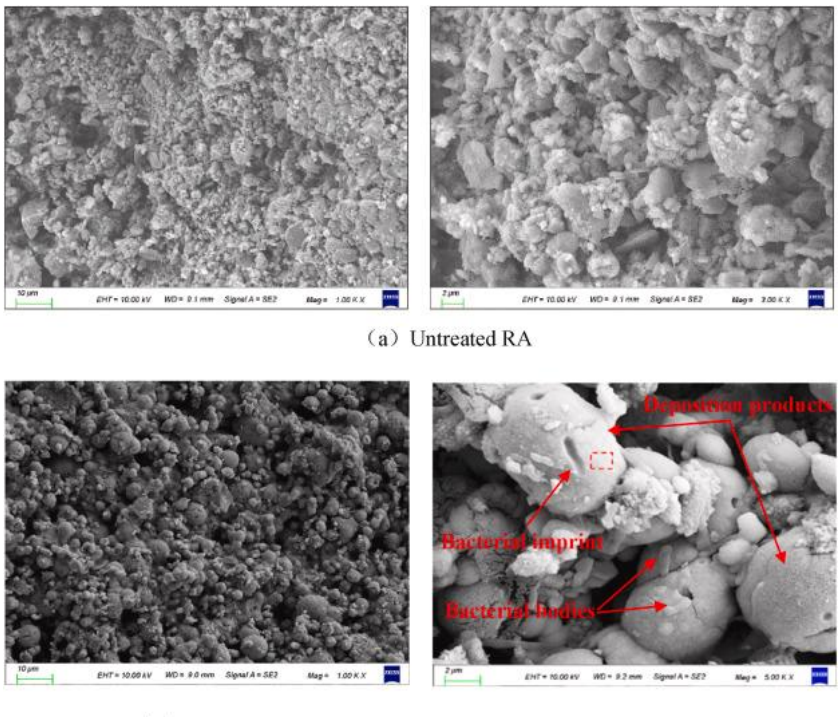


Figure 2. Surface micromorphology of RA before and after microbial mineralization treatment.

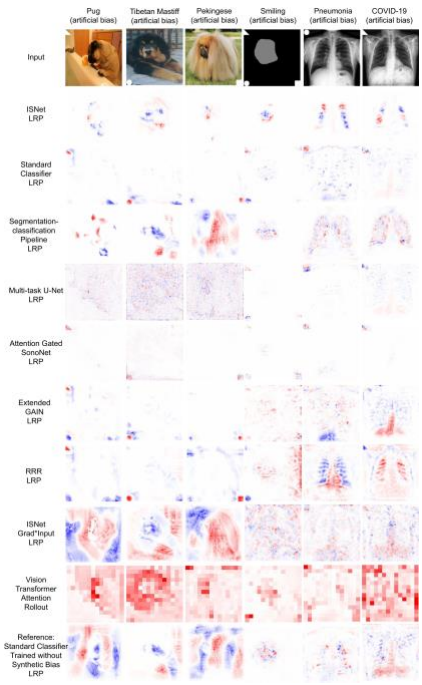
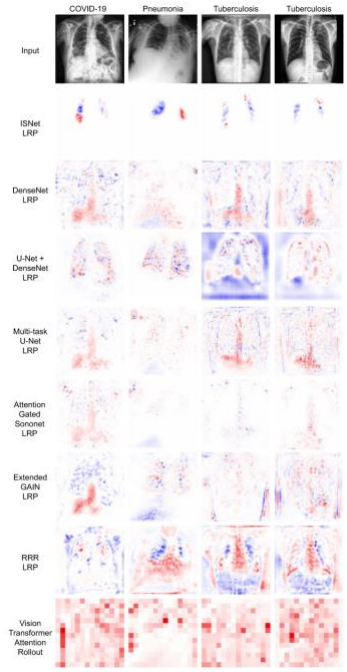
VOCAB: (w/definition)	Recycled aggregate (RA): Aggregates obtained by crushing and processing construction waste. Self-healing concrete: Concrete capable of autonomously repairing cracks via

	proposed biomineralization. Sporosarcina pasteurii: Bacteria used for microbial-induced CaCO3 deposition. Crack width repair ratio: Percentage reduction of crack width after healing. Relative permeability coefficient: Measure of water flow through cracked concrete (lower values indicate better sealing). Chloride ion penetration resistance: Measure of concrete's durability against corrosion. Vacuum impregnation: A technique to load liquids or bacteria into porous materials under reduced pressure.
Cited references to follow up on	• De Muynck, W., De Belie, N., & Verstraete, W. (2010). Microbial carbonate precipitation in construction materials: A review. *Ecological Engineering*, *36*, 118–136. https://doi.org/10.1016/j.ecoleng.2009.02.001 • Al-Salloum, Y., Hadi, S., Abbas, H., Almusallam, T., & Moslem, M. A. (2017). Bio-induction and bioremediation of cementitious composites using microbial mineral precipitation: A review. *Construction and Building Materials*, *154*, 857–876. https://doi.org/10.1016/j.conbuildmat.2017.08.064 • Wiktor, V., & Jonkers, H. M. (2011). Quantification of crack-healing in novel bacteria-based self-healing concrete. *Cement and Concrete Composites*, *33*(7), 763–770. https://doi.org/10.1016/j.cemconcomp.2011.03.002
Follow up Questions	How does the microstructure of biomineralization-modified aggregates affect activity during crack healing? How does the interaction between treated aggregates and the concrete matrix influence calcium carbonate deposition? Could adjusting the replacement ratio of treated aggregates optimize the efficiency of self-healing?

Article #5 Notes: Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization

Source Title	Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization
Source citation (APA Format)	Bassi, P. R. A. S., Dertkigil, S. S. J., & Cavalli, A. (2024). Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications*, *15*, 291. https://doi.org/10.1038/s41467-023-44371-z
Original URL	https://www.nature.com/articles/s41467-023-44371-z
Source type	Journal Article
Keywords	Deep neural networks, generalization, robustness, background bias, Layer-wise Relevance Propagation
#Tags	#DNN #BackgroundBias
Summary of key points + notes (include methodology)	**Introduction:** • Background features in images can create bias in image classification using deep neural networks. • This often causes shortcut learning. • The paper suggests a way to reduce background bias by improving LRP heatmaps during training: ISNET **Methods:** • Training images had either natural or synthetic background bias and were paired with masks showing the true object area.

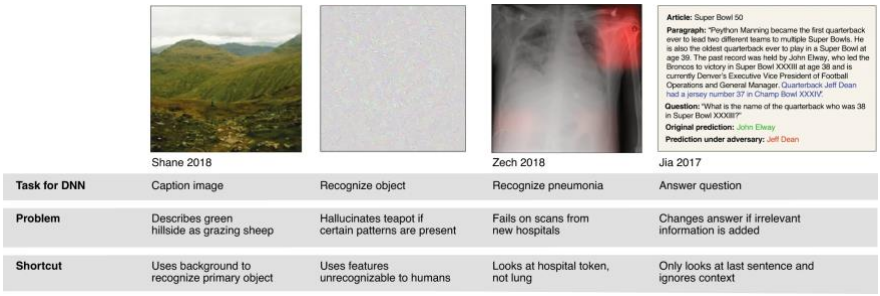
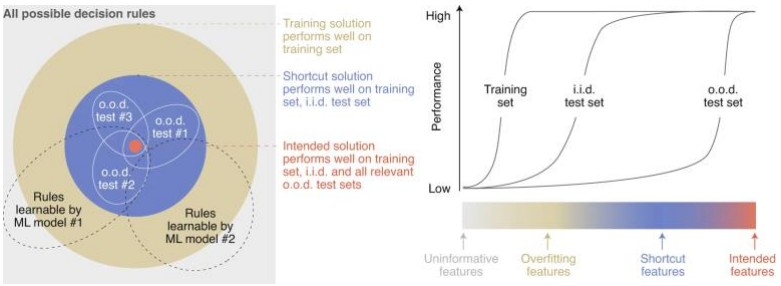
	• During the forward pass, LRP heatmaps showed how much each pixel affected the classification. • A joint loss combined normal classification loss and BRM Loss to punish attention to background areas. • After training, the LRP optimization was removed, and ISNet was tested on Out-of-Distribution (OOD) data. **Results:** • ISNet successfully ignored both synthetic and real background bias. • ISNet performed better than all benchmark models on OOD test datasets. • In COVID-19 detection, ISNet achieved the highest OOD F1 score: 0.773 plus or minus 0.009. **Implications / Conclusion:** • ISNet can be used with almost any classification architecture. • The main advantage is that it does not make the model slower after training. • This makes ISNet fast and efficient for real-world use.
Research Question/Problem/ Need	How can the generalization of deep neural networks in image classification be improved by mitigating the influence of background bias that lead to shortcut learning?

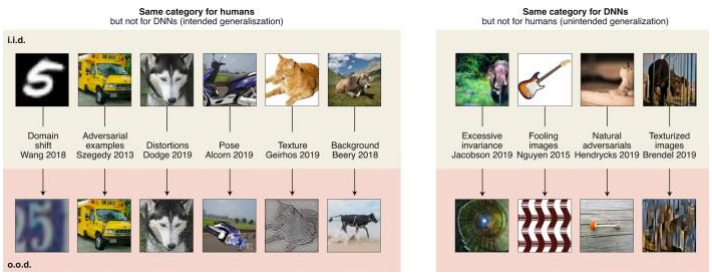
Important Figures	 Fig. 1: Heatmaps (Layer-wise Relevance Propagation/LRP for convolutional networks and attention rollout for Vision Transformer) for positive COVID-19 and Pneumonia X-rays and photographs, extracted from the synthetically biased test datasets (biased test).  Fig. 2: Heatmaps (Layer-wise Relevance Propagation/LRP for convolutional networks and attention rollout for Vision Transformer) for positive COVID-19, Pneumonia, and tuberculosis.
VOCAB: (w/definition)	• Layer-wise relevance propagation = heatmap that shows how much

	each part of the image affects the DNN's decision • Background Relevance Minimization (BRM) = loss function trained on heatmap made from LRP • Background Bias: Features in the background of an image that accidentally correlate with the image's class, causing a DNN to learn a misleading relationship. • Shortcut Learning: The phenomenon where a DNN achieves high performance by learning to rely on simple features that happen to be correlated with the class in the training data. • Out-of-Distribution (OOD) Generalization: A model's ability to perform accurately on new data that comes from a distribution different from the training data. • ISNet (Ignored-Segmentation Network): The proposed classifier architecture and training strategy that uses BRM to learn to ignore background features.
Cited references to follow up on	• Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z • Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140
Follow up Questions	How does the performance of the ISNet compare to other background bias mitigation techniques on a broader range of datasets beyond medical imaging and MNIST? What is the complexity or required detail of the initial foreground mask needed for the BRM training? What is the increase in training time or resource consumption due to the LRP calculation and BRM optimization compared to a standard training loop?

Article #6 Notes: Shortcut learning in deep neural networks

Source Title	Shortcut learning in deep neural networks
Source citation (APA Format)	Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z
Original URL	https://www.nature.com/articles/s42256-020-00257-z#citeas
Source type	Journal Article
Keywords	Shortcut Learning, Deep Neural Networks, Generalization, Dataset Bias, Out-Of-Distribution(OOD)
#Tags	#ShortcutLearning #DNN #XAI
Summary of key points + notes (include methodology)	**Introduction:** • DNNs often rely on superficial dataset correlations (“shortcuts”) rather than task-relevant features. • This can give high in-distribution performance but poor generalization to new data. **Generic Research Need:** • Evaluation metrics may overestimate true model understanding. • Studying shortcuts helps identify model weaknesses and improve robustness. **Methodology :** • They tested multiple DNN architectures on standard datasets with controlled modifications to highlight shortcut cues.

	• Models on OOD datasets were evaluated. They compared performance to human reasoning. • Applied interventions like data augmentation and feature disentanglement to reduce shortcut reliance. **Results:** • DNNs prioritize simple cues like texture over global shape. • Removing superficial correlations can drastically reduce performance. • Humans rely on robust features, highlighting DNN shortcut dependence. **Implications / Conclusion:** • Shortcut learning explains why DNNs can appear accurate yet lack true understanding. • Encourages improved architecture design, training strategies, and dataset construction.
Research Question/Problem/ Need	To what extent do DNNs rely on superficial dataset cues rather than robust features, and how does this affect OOD generalization?
Important Figures	 Figure 1. Examples of shortcut learning.  Figure 2. Taxonomy of decision rules.

	 Figure 3. Humans and DNNs both generalize, but they generalize very differently.
VOCAB: (w/definition)	• Shortcut learning: When a model relies on superficial dataset cues instead of robust task-relevant features. • Out-of-Distribution (OOD): Data that differs from the training distribution, used to evaluate generalization. • Dataset bias: Non-uniform or misleading correlations in the training data that a model may exploit. • Generalization: Ability of a model to perform well on unseen data outside the training distribution. • Feature disentanglement: Training method to encourage models to learn independent, meaningful features rather than shortcuts.
Cited references to follow up on	• He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034). *IEEE*. https://doi.org/10.1109/ICCV.2015.123 • Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1 • Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision* (pp. 456–473). *Springer*. https://doi.org/10.1007/978-3-030-01246-5_28
Follow up Questions	• How can training paradigms be systematically designed to minimize shortcut learning without reducing performance? • Which dataset features most commonly lead to shortcut reliance across different domains? • Can shortcut learning explain vulnerability to adversarial attacks or model brittleness?

	• How does shortcut learning vary across DNN architectures and depth?

Article #7 Notes: SLRP: Improved heatmap generation via selective layer-wise relevance propagation

Source Title	SLRP: Improved heatmap generation via selective layer-wise relevance propagation
Source citation (APA Format)	Jung, Y.-J., Han, S.-H., & Choi, H.-J. (2021). SLRP: Improved heatmap generation via selective layer-wise relevance propagation. *Electronics Letters*, *57*(10). doi:10.1049/ell2.12061
Original URL	https://ietresearch-onlinelibrary-wiley-com.ezpv7-web-p-u01.wpi.edu/doi/10.1049/ell2.12061
Source type	Journal Article
Keywords	Explainable AI, Heatmap, Layer-wise relevance propagation (LRP)
#Tags	#LRP #XAI
Summary of key points + notes (include methodology)	**Introduction:** • Deep learning models are very accurate but hard to understand because they are complex and non-linear. • Explainable AI tries to make model decisions easier to interpret without lowering performance. • Standard LRP methods often create heatmaps that are noisy and unclear. **Problem:** • Current methods like LRP, CLRP, and SGLRP still have noise and poor class separation. • A better method is needed to make heatmaps clearer while keeping object

	details and class focus. **Methodology:** • They created Selective Layer-Wise Relevance Propagation (SLRP), which mixes relevance and gradient ideas. • Only activations with positive gradients for the target class are passed through. • These activations are weighted by gradient strength, avoiding zero-sum limits. • They tested it on VGG16 using ImageNet 2012 validation data **Results:** • SLRP made heatmaps clearer, less noisy, and more focused on the right class than LRP. • It did better than CLRP and SGLRP in keeping full target objects and ignoring background noise. • Tests showed better localization of important image areas. **Implications / Conclusion:** • SLRP makes CNNs easier to understand by keeping only useful activations. • It gives a practical XAI method that improves visual explanations without reducing accuracy. • The method can be scaled and potentially used in other vision-based AI systems that need clear decision-making.
Research Question/Problem/ Need	How can heatmap-based explainability methods in deep learning be improved to produce clearer, class-discriminative, and object-preserving visualizations without increasing model complexity?

Important Figures

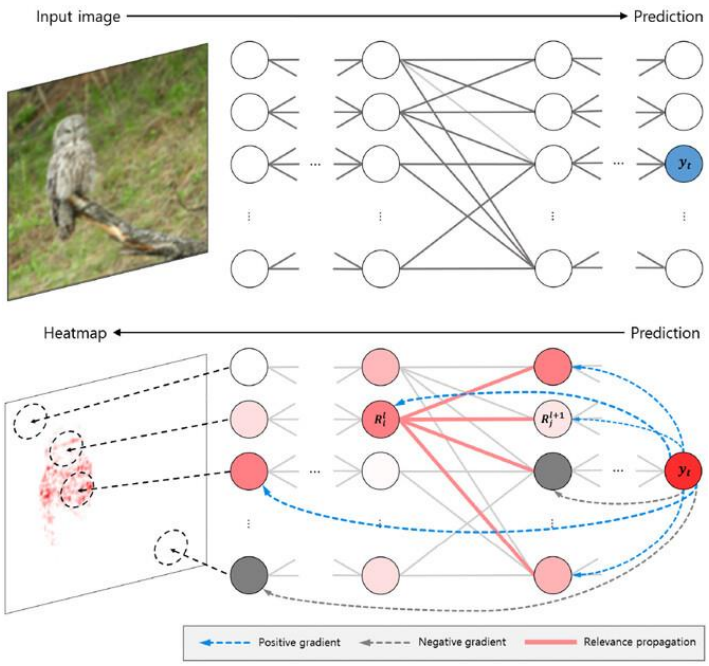


Figure 1. Overview of the proposed SLRP

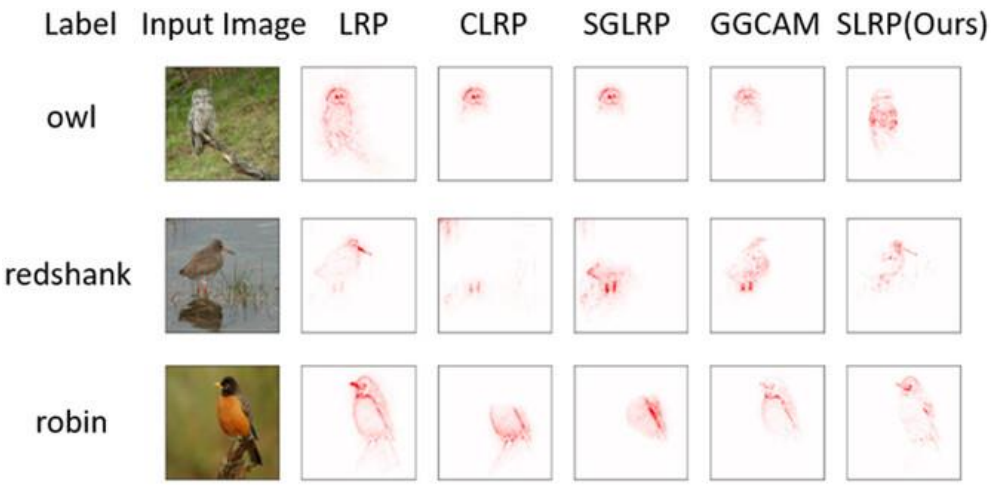


Figure 2. Comparison with other methods with poor CLRP and SGLRP results

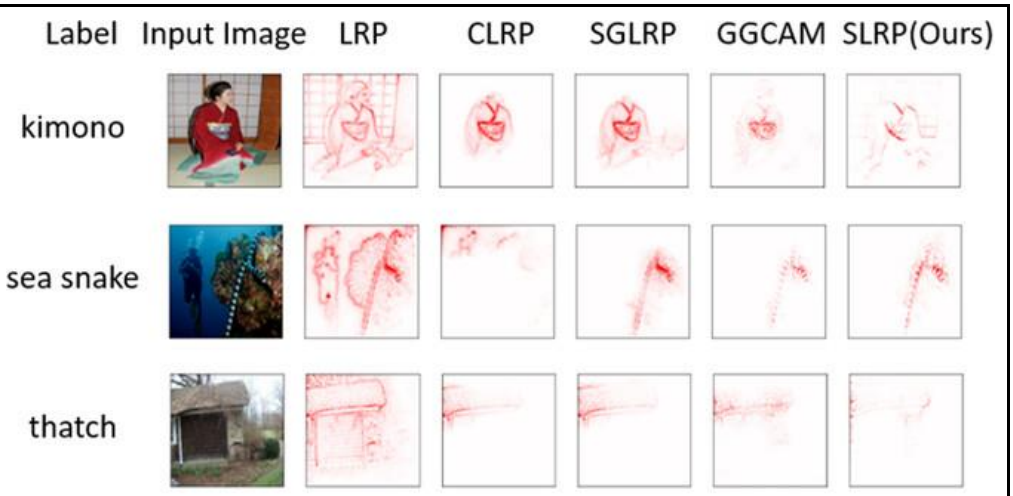


Figure 3. Comparison with other methods with poor LRP results

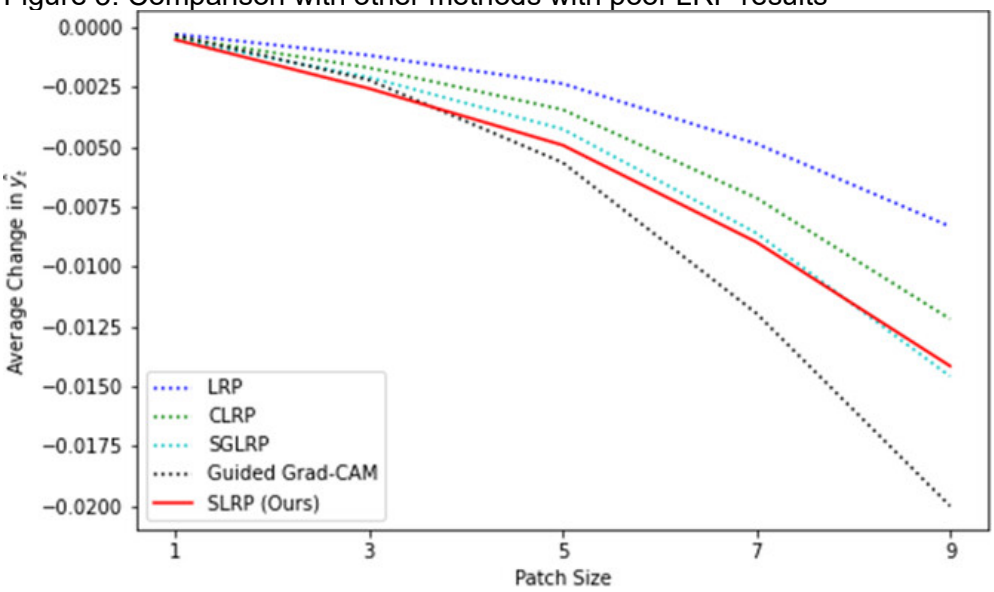


Figure 4. Average change in probability of the target class (y)

VOCAB: (w/definition)

Explainable AI (XAI): Field focused on making AI model decisions understandable to humans using reprensentative techniques.

Layer-Wise Relevance Propagation (LRP): Method to see how much each pixel contributes to a model's prediction.

Selective Layer-Wise Relevance Propagation (SLRP): A Modified LRP that only uses positively contributing activations for clearer heatmaps.

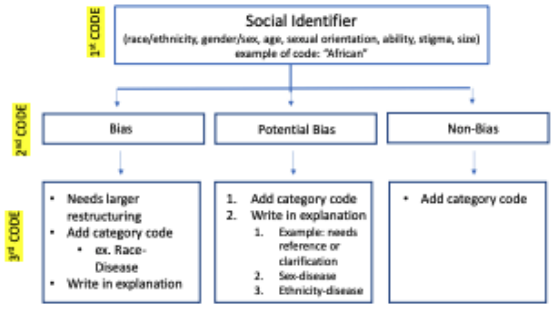
Global Average Pooling (GAP): A pooling operation that summarizes activation maps at the channel level.

Gradient: Measure of sensitivity showing how changes in input affect the model's output.

Cited references to follow up on	• Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140 • Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., & Sclaroff, S. (2017). Top-down neural attention by excitation backprop. *International Journal of Computer Vision, 126*(10), 1084–1102. https://doi.org/10.1007/s11263-017-1042-2
Follow up Questions	Could incorporating adaptive gradient thresholds further refine SLRP's ability to distinguish subtle class features? How would SLRP perform when applied to transformer-based or multimodal networks beyond CNNs? Could selective activation weighting be extended to temporal models, such as RNNs or video-based classifiers? COuld integrating SLRP with uncertainty estimation techniques improve model trustworthiness in critical systems?

Article #8 Notes: Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes

Source Title	Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes
Source citation (APA Format)	Salavati, C., Song, S., Diaz, W. S., Hale, S. A., Montenegro, R. E., Murai, F., & Dori-Hacohen, S. (2024). Reducing biases towards minoritized populations in medical curricular content via artificial intelligence for fairer health outcomes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 1269–1280. https://doi.org/10.1609/aies.v7i1.31722
Original URL	https://dl.acm.org/doi/10.5555/3716662.3716773
Source type	Journal Article
Keywords	Artificial Intelligence, Machine Learning, Bias Detection, Bisinformation, Fairness, Health Equity.
#Tags	#Bisinformation #MachineLearning
Summary of key points + notes (include methodology)	**Introduction:** • Biased information is often taught in medical training, which leads to unfair health outcomes. • Checking content for bias by hand is slow and takes a lot of work. **General Research Need:** • A faster and more scalable way is needed to find and mark biased text in medical materials. • This would help remove bias and make healthcare education fairer. **Methodology:** • They started the BRICC (Bias Reduction in Instructional Curricula Content) project using machine learning to detect bias. • Medical experts labeled text samples for bias related to gender, age, ethnicity, race, and other social factors. • They tested three types of machine learning models using a Transformer to predict overall and specific kinds of bias. • The system worked with human experts: the AI flagged possible bias,

	and people reviewed and fixed it. **Results:** • The General Binary Classifier was very effective at finding biased content. • It performed better than the other multi-task and specific-type models tested. **Implications / Conclusion:** • Machine learning can help find biased text quickly and accurately in medical training materials. • This method can make the debiasing process much faster and help schools improve fairness and health outcomes.
Research Question/Problem/ Need	Research Question: Is it possible to develop and evaluate Artificial Intelligence models capable of accurately and systematically detecting and flagging biased content in medical educational materials to enable an efficient expert-in-the-loop system for curricular debiasing.
Important Figures	 Figure 1: BRICC Coding/Labeling Procedure. If a social identifier is present in an excerpt, annotators coded the sample in 3 levels. 1st code: tokens denoting social identifier (e.g. “African American”); 2nd code: bias label (“bias”, “potential bias” or “non-bias”); and, if applicable, 3rd code: identity types (e.g., race, gender) that are relevant to the bias, along with an explanation for biased quotes.

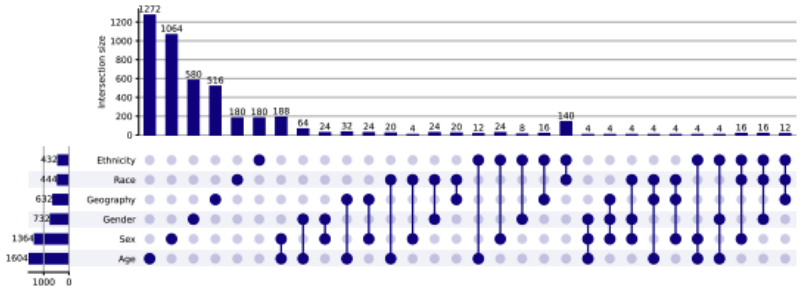
<table>
<tr>
<td></td>
<td>


Figure 2: An Upset plot detailing the intersection between the biased quotes. A filled in circle indicates the inclusion of a specific bias type in the intersection size.

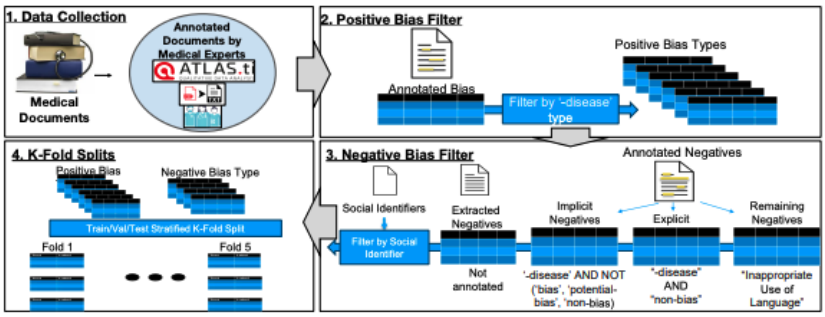


Figure 3: An overview of the proposed method from Data annotating by experts to the data splits. In part 1, we collected an annotated corpus from a team of medical experts and consolidated this to one file. In part 2, we filter the positive bias types by their respective type specific -'disease'. In part 3, we filter the negative bias subsets by different labeled conditions as well as by social identifiers for respective positive bias. In part 4, we split our data into training, validation, and testing sets by using a K-Fold split.
</td>
</tr>
<tr>
<td>VOCAB: (w/definition)</td>
<td>
Bisinformation: Biased medical information that continues to be taught in medical curricula.

BRICC: The acronym for the Bias Reduction in Instructional Curricula Content initiative.

Social Identifiers: The characteristics used to categorize groups in the study, for example race/ethnicity.

Expert-in-the-loop: A system design where the AI model flags potentially biased sentences. They are then passed to human domain experts for final review and correction.
</td>
</tr>
<tr>
<td>Cited references to follow up on</td>
<td>

Acquaviva, K. D., & Mintz, M. (2010). Perspective: Are we teaching racial profiling? The dangers of subjective determinations of race and ethnicity in case presentations. Academic Medicine, 85(4), 702–705.

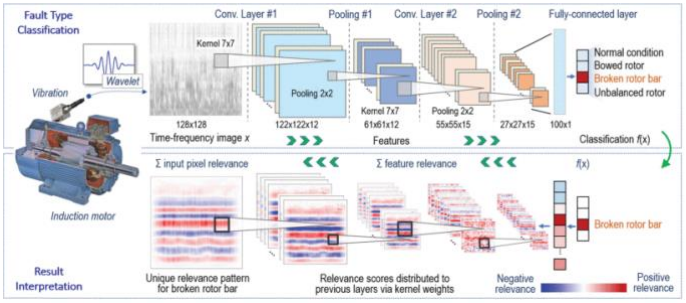
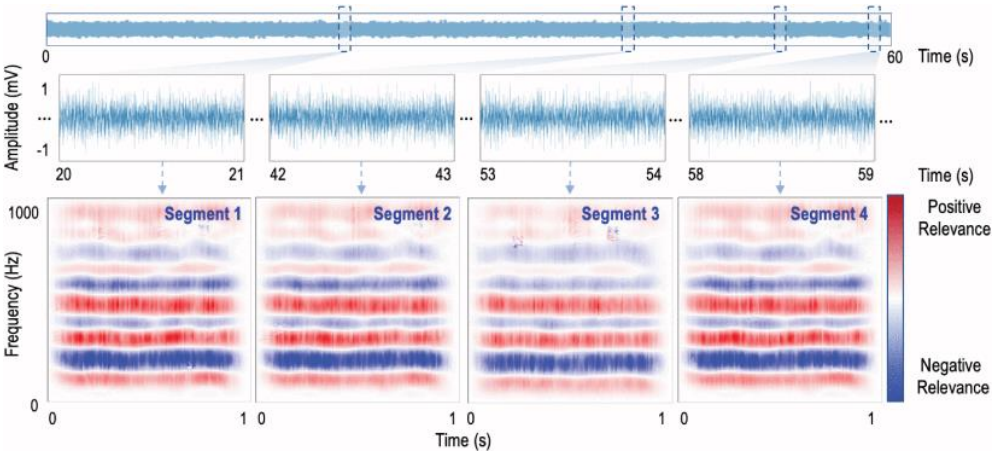
</td>
</tr>
</table>

	• Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., … Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. doi:10.1016/j.inffus.2023.03.008 • Ali-Khan, S. E., et al. (2011). The use of race, ethnicity, and ancestry in human genetic research. *The HUGO Journal*, *5*, 47–63.
Follow up Questions	How will the BRICC tool be made available (e.g., open-source tool, API, commercial software) for other medical institutions to adopt? What is the long-term plan for maintaining and updating the BRICC dataset to reflect evolving language and new forms of bias?

Article #9: Interpretable Convolutional Neural Network Through Layer-wise Relevance Propagation for Machine Fault Diagnosis

Source Title	Interpretable Convolutional Neural Network Through Layer-wise Relevance Propagation for Machine Fault Diagnosis
Source citation (APA Format)	Grezmak, J., Zhang, J., Wang, P., Loparo, K. A., & Gao, R. X. (2020). Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. *IEEE Sensors Journal, 20*(6), 3172–3181.

	https://doi.org/10.1109/JSEN.2019.2961517
Original URL	https://ieeexplore.ieee.org/document/8930493
Source type	Journal Article
Keywords	Convolutional Neural Network (CNN), Explainable AI (XAI), Layer-wise Relevance Propagation (LRP), Time-Frequency Analysis
#Tags	#XAI #MachineLearning #LRP #CNN
Summary of key points + notes (include methodology)	**Introduction:** • CNNs are often used to detect machine faults because they can learn complex patterns from data. • However, CNNs are like "black boxes," so it is hard to know why they make certain predictions. • This lack of clarity reduces trust and limits their use in important industries. **Research Need:** • CNN decisions need to be more understandable to ensure safety and reliability. • Knowing which input features affect predictions can help check and improve model accuracy. **Methodology:** • They trained a CNN on time-frequency spectra made from vibration signals of an induction motor. • LRP was used to create heatmaps showing which input areas affected each prediction the most. • They compared results from time-frequency images, raw time series, and DFT data to test interpretability and consistency. **Results:** • LRP heatmaps explained CNN predictions by showing patterns that matched known fault behaviors. • Time-frequency spectra made the model's reasoning clearer and more consistent than using raw data. • The approach showed which parts of the signal mattered, helping them understand how faults were classified. **Implications / Conclusion:**

	• Combining LRP with CNNs makes them more transparent and trustworthy. • This builds confidence in AI-based fault detection for industrial use. • Using time-frequency inputs with LRP creates a clearer way to monitor machine health.
Research Question/Problem/ Need	Research Question: How can Layer-wise Relevance Propagation be used to visually and quantitatively interpret the decision-making process of a CNN trained on time-frequency images of vibration signals for diagnosing machine fault?
Important Figures	 Figure 1. Flowchart of interpretation of CNN through LRP for motor fault diagnosis.  Figure 2. Relevance score heatmaps corresponding to different time-frequency image samples for motor with broken rotor bars.
VOCAB: (w/definition)	Convolutional Neural Network (CNN): A type of deep learning model highly effective for processing data with a grid-like topology Black Box Model: A machine learning model whose internal workings are not transparent or easily understood by humans. Time-Frequency Spectra: A visual representation of a signal that shows how the frequency content of the signal changes over time Machine Fault Diagnosis: The process of identifying the type and severity of

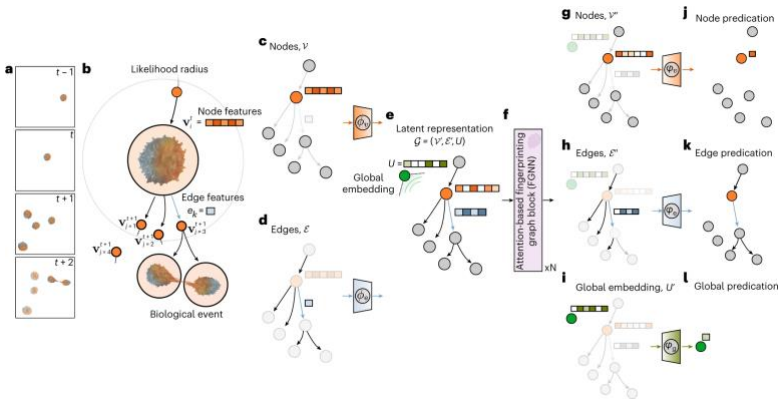
	mechanical failure using sensor data.
Cited references to follow up on	• Yin, S., Li, X., Gao, H., & Kaynak, O. (2015). Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics*, *62*(1), 657–667. https://doi.org/10.1109/TIE.2014.2318254 • Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing, 7*(3), 197–387. https://doi.org/10.1561/2000000039
Follow up Questions	How did the LRP heatmaps specifically correlate with known physical fault characteristics? How does the computational cost of integrating LRP compare to the cost of the forward pass of the CNN during inference? Can LRP be applied to 1D CNNs using raw time series inputs?

Article #10: Geometric deep learning reveals the spatiotemporal features of microscopic motion

Source Title	Geometric deep learning reveals the spatiotemporal features of microscopic

	motion
Source citation (APA Format)	Pineda, J., Midtvedt, B., Bachimanchi, H., Noé, S., Midtvedt, D., Volpe, G., & Manzo, C. (2023). Geometric deep learning reveals the spatiotemporal features of microscopic motion. *Nature Machine Intelligence*, *5*, 71–82. https://doi.org/10.1038/s42256-022-00595-0
Original URL	https://www.nature.com/articles/s42256-022-00595-0
Source type	Journal Article
Keywords	Geometric deep learning, graph neural networks (GNN), trajectory linking, motion characterization, spatiotemporal modelling, attention mechanisms, microscopy
#Tags	#DNN #Microscopy

Summary of key points + notes (include methodology)	Introduction
	• Advanced microscopes let scientists watch how cells and molecules move in good detail.
	• Older tracking methods don't work well in busy or noisy images.
	• The researchers made MAGIK, a deep learning model that learns motion directly from video data without manual setup.
	
	Problem / Research Need
	• Many current tools need manual tuning for each experiment.
	• They often fail when the data is crowded or unclear.
	• A single, easy-to-use model is needed to handle tracking and motion prediction together.
	
	Methodology
	• MAGIK essentially turns motion data into a graph:
	  ○ Each detection is a node
	  ○ Connections between detections are edges
	• A graph neural network with gated self-attention helps the model learn both local and overall motion patterns.
	• It was trained to link moving objects and identify motion types.
	
	Results
	• MAGIK linked objects almost perfectly, even in difficult images.

	• It accurately predicted how objects moved and what kind of motion was happening. • Extra tests showed that using gated attention improved accuracy. Implications / Conclusion • MAGIK is a dependable and simple way to study motion in microscopic data. • It works better than older tracking tools in crowded or noisy scenes. • The same design could be used for other motion-tracking problems beyond microscopy.
Research Question/Problem/ Need	Can a geometric deep-learning model operate directly on spatiotemporal detection graphs to jointly infer trajectories and motion parameters in complex biological imaging data without manual tuning?
Important Figures	 Figure 1. Spatiotemporal characterization of trajectories using MAGIK.

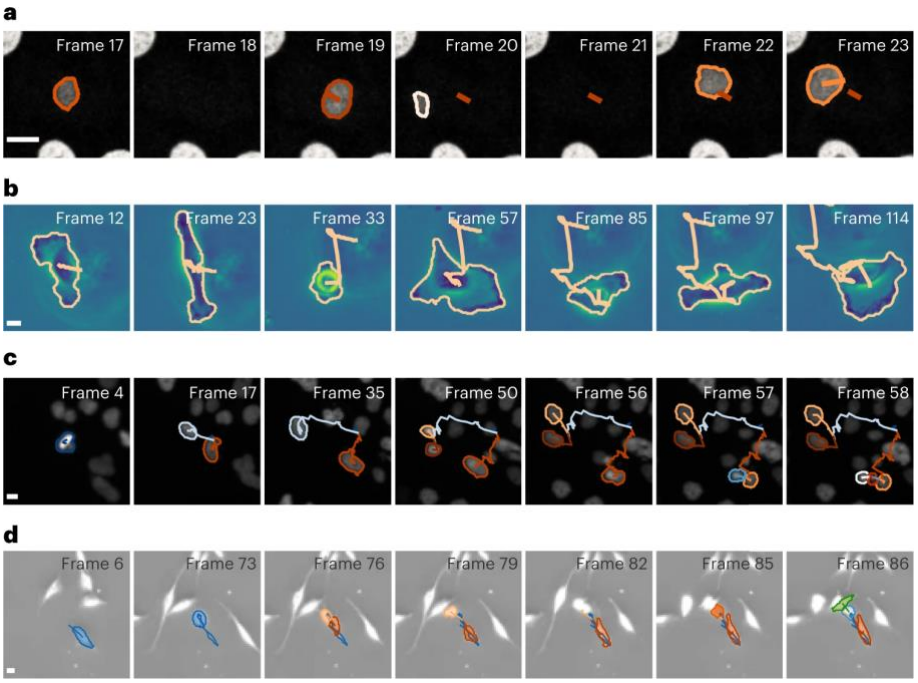


Figure 2. MAGIK reliably links trajectories in various experimental scenarios.

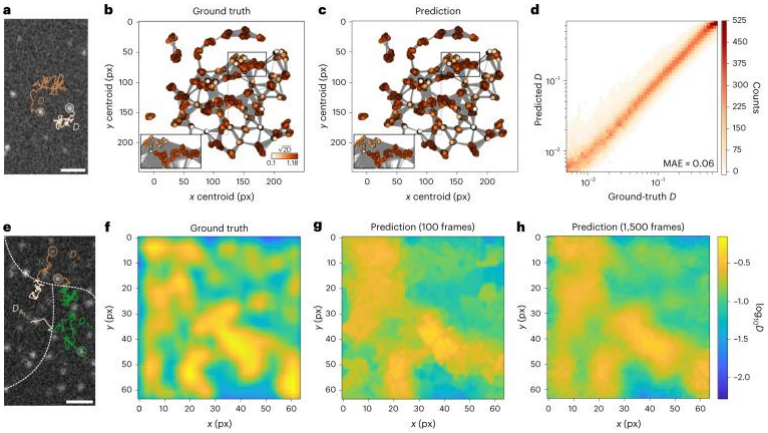
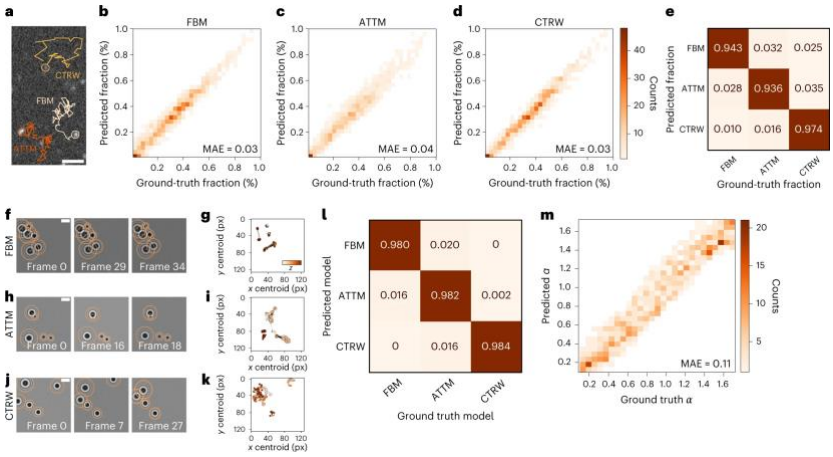


Figure 3. MAGIK reliably links trajectories in various experimental scenarios.

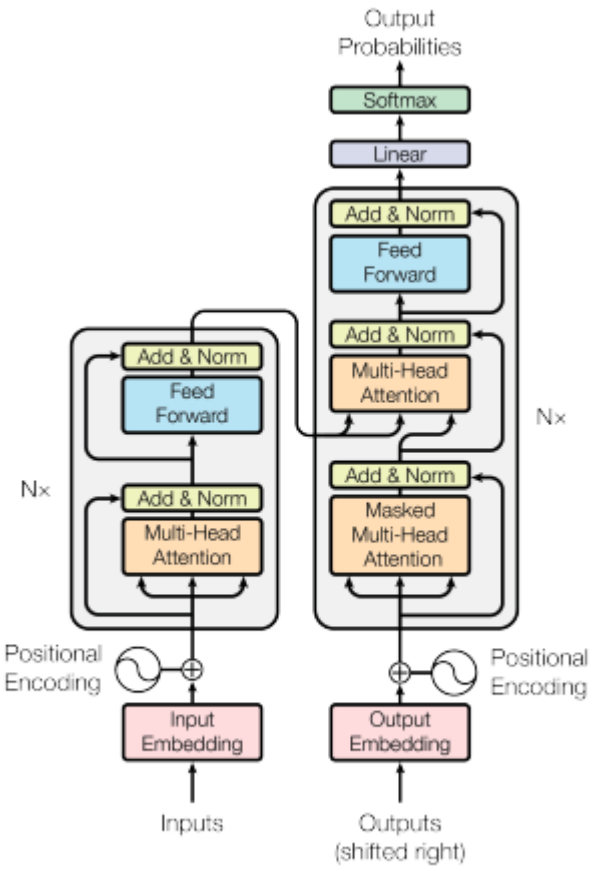
	 Figure 4. MAGIK estimates local and global dynamic properties at the ensemble and single-object levels.
VOCAB: (w/definition)	Geometric deep learning: ML techniques that generalize neural networks to non-Euclidean structures like graphs or manifolds. Graph Neural Network (GNN): Neural architecture where data nodes and edges encode relational information, and message passing updates representations. Trajectory linking: The task of associating detections across frames into coherent object paths. Node regression: Predicting continuous object-level attributes (e.g. diffusion coefficient) directly on graph nodes. Global attribute estimation: Inferring properties of the entire system (e.g. proportion of motion types) as a whole graph-level task. Gated self-attention: Attention mechanism that weights features via gating to modulate influence of different nodes. Fingerprinting graph block (FGNN): A module employed within GNN to encode relational and node-specific updates iteratively. Tracking accuracy (TRA): Metric comparing predicted trajectories vs true trajectories, normalized by graph matching cost. Mean Absolute Error (MAE): Metric for regression tasks measuring average difference between predictions and ground truth.
Cited references to follow up on	• Brückner, D. B., et al. (2021). Learning the dynamics of cell–cell interactions in confined cell migration. Proceedings of the National Academy of Sciences of the *United States of America, 118*, 46-52. https://doi.org/10.1073/pnas.2016602118

	• Helgadottir, S., Argun, A., & Volpe, G. (2019). Digital video microscopy enhanced by deep learning. *Optica, 6*(5), 506–513. https://doi.org/10.1364/OPTICA.6.000506 • Berg, S., et al. (2019). Ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods, 16*, 1226–1232. https://doi.org/10.1038/s41592-019-0582-9
Follow up Questions	How well does MAGIK handle extreme behavior changes (e.g. sudden direction shifts or merging/splitting) in dense systems? Could MAGIK's learned attention weights reveal biologically meaningful interactions (e.g. cell-cell coupling) beyond mere tracking? How transferable is a MAGIK model trained on one cell type or microscopy modality to another without re-training?

Article #11: Attention is all you need

Source Title	Attention Is All You Need
Source citation (APA Format)	Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03762
Original URL	https://arxiv.org/abs/1706.03762
Source type	Journal Article
Keywords	Transformer, Self-Attention, Neural Machine Translation, Encoder-Decoder, Deep Learning
#Tags	#Transformer #Attention #DNN #NLP

Summary of key points + notes (include methodology)	Introduction:
	
	Dominant models (RNNs, LSTMs) process sequences sequentially, which has two negative effects:
	
	• It prevents parallelization
	• It makes learning long-range dependencies difficult.
	
	Methodology:
	
	• Proposed the Transformer
	  ○ Network architecture based on attention mechanism, removed recurrence
	  ○ Also removed need for convolutions entirely.
	• Architecture: Follows an encoder-decoder structure using stacked self-attention and fully connected layers .
	• Self-Attention: This is the beauty of transformers and what allows them to be so powerful in relation to RNNs and LSTMs.
	  ○ It computes relationships between all words in a sequence simultaneously. This is called the "Scaled Dot-Product Attention"
	• Multi-Head Attention: Runs multiple attention layers in parallel
	• Positional Encoding: Since there is no recurrence, sine and cosine functions are added to embeddings to notify the order of tokens .
	
	Results:

	• The model achieved state-of-the-art results on WMT 2014 English-to-German and English-to-French translation tasks • Training was significantly faster (3.5 days on 8 GPUs) compared to previous best models. Implications / Conclusion: Transformers enable massive parallelization during training. They provide a new, superior foundation for sequence transduction tasks, replacing RNNs.
Research Question/Problem/ Need	Can a sequence transduction model rely solely on attention mechanisms to achieve state-of-the-art results while being more parallelizable and faster to train than Recurrent Neural Networks?
Important Figures	 Figure 1: The Transformer - model architecture.

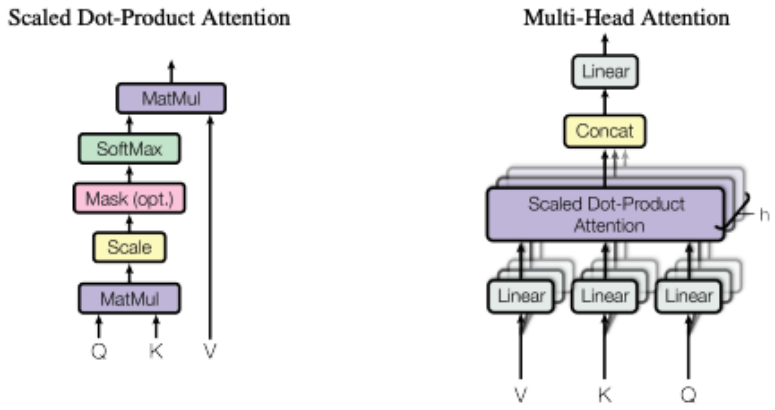
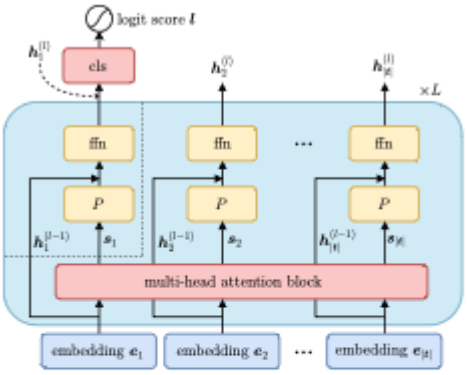
<table>
<tr><td></td><td> 

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.</td></tr>
<tr><td>VOCAB: (w/definition)</td><td>• Transformer: A neural network architecture that relies entirely on self-attention mechanisms.
• Self-Attention: An attention mechanism relating different positions of a single sequence to calculate a representation of the sequence .
• Multi-Head Attention: A technique that linearly projects queries, keys, and values multiple times to perform attention in parallel, allowing the model to capture different types of information.
• Positional Encoding: A method of showing information about the relative or absolute position of tokens in the sequence, using sine and cosine functions.</td></tr>
<tr><td>Cited references to follow up on</td><td>• Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv. https://arxiv.org/abs/1607.06450
• Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv. https://arxiv.org/abs/1409.0473
• Britz, D., Goldie, A., Luong, M.-T., & Le, Q. V. (2017). Massive exploration of neural machine translation architectures. arXiv. https://arxiv.org/abs/1703.03906</td></tr>
<tr><td>Follow up Questions</td><td>1. How does the lack of recurrence impact the model's ability to generalize to sequence lengths much longer than those seen during training?
2. Can this architecture be effectively applied to images or audio?</td></tr>
</table>

Article #12: Attention Mechanisms Don't Learn Additive Models: Rethinking Feature Importance for Transformers

Source Title	Attention Mechanisms Don't Learn Additive Models: Rethinking Feature Importance for Transformers
Source citation (APA Format)	Leemann, T., Fastowski, A., Pfeiffer, F., & Kasneci, G. (2025). Attention Mechanisms Don't Learn Additive Models: Rethinking Feature Importance for Transformers. *Transactions on Machine Learning Research*, In press. https://arxiv.org/abs/2405.13536
Original URL	https://arxiv.org/abs/2405.13536
Source type	Journal Article
Keywords	Transformers, Explainable AI, Softmax, Fidelity, Interpretability
#Tags	#XAI #Transformer #Softmax #Fidelity

Summary of key points + notes (include methodology)	Problem: Traditional XAI methods (LIME, SHAP) rely on linear or additive surrogate models and they assume features contribute independently.
	
	Proposed Solution: The Softmax-Linked Additive Log Odds Model (SLALOM) is introduced as a novel surrogate model explicitly designed to align with the Transformer's function space.
	
	Methodology:
	
	• SLALOM represents a token's role using two scores
	  ○ Token Value *(v*): Describes the independent effect of the token towards a classification.
	  ○ Token Importance *(s):* Provides the token's interaction weight

	when combined with others in a sequence. • The functional form incorporates a softmax link on the importance scores • Algorithms: Two fitting routines are proposed: SLALOM-eff (efficient, uses short sequences) and SLALOM-fidel (maximized fidelity, uses token removal/perturbations). Key Theoretical Results: Learnability (Fidelity): Transformers can be parameterized to reflect SLALOM. Recovery: SLALOM parameters *(s, v)* can be uniquely recovered from model outputs Fidelity: Proposed fidelity was 58% greater than SHAP and 72% greater than LIME, two competing XAI methods.
Research Question/Problem/ Need	Why do conventional feature attribution methods fail to explain the non-linear transformer architecture, and can a novel surrogate model be designed to fix this?
Important Figures	 Figure 2: **Transformer architecture.** In each layer $l=1,\ldots,L$, input embeddings $\boldsymbol{h}_i^{(l-1)}$ for each token i are transformed into output embeddings $\boldsymbol{h}_i^{(l)}$. When detaching the part prior to the classification head ("cls"), we see that the output only depends on the last embedding $\boldsymbol{h}_1^{(L-1)}$ and attention output $\boldsymbol{s}_1$.

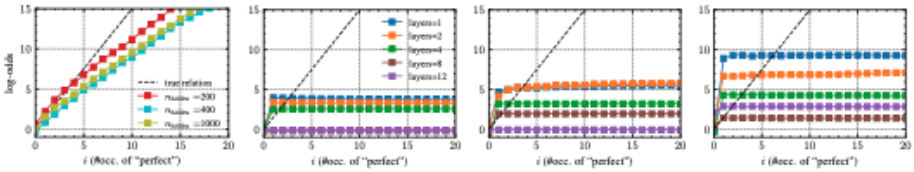
<table>
<tr>
<td></td>
<td>

(a) Fully-Connected Model (b) BERT (c) DistilBERT (d) GPT-2

Figure 3: Transformers fail to learn linear models. We train different models on a synthetically sampled dataset where the log odds obey a linear relation to the features. Fully connected models (2-layer ReLU networks with different hidden layer widths) capture the linear form of the relationship well despite some estimation error (a). However, common transformer models fail to model this relationship and output almost constant values (b)-(d). This does not change with more layers.

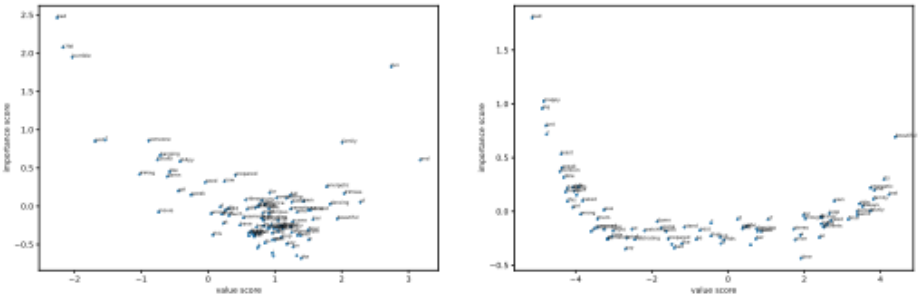

(a) BERT (b) GPT-2

Figure 12: Full scatter plots of SLALOM scores for the sample shown in the main paper (please zoom in for details). We observe that words like “bad” or “fun” get assigned high importance scores and value scores of high magnitude (albeit with different signs) by SLALOM.
</td>
</tr>
<tr>
<td>VOCAB: (w/definition)</td>
<td>

Token Value: The absolute contribution score of a token
Token Importance: The score that determines a token's weight relative to other tokens in the sequence
Fidelity: The property quantifying how well an explanation's predicted output under perturbation matches the complex model's actual output (low error = high fidelity)
Recovery Property: The ability to uniquely re-identify the original model's parameters by fitting the surrogate model

</td>
</tr>
<tr>
<td>Cited references to follow up on</td>
<td>

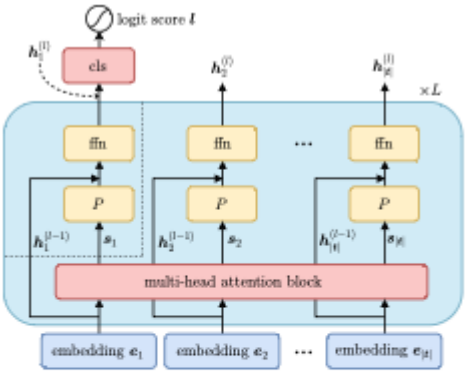
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
Li, J., & Ma, J. (2025). Mixture cure semiparametric additive hazard models under partly interval censoring—a penalized likelihood approach. Statistics and Computing, 35(4), 92.

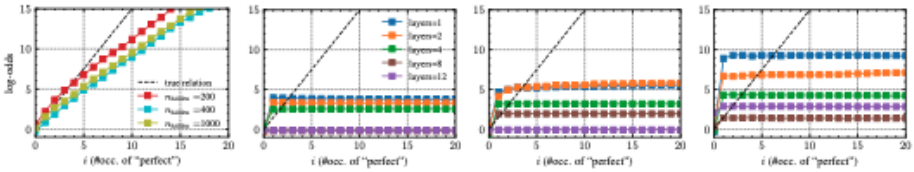
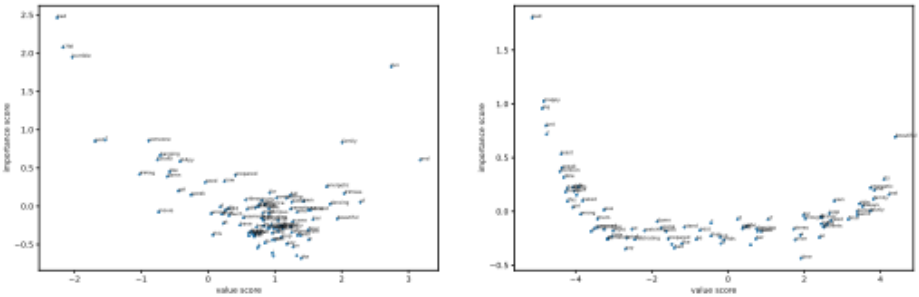
</td>
</tr>
<tr>
<td>Follow up Questions</td>
<td>1. How can a higher order SLALOM be developed to address the acknowledged</td>
</tr>
</table>

	limitation that SLALOM only operates at the token level? 2. Can the use of linearized SLALOM scores be used with the finding that Importance scores better predict human attention?

Article #13: "Why Should I Trust You?" Explaining the Predictions of Any Classifier

Source Title	"Why Should I Trust You?" Explaining the Predictions of Any Classifier
Source citation (APA Format)	Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. http://dx.doi.org/10.1145/2939672.2939778
Original URL	http://dx.doi.org/10.1145/2939672.2939778
Source type	Journal Article
Keywords	Explainable AI (XAI), LIME, Interpretability, Model-Agnostic, Trust, Local Fidelity, Submodular Pick
#Tags	#XAI #LIME #MachineLearning #Interpretability
Summary of key points + notes (include methodology)	Introduction: Trust is fundamental for taking action on predictions and deploying models. Trust is defined in two ways: trusting a prediction (individual) and trusting a model (global). Accuracy metrics on validation data can be misleading due to data leakage or dataset shift. Methodology: • Goal: Identify an interpretable model over an interpretable representation that is *locally faithful* to the classifier. • Proposed mechanism: ○ 1. Interpretable Representation: Converts complex features into binary vectors that shows presence or absence of understandable components.

	○ 2. Sampling: Perturbs the instance by randomly removing components to create a dataset of perturbed samples. ○ 3. Weighting: Weights these samples by proximity to the original instance using a kernel. ○ 4. Optimization: Fits a linear model (using K-LASSO) to minimize the locally weighted loss. Methodology: • Addressed the "trusting the model" problem by selecting a representative set of instances to explain. • Uses submodular optimization (greedy algorithm) to select some instances that cover the most important features without redundancy. Results: • Fidelity: LIME achieved >90% recall in recovering truly important features of interpretable models, outperforming Parzen and Greedy baselines. In a "Wolf vs. Husky" task, LIME revealed the model was spuriously relying on snow in the background rather than animal features
Research Question/Problem/ Need	How can we faithfully explain the predictions of any black-box to allow human users to assess trust in individual predictions and the model as a whole?
Important Figures	 Figure 2: **Transformer architecture.** In each layer $l=1,\ldots,L$, input embeddings $\boldsymbol{h}_i^{(l-1)}$ for each token i are transformed into output embeddings $\boldsymbol{h}_i^{(l)}$. When detaching the part prior to the classification head ("cls"), we see that the output only depends on the last embedding $\boldsymbol{h}_1^{(L-1)}$ and attention output $\boldsymbol{s}_1$.

	 (a) Fully-Connected Model (b) BERT (c) DistilBERT (d) GPT-2 Figure 3: **Transformers fail to learn linear models.** We train different models on a synthetically sampled dataset where the log odds obey a linear relation to the features. Fully connected models (2-layer ReLU networks with different hidden layer widths) capture the linear form of the relationship well despite some estimation error (a). However, common transformer models fail to model this relationship and output almost constant values (b)-(d). This does not change with more layers.  (a) BERT (b) GPT-2 Figure 12: Full scatter plots of SLALOM scores for the sample shown in the main paper (please zoom in for details). We observe that words like "bad" or "fun" get assigned high importance scores and value scores of high magnitude (albeit with different signs) by SLALOM.
VOCAB: (w/definition)	• LIME (Local Interpretable Model-agnostic Explanations): A technique that explains the predictions of any classifier by approximating it locally with an interpretable model. • Model-Agnostic: The ability to explain any machine learning model by treating it as a black box. • Local Fidelity: The property that an explanation must correspond to how the model behaves in the vicinity of the instance being predicted • Interpretable Representation: A representation of data understandable to humans, distinct from the features used by the model.
Cited references to follow up on	• Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research, 11,* 1803-1831. • Wang, F., & Rudin, C. (2015, February). Falling rule lists. *In Artificial intelligence and statistics* (pp. 1013-1022). PMLR.
Follow up Questions	1. How sensitive is the explanation to the choice of kernel width and the number of perturbed samples?

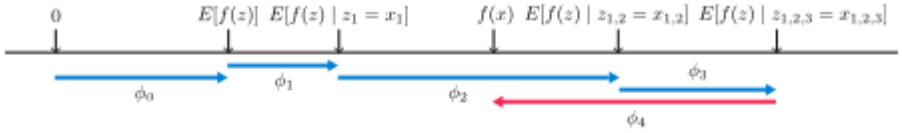
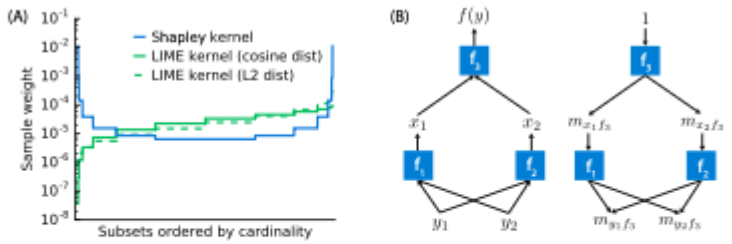
	2. Can LIME fail in scenarios where the local decision boundary is highly non-linear? 3. How does the "interpretable representation" definition limit the types of insights LIME can provide. For example, can it explain interactions between pixels if super-pixels are used?

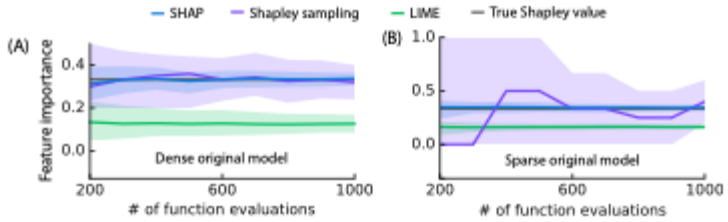
Article #14: A Unified Approach to Interpreting Model Predictions

Source Title	A Unified Approach to Interpreting Model Predictions
Source citation (APA Format)	Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

	https://arxiv.org/abs/1705.07874
Original URL	https://arxiv.org/abs/1705.07874
Source type	Journal Article
Keywords	SHAP, Shapley Values, Explainable AI (XAI), Feature Attribution, Model-Agnostic, Game Theory
#Tags	#XAI #SHAP #MachineLearning #FeatureAttribution

Summary of key points + notes (include methodology)	Introduction/Problem:
	
	Complex models offer high accuracy but low interpretability. Existing explanation methods (LIME, DeepLIFT, LRP) lack a unified theoretical foundation, making it unclear which to prefer.
	
	Methodology (Unified Framework):
	
	• Additive Feature Attribution Methods: Defined a class of explanation models where the prediction is approximated by a linear function of binary variables.
	• Unification: Showed that LIME, DeepLIFT, LRP, and Shapley Regression Values all fall into this class.
	
	Theoretical Guarantee (Theorem 1):
	
	• Proved there is a unique solution in this class that satisfies three desirable properties:
	  ○ Local Accuracy
	  ○ Missingness
	  ○ Consistency (if a model changes to rely more on a feature, its attribution should not decrease) .
	• This unique solution is the Shapley Value from cooperative game theory.
	
	Proposed Method (SHAP):
	
	• SHAP (SHapley Additive exPlanations): Assigns each feature an importance value representing the change in expected model prediction when conditioning on that feature .

	• Kernel SHAP (Model-Agnostic): Uses a specific weighted linear regression (derived from game theory) to estimate Shapley values efficiently, requiring fewer model evaluations than standard sampling . • Deep SHAP: Adapts DeepLIFT to approximate SHAP values for deep neural networks by recursively passing multipliers backwards . Results: Kernel SHAP is more sample-efficient than Shapley sampling and LIME. • Human Intuition: User studies confirmed that SHAP values align better with human explanations than LIME or DeepLIFT. • Fidelity: SHAP provided better discrimination between output classes compared to DeepLIFT and LIME.
Research Question/Problem/ Need	How can we unify the various existing feature attribution methods into a single framework with theoretical guarantees to ensure explanations are consistent, locally accurate, and aligned with human intuition?
Important Figures	 Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.  Figure 2: (A) The Shapley kernel weighting is symmetric when all possible z' vectors are ordered by cardinality there are 2^{15} vectors in this example. This is distinctly different from previous heuristically chosen kernels. (B) Compositional models such as deep neural networks are comprised of many simple components. Given analytic solutions for the Shapley values of the components, fast approximations for the full model can be made using DeepLIFT's style of back-propagation.

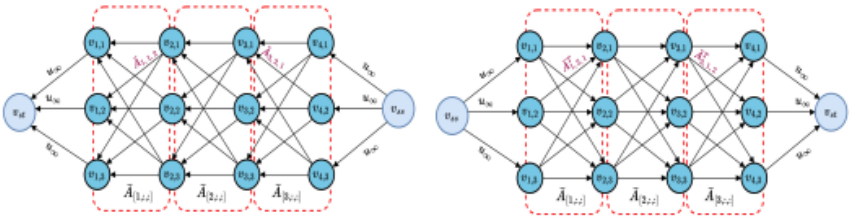
	 Figure 3: Comparison of three additive feature attribution methods: Kernel SHAP (using a debiased lasso), Shapley sampling values, and LIME (using the open source implementation). Feature importance estimates are shown for one feature in two models as the number of evaluations of the original model function increases. The 10th and 90th percentiles are shown for 200 replicate estimates at each sample size. (A) A decision tree model using all 10 input features is explained for a single input. (B) A decision tree using only 3 of 100 input features is explained for a single input.
VOCAB: (w/definition)	SHAP: A unified measure of feature importance that assigns each feature a Shapley value, representing its contribution to the prediction. Additive Feature Attribution: An explanation model that is a linear function of binary variables (feature presence/absence). Shapley Value: A game-theoretic concept that distributes the total payout among players based on their contributions across all possible coalitions. Kernel SHAP: A model-agnostic method that estimates SHAP values using weighted linear regression with a specific "Shapley kernel." Consistency: A property requiring that if a model changes so a feature's contribution increases, its attributed importance should not decrease.
Cited references to follow up on	• Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMlR. • Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* II (Annals of Mathematical Studies, *Vol. 28*, pp. 307–317). Princeton University Press.
Follow up Questions	1. Why exactly does the assumption of feature independence simplify computation?

	2. How does the computational cost of Kernel SHAP scale with the number of features, given the regression complexity is without approximation? 3. In what specific scenarios would LIME be preferable to SHAP, given that SHAP has better theoretical properties but might be computationally heavier?

Article #15: Generalized Attention Flow: Feature Attribution for Transformer Models via Maximum Flow

Source Title	Generalized Attention Flow: Feature Attribution for Transformer Models via Maximum Flow
Source citation (APA Format)	Azarkhalili, B., & Libbrecht, M. W. (2025, July). Generalized Attention Flow: Feature Attribution for Transformer Models via Maximum Flow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 19954-19974).
Original URL	https://arxiv.org/abs/2502.15765
Source type	Journal Article(Preprint)
Keywords	Generalized Attention Flow (GAF), Feature Attribution, Transformer, Maximum Flow, Shapley Values, Log Barrier Method
#Tags	#XAI #Transformer #MaxFlow #ShapleyValues #Attention

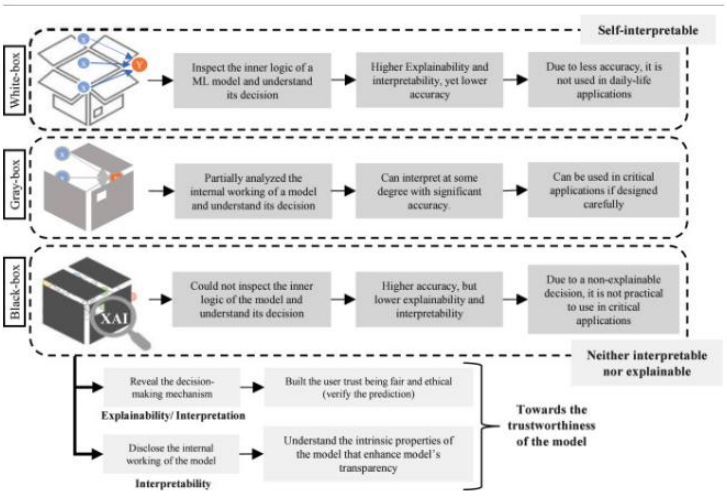
Summary of key points + notes (include methodology)	Problem: Current attention-based attribution methods (like Attention Flow) and gradient methods fail to capture pairwise feature interactions and often violate key XAI axioms (symmetry, efficiency). In addition, the standard Maximum Flow formulation for attribution yields non-unique solutions, making the explanations unreliable. Proposed Solution (GAF): The authors introduce Generalized Attention Flow (GAF), which models the Transformer as a flow network but solves a regularized maximum flow problem to ensure uniqueness and theoretical soundness. Methodology: Information Tensor: Instead of just raw attention, GAF uses three types of tensors to define edge capacities: 1. Attention Flow (Raw weights). 2. Attention Grad Flow (Gradients of weights). 3. Attention times Grad Flow (Element-wise product). To fix the non-uniqueness of Max Flow, they apply the Log Barrier method, creating a strictly convex optimization problem. Axiomatic Guarantee: They prove that the unique solution obtained via this regularization corresponds exactly to Shapley Values, satisfying efficiency, symmetry, nullity, and linearity. Results: Evaluated on SST2, IMDB, Yelp, Amazon, and AG News datasets. • AGF consistently outperformed state-of-the-art baselines (SHAP, IG, LRP, Raw Attention) in AOPC (Area Over Perturbation Curve) and Log odds metrics.
Research Question/Problem/ Need	How can we resolve the non-uniqueness of flow-based feature attribution in Transformers to produce explanations that capture feature interactions and satisfy the theoretical axioms of Shapley values?

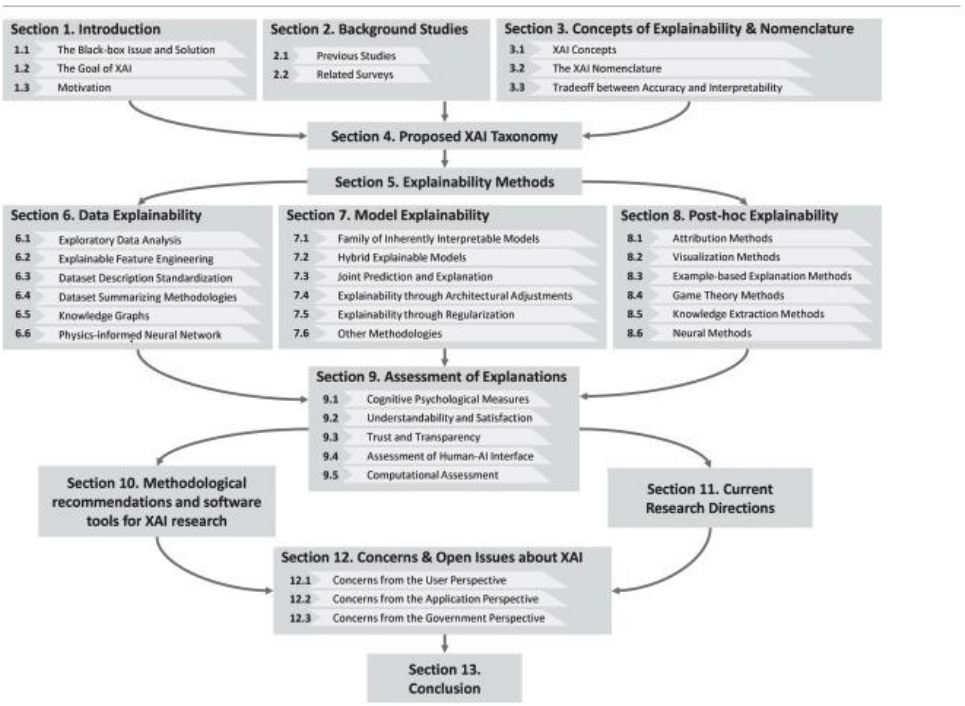
Important Figures	 (a) Schematic information flow created via Algorithm 1.(b) Schematic information flow created via Algorithm 2. Figure 1: Schematics overview of Generalized Attention Flow created using Algorithm 1 and Algorithm 2.
VOCAB: (w/definition)	Generalized Attention Flow (GAF): A framework that treats Transformer attention as a network flow problem, using regularization to ensure unique and axiomatic explanations. Information Tensor: A generalization of attention weights that determines the capacity of edges in the flow graph. Log Barrier Method: An optimization technique that adds a penalty term to the objective function, forcing the solution to be unique and strictly inside the feasible region. AOPC (Area Over Perturbation Curve): An evaluation metric that measures how much the model's accuracy drops when the "most important" tokens are masked (Higher is better). Shapley Values: A game-theoretic method for assigning unique importance scores that satisfy fairness axioms: efficiency, symmetry, nullity, linearity
Cited references to follow up on	• Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928* • Ethayarajh, K., & Jurafsky, D. (2021). Attention flows are Shapley value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 49–54). Association for Computational Linguistics. https://aclanthology.org/2021.acl-short.8/ • Qiang, Y., Pan, D., Li, C., Li, X., Jang, R., & Zhu, D. (2022). AttCAT: Explaining transformers via attentive class activation tokens. *In Advances in Neural Information Processing Systems, 35*.

Follow up Questions	1. Why specifically does the AGF variant outperform the pure Gradient (GF) or pure Attention (AF) variants empirically? 2. The paper notes that computational cost increases with sequence length; How does the runtime compare to SLALOM (from Article #12) which claims high efficiency?

Article #16: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence

Source Title	Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence
Source citation (APA Format)	Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion, 99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805
Original URL	https://doi.org/10.1016/j.inffus.2023.101805
Source type	Journal Article
Keywords	Explainable AI (XAI), Trustworthy AI, Deep Learning, Interpretability, Transparency, Human-centric AI
#Tags	#XAI #Transformer #MaxFlow #Attention
Summary of key points + notes (include methodology)	The authors conducted a systematic review of existing XAI literature to identify a roadmap for transitioning from basic explainability to "Trustworthy AI". • Taxonomy of XAI: The paper classifies XAI into ante-hoc and post-hoc

	explanations for "black-box" models, like LIME or SHAP. • Performance-Interpretability Trade-off: High-performance models such as Transformers and Deep Neural Networks are inherently less interpretable, while simple models are transparent but less accurate. • Post-hoc Local Explanations: Focuses on explaining specific individual predictions, which is the category my project (SLALOM) falls into. • Defines "Trustworthy AI" as requiring more than just math. It must include robustness, fairness, and human-understandable evaluation.
Research Question/Problem/ Need	Current XAI methods are often "model-centric" and lack "human-centric" design, making explanations difficult for non-expert users to interpret.
Important Figures	 Fig. 1. A comparison of white-box, gray-box, and black-box models. On the one hand, white-box models are interpretable by design thus making their outputs easier to understand but less accurate. In addition, gray-box models yield a good interpretability-accuracy tradeoff. On the other hand, black-box models are more accurate but less interpretable. More complex XAI techniques are required for creating trustworthy models.

	 Fig. 2. Detailed overview of the different sections and topics covered in the survey for easing its readability.
VOCAB: (w/definition)	• Ante-hoc Interpretability: Designing models that are simple enough to be understood without external tools. • Trustworthy AI: An AI system that is lawful, ethical, and technically robust • Fidelity: The accuracy with which a surrogate explanation model fits the actual behavior of the original complex model.
Cited references to follow up on	• Georgiev, P., Bhattacharya, S., Lane, N. D., & Mascolo, C. (2017). Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(3), 1-19 • Jain, A., Koppula, H. S., Raghavan, B., Soh, S., & Saxena, A. (2015). Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3182-3190).

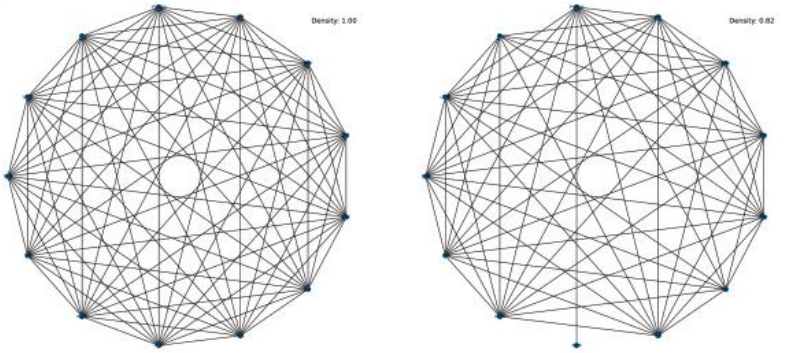
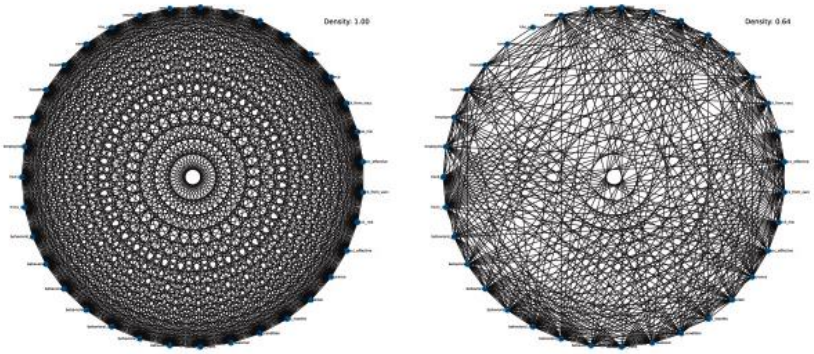
Follow up Questions	Why specifically does the AGF variant outperform the pure Gradient (GF) or pure Attention (AF) variants empirically? The paper notes that computational cost increases with sequence length. How does the runtime compare to SLALOM (from Article #12) which claims high efficiency?

Article #17: ShapG: New feature importance method based on the Shapley value

Source Title	ShapG: New feature importance method based on the Shapley value
Source citation (APA Format)	Zhao, C., Liu, J., & Parilina, E. (2025). ShapG: New feature importance method based on the Shapley value. E*ngineering Applications of Artificial Intelligence, 148*, 110409. https://doi.org/10.1016/j.engappai.2025.110409
Original URL	https://doi.org/10.1016/j.engappai.2025.110409
Source type	Journal Article
Keywords	Explainable artificial intelligence method, Shapley value, Complex artificial intelligence model, Feature importance
#Tags	#XAI #ShapleyValue #FeatureImportance #ShapG #ModelAgnostic #GlobalExplanations
Summary of key points + notes (include methodology)	The paper introduces ShapG (Explanations based on Shapley value for Graphs), which utilizes graph theory to approximate Shapley values more efficiently than traditional methods like KernelSHAP. • Graph Construction: It defines an undirected graph where nodes are features and edges represent Pearson correlation coefficients.

	• Density Reduction: To speed up calculations, the algorithm reduces the graph's density into a sparse, connected graph by keeping only the strongest correlations. • A novel sampling approach is implemented that only considers "reachable nodes" within a certain depth rather than all possible coalitions, drastically reducing computational complexity. • The method was tested on regression ("housing price") and classification datasets using R^2 and F1 scores as characteristic functions. Key Results: ShapG provides more accurate explanations and exhibits "obvious advantages" in running time compared to other cooperative game theory-based XAI tools.
Research Question/Problem/ Need	There is a critical requirement for efficient global explanation methods that can handle complex "black-box" models (like deep neural networks) in high-stakes industries.

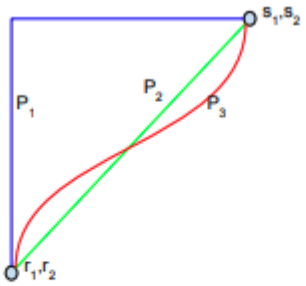
Important Figures	**Algorithm 2** Calculation of the Shapley Value based on graph G'
	Require: A graph $G'(\mathcal{M}, \mathcal{E})$ with $M = \lvert\mathcal{M}\rvert$ nodes
	Ensure: Shapley value component $\phi(i)$ for each node $i \in \mathcal{M}$
	for all nodes $i \in \mathcal{M}$ **do**
	  Initialize $\phi(i) \leftarrow 0$
	end for
	for all nodes $i \in \mathcal{M}$ **do**
	  for all subsets $S \subseteq \mathcal{M} \setminus \{i\}$ **do**
	  Compute $v(S) \leftarrow f(S)$
	  Compute $v(S \cup \{i\}) \leftarrow f(S \cup \{i\})$
	  $\Delta v(S, i) \leftarrow v(S \cup \{i\}) - v(S)$
	  coeff $\leftarrow \frac{S!*(M-S-1)!}{M!}$
	  $\phi(i) \leftarrow \phi(i) +$ coeff $\cdot \Delta v(S, i)$
	  end for
	end for
	  return $\phi(i)$ for all $i \in \mathcal{M}$
	
	Figure 1: Algorithm 2 describes the calculation of the Shapley value of features based on characteristic function defined by (3) following the steps described above.5 Therefore, Algorithm 2 is graph independent

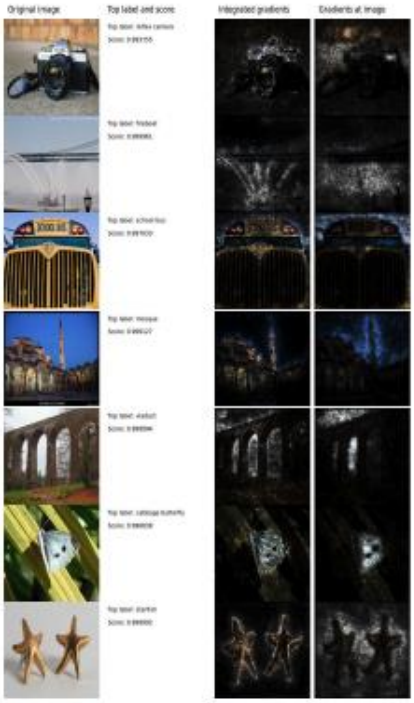
	Fig. 2. Heatmap of Pearson correlation coefficients for the "H1N1" dataset.  (a) Original graph G (b) Reduced graph G' computed by Algorithm 1 Fig. 3. Graph connecting features in "housing price" dataset.  (a) Original graph G (b) Reduced graph G' computed by Algorithm 1
VOCAB: (w/definition)	• Pearson Correlation Coefficient: The metric used to weight edges between feature nodes in the initial graph construction. • Slope Score (S): A metric representing the weighted decrease in model performance as features are removed in order of importance. • Generalized Coupon Collector's Problem: The probabilistic framework used to determine the required number of samples for the graph-based algorithm
Cited references to follow up on	• Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion, 99*, 101805. • Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics, 26*(10), 1340-1347. • Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI

	in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences, 11*(11), 5088.
Follow up Questions	• The paper notes that ShapG currently requires tabular data; can the hierarchical aggregation I am building for SLALOM be adapted to make ShapG compatible with non-tabular Transformer inputs?. • Would using Spearman correlation instead of Pearson (as suggested in Remark 2) improve the model's ability to capture non-linear token interactions in my project?.

Article #18: Axiomatic Attribution for Deep Networks

Source Title	Axiomatic Attribution for Deep Networks
Source citation (APA Format)	Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning, 70*, 3319-3328.
Original URL	https://arxiv.org/abs/1703.01365

Source type	Journal Article
Keywords	Deep learning, attribution, Integrated Gradients, axioms, interpretability
#Tags	#XAI #IntegratedGradients #AxiomaticAttribution #DeepLearning #Interpretability
Summary of key points + notes (include methodology)	The author of this paper introduces Integrated Gradients (IG), a method that attributes a deep network's prediction to its input features by integrating the gradients along a straight-line path from a baseline input to the actual input. Here are the main points: • Axiomatic Approach: The authors identify two fundamental axioms that attribution methods should satisfy: Sensitivity and Implementation Invariance. • The "Saturation" Problem: IG is designed to solve the issue where gradients become zero (saturated) even if a feature is important ○ This is typically a common failure in standard saliency maps. • Completeness: IG satisfies the Completeness axiom, meaning the sum of attributions for all features exactly equals the difference between the model's output for the input and the output for the baseline. • Efficiency: The method requires no modification to the original network and is simple to implement using only a few calls to a standard gradient operator.
Research Question/Problem/ Need	Need: A reliable, theoretically sound method to explain deep neural network predictions that works across different domains without changing the model architecture
Important Figures	 *Figure 1.* Three paths between an a baseline (r_1, r_2) and an input (s_1, s_2). Each path corresponds to a different attribution method. The path P_2 corresponds to the path used by integrated gradients.

	 *Figure 2.* **Comparing integrated gradients with gradients at the image.** Left-to-right: original input image, label and softmax score for the highest scoring class, visualization of integrated gradients, visualization of gradients*image. Notice that the visualizations obtained from integrated gradients are better at reflecting distinctive features of the image.
VOCAB: (w/definition)	• Sensitivity (Axiom): If for every input and baseline that differ in one feature but have different predictions, the differing feature must receive a non-zero attribution. • Path Integral: The mathematical core of IG ○ Calculating the integral of gradients along a straight-line trajectory in the input space from a baseline to the actual input. • Implementation Invariance (Also an axiom): Two networks that always produce the same output for the same input should receive the same attribution scores. • Baseline Input: A neutral reference point against which the actual input is compared to calculate importance.
Cited references to follow up on	• Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. • Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034.*

Follow up Questions	• Can the "Completeness" property of IG be used as a ground-truth metric to validate the faithfulness of my phrase-level aggregations? • Since IG relies on a baseline, what is the optimal "neutral" baseline for natural language phrases: an empty string, or a sequence of pad tokens?

Article #19: The Explainability of Transformers: Current Status and Directions

Source Title	The Explainability of Transformers: Current Status and Directions
Source citation (APA Format)	Fantozzi, P., & Naldi, M. (2024). The explainability of transformers: Current status and directions. *Computers, 13*(4), 92. https://doi.org/10.3390/computers13040092
Original URL	https://doi.org/10.3390/computers13040092
Source type	Journal Article
Keywords	Explainability, transformers, visual transformers, natural language processing, interpretability, deep learning
#Tags	#XAI #Transformers #LiteratureReview #Interpretability #NLP #DeepLearning
Summary of key points + notes (include methodology)	• Methodology: The authors conducted a systematic literature review using the Scopus database, identifying 279 initial records and selecting 95 relevant papers. • Taxonomy of Methods: The paper categorizes XAI techniques for transformers into four primary classes based on their architecture: ○ Activation, Attention, Gradient, and Perturbation. • Current Trends: While attention-based methods are the most frequently employed (appearing in 38 of the analyzed papers), hybrid methods specifically those combining Activation + Attention (like LRP-Rollout) receive significantly higher citation count. • There is a growing trend toward using visual tools (e.g., BertViz, Ecco) to communicate complex feature importance data to non-expert users, though these tools often use existing mathematical backends like gradients or attention. • Although transformers were originally designed for NLP, explainability research has expanded rapidly into Computer Vision (Visual Transformers) and multi-modal tasks.
Research Question/Problem/ Need	Problem: The "black-box" nature of deep learning creates a trust deficit. In addition, existing literature is fragmented.

Important Figures

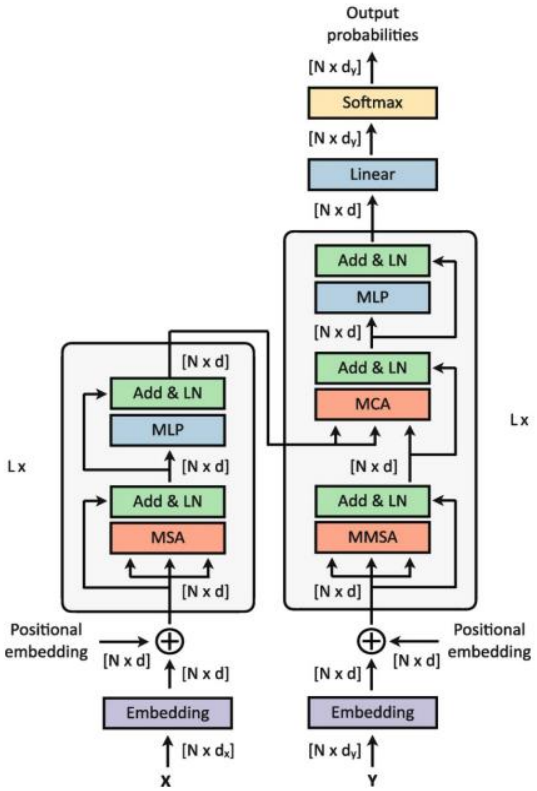


Figure 1. Transformer architecture as shown in [29] (the image is available at https://www.ncbi.nlm.nih.gov/books/NBK597474/figure/ch6.Fig3/?report=objectonly (accessed on 1 March 2024) and is licensed under the terms of the Creative Commons Attribution 4.0 International License).

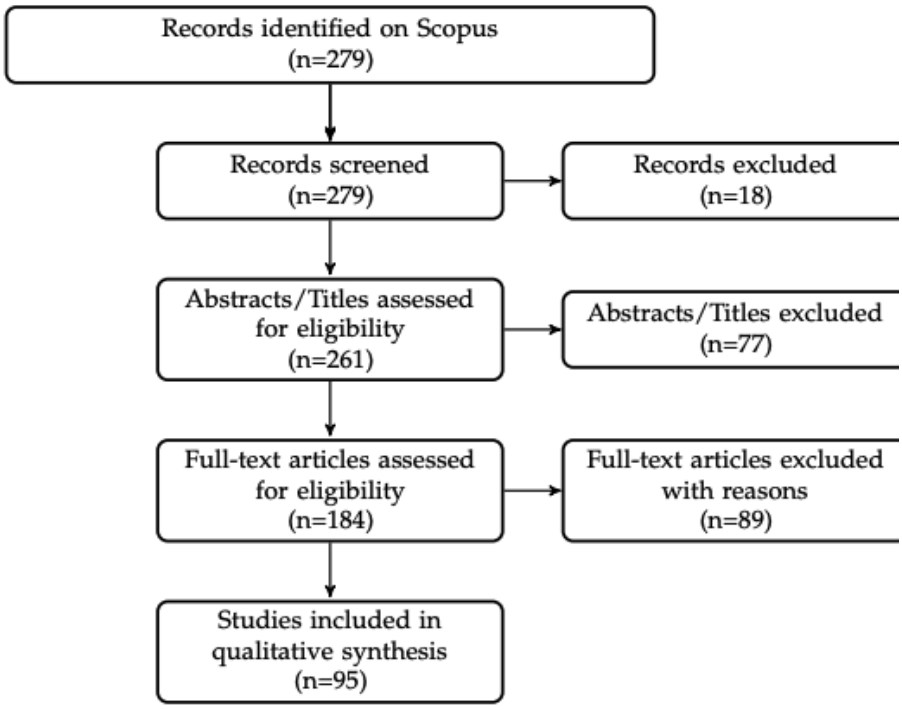
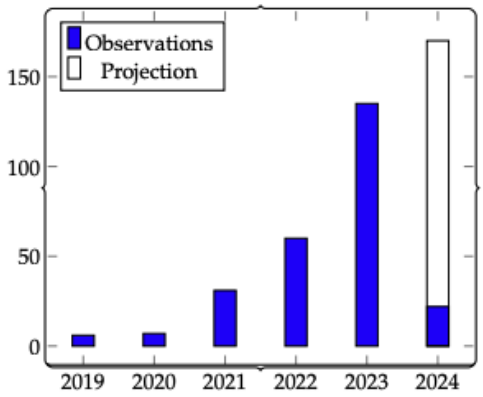
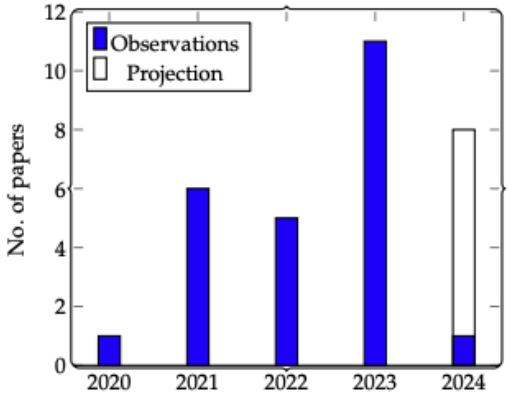


Figure 2. Systematic selection flowchart.

	 **Figure 3.** Number of papers over time.  **Figure 4.** Number of papers introducing new methods by year of publication.

VOCAB: (w/definition)	• LRP (Layer-wise Relevance Propagation): An activation-based method that redistributes the model's output score backward through the layers to identify the relevance
	• Saliency Map: A heatmap visualization used to show the relative importance of different features to the model's final prediction.
	• Model-Agnostic: Explainability methods that can be applied to any machine learning architecture because they do not require knowledge of the internal code or structure.
	• Backpropagation: The standard algorithm for training neural networks that calculates the gradient of a loss function with respect to the network's weights.
	• Encoder/Decoder: The two main sections of the original transformer
	  ○ the encoder compresses input information
	  ○ the decoder builds the output sequence

Cited references to follow up on	• Abbruzzese, R., Alfano, D., & Lombardi, A. (2023). REMOAC: A retroactive explainable method for OCR anomalies correction in legal domain. In *Frontiers in Artificial Intelligence and Applications*.

	• Abdalla, M. H. I., Malberg, S., Dementieva, D., Mosca, E., & Groh, G. (2023). A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers. *Information, 14*(9), 522. • Abnar, S., & Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190–4197).
Follow up Questions	How can the "cognitive mismatch" identified in my project proposal be evaluated using the human-centric benchmarks mentioned in the conclusion of this survey? Since hybrid methods (Activation + Attention) are shown to be more influential (highly cited), would combining my phrase-level scores with activation weights provide a more robust explanation? Does that claim the attention is "not a synonym for explainability" show that standard attention-based XAI methods are flawed?

Article #20: Graph-based Integrated Gradients for Explaining Graph Neural Networks

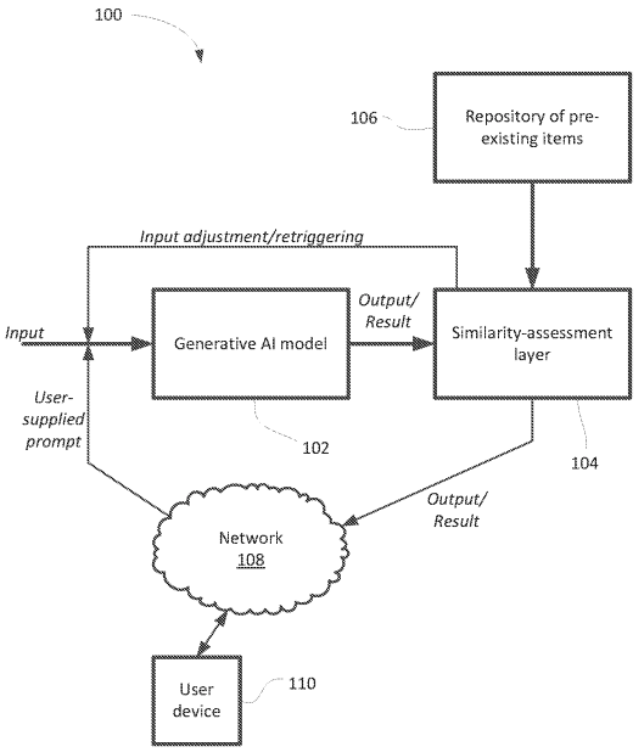
Source Title	Graph-based Integrated Gradients for Explaining Graph Neural Networks
Source citation (APA Format)	Simpson, L., Millar, K., Cheng, A., Lim, C. C., & Chew, H. G. (2025, November). Graph-Based Integrated Gradients for Explaining Graph Neural Networks. In *Australasian Joint Conference on Artificial Intelligence* (pp. 150-162). Singapore: Springer Nature Singapore.
Original URL	https://arxiv.org/abs/2509.07648v1
Source type	Journal Article

Keywords	Explainable AI, XAI, Graph-XAI, Integrated Gradients
#Tags	#XAI #GraphNeuralNetworks #IntegratedGradients #GNN #Interpretability #GraphXAI
Summary of key points + notes (include methodology)	The paper introduces Graph-Based Integrated Gradients (GB-IG), which extends Integrated Gradients (IG) method to the discrete and structured domain of graphs. GB-IG replaces the single straight-line path used in Euclidean IG with the set of all shortest paths between a baseline node and the target node. This incorporates the discrete graph structure into the explanation. • Discretized Integration: The method utilizes a discretized integration formula where the attribution is calculated as the product of the feature difference between successive nodes on a path • Path Aggregation: Attributions are averaged across every identified shortest path • Baseline Selection: To manage the complexity of multiple potential baseline nodes, the authors developed an information-theoretic approach. They calculate path "information" based on the degrees of nodes along a path and then select the baseline that maximizes this entropy. • Evaluation Framework: Performance was benchmarked on four synthetic datasets using a three-layer GCN. Metrics included fidelity , sparsity and the Jaccard Index (overlap with ground-truth motifs).
Research Question/Problem/ Need	Problem: Standard explainability methods like Integrated Gradients are designed for continuous data and cannot handle the discrete, non-Euclidean geometry of graphs when "straight-line" interpolation is not defined.
Important Figures	 **Fig. 2.** Example motifs from ShapeGGen library. Red: node to explain, Yellow: important nodes in the motif, Purple: non-important nodes in the graph. Edge colours are for edge importance which are not considered in this work. **Left**: House motif, **Right**: Circle motif.

VOCAB: (w/definition)	• Path-based Completeness: A new axiom introduced in this paper stating that the sum of attributions equals the output difference when accumulated over all paths in a graph. • Homophily Coefficient: A measure determining the degree to which nodes with similar features or labels are connected within a graph. • Adjacency Matrix: A square binary matrix used to represent the set of edges between all nodes in a graph. • Transductive Model: A type of model where the graph structure is assumed to be both fixed and known during both training and inference. • Shortest Path: On a graph, the path between two nodes that consists of the minimum possible number of edges.
Cited references to follow up on	• Agarwal, C., Queen, O., Lakkaraju, H., & Zitnik, M. (2023). Evaluating explainability for graph neural networks. *Scientific Data, 10*(1), 144. • Bordt, S., Uddeshya, U., Akata, Z., & von Luxburg, U. (2023). The Manifold Hypothesis for Gradient-Based Explanations. *Proceedings of the 2023 IEEE/CVF CVPRW*. • Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*.
Follow up Questions	• Does the "path-based completeness" axiom hold for graphs that are directed or have weighted edges where the shortest path may be more complex to define? • How does the performance of GB-IG change when the graph is highly dense, leading to an increase in the number of potential shortest paths? • In Section 5.1, the paper notes that thresholds for "important" nodes are not standardized in literature. How sensitive are the GB-IG fidelity results to changes in the 0.8 threshold?

Article #21: Similarity-based Generative AI Output Filtering

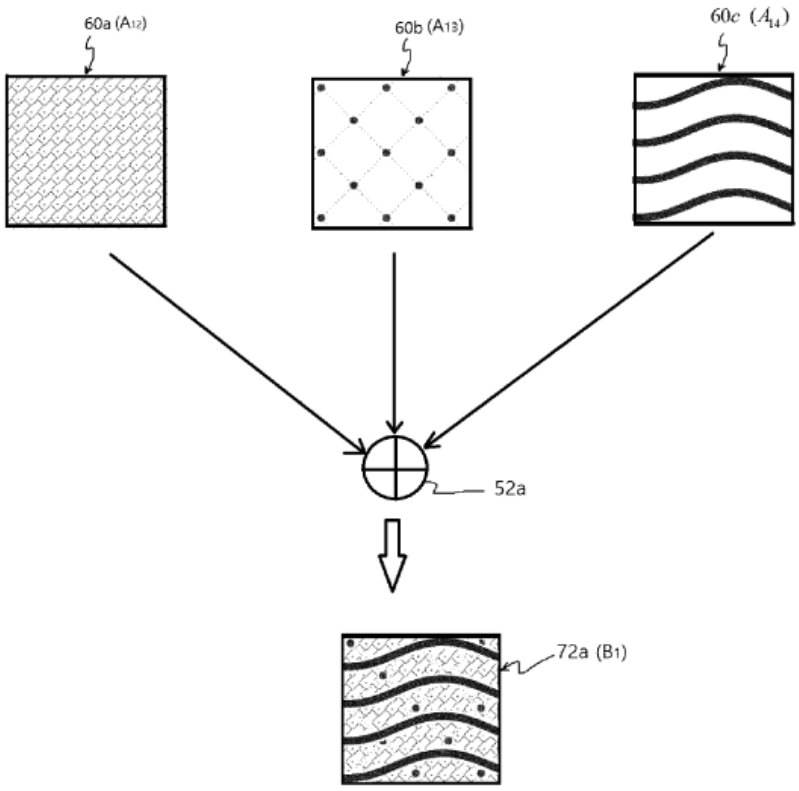
Source Title	Similarity-based Generative AI Output Filtering
Source citation (APA Format)	Padgett, N. L., & Adams, A. (2024). *Similarity-based generative AI output filtering* (U.S. Patent Application No. 18/313,688). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20240160902A1/en
Original URL	https://patents.google.com/patent/US20240160902A1/en
Source type	Patent
Keywords	Generative AI, Output Filtering, Similarity Measure, Intellectual Property, Rote Learning, Prompt Template
#Tags	#XAI #GenerativeAI #OutputFiltering #IntellectualProperty #SimilarityMeasure #RoteLearning #PromptEngineering
Summary of key points + notes (include methodology)	Key Points: The patent describes a system to filter generative AI outputs that are "too similar" to pre-existing content, addressing problems like "rote learning" and potential intellectual property infringement. Methodology: • Similarity Assessment: A dedicated layer calculates distance metrics between the AI's output and items in a repository of pre-existing content . • Iterative Refinement: If excessive similarity is detected, the system automatically adjusts the input or alters model parameters like seed values and noise injection. This is done to steer the model toward a unique result. • Two-Stage Training: A different method involves using a first model to generate a large set of outputs, filtering them for uniqueness, and using that "safe" filtered set to train a second generative model for production use .
Research Question/Problem/ Need	Problem: Generative AI models often produce "problematic" outputs that replicate portions of their training data or closely match existing intellectual property, creating legal and reputational risks .

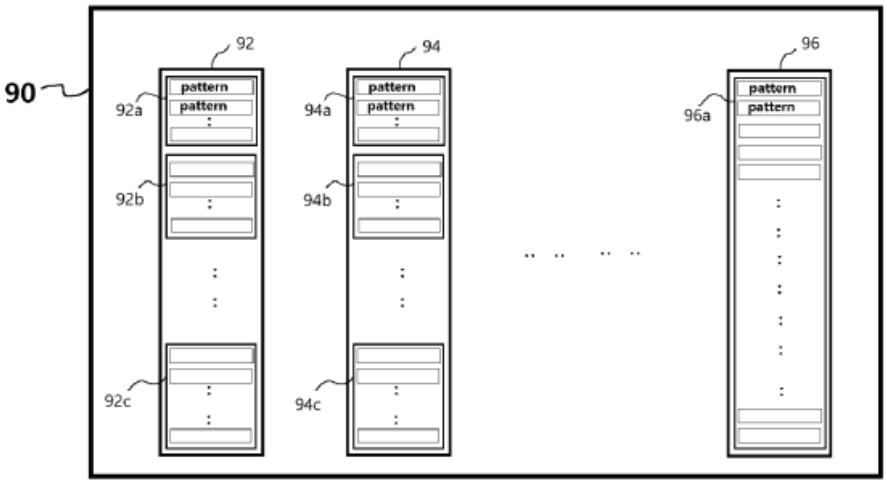
Important Figures	 FIG. 1
VOCAB: (w/definition)	Rote Learning Problem: A scenario where an AI model replicates a portion of its training data in its output instead of generating original content. Similarity-Assessment Layer: A computer-implemented layer that evaluates AI output against a repository to determine if it is too close to existing content. Distance Metric: A mathematical measurement (like Euclidean or cosine distance) used to quantify the similarity between two data items .
Cited references to follow up on	N/A
Follow up Questions	How does the system adjust the similarity threshold in real-time to avoid potential infinite loops when a prompt naturally leads to highly standardized outputs ? What specific "noise" injection methods are most effective at steering the model's convergence away from a problematic output?

Article #22: Artificial Intelligence Neural Network Apparatus and Data Classification Method with Visualized Feature Vector

Source Title	Artificial Intelligence Neural Network Apparatus and Data Classification Method with Visualized Feature Vector
Source citation (APA Format)	Yoo, J. C. (2022). *Artificial intelligence neural network apparatus and data classification method with visualized feature vector* (U.S. Patent No. 11,288,545). U.S. Patent and Trademark Office.
Original URL	https://patents.google.com/patent/US11288545B2/en?oq=11%2c288%2c545
Source type	Patent
Keywords	Artificial Intelligence, Neural Network, Feature Vector, Visualization, Data Classification, Cross-Correlation
#Tags	#XAI #NeuralNetworks #DataVisualization #FeatureVectors #MachineLearning #Classification

Summary of key points + notes (include methodology)	• Methodology:
	  ○ The system classifies data by converting abstract feature vectors into two-dimensional "feature vector images".
	  ○ It calculates cross-correlations between individual elements (i times i) of an N-dimensional feature vector to create a series of pattern images.
	  ○ These individual patterns are synthesized into "local pattern images" and finally into a comprehensive "feature vector image"
	  ○ Finally, it is processed by a deep-learned AI for classification[777].
	• Visual Classification: Unlike standard numerical processing, this patent treats the relationships between features as visual patterns
	• This allows for classification based on the spatial and intensity relationships represented in the generated images.

Research Question/Problem/ Need	Need: An apparatus that can visualize these internal data relationships into a standardized two-dimensional format for providing a visual basis for understanding data patterns
Important Figures	[FIG. 2C] 

	[FIG. 3A] 
VOCAB: (w/definition)	• Feature Vector: An N-dimensional vector of numerical values where each element represents a specific characteristic or "feature" of the input data. • Element Visualizer: A component of the apparatus that maps the relationship between two specific elements of a feature vector into a two-dimensional visual space.
Cited references to follow up on	N/A
Follow up Questions	• How does the apparatus handle feature vectors with extremely high dimensions (N > 1000) where the amount of synthesized cross-correlation images would increase exponentially? • Is the choice of two-dimensional mapping for the feature vector image optimized for specific types of neural network architectures, e.g. CNNs? • In what ways does the "customized weight" in Claim 5 differ from standard weight learning in traditional non-visual neural networks?