Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

ICML 2024

Position: Embracing Negative Results in Machine Learning

Florian Karl[1, 2, 3], Lukas Malte Kemeter[1], Gabriel Dax[1], Paulina Sierak[1]

1 Fraunhofer Institute for Integrated Circuits IIS, Fraunhofer IIS, Nuremberg, Germany
2 Ludwig-Maximilians-Universität München, Munich, Germany
3 Munich Center for Machine Learning, Munich, Germany

Our position: „We argue that machine learning research is at a point where it should encourage or even welcome the publication of negative results.”

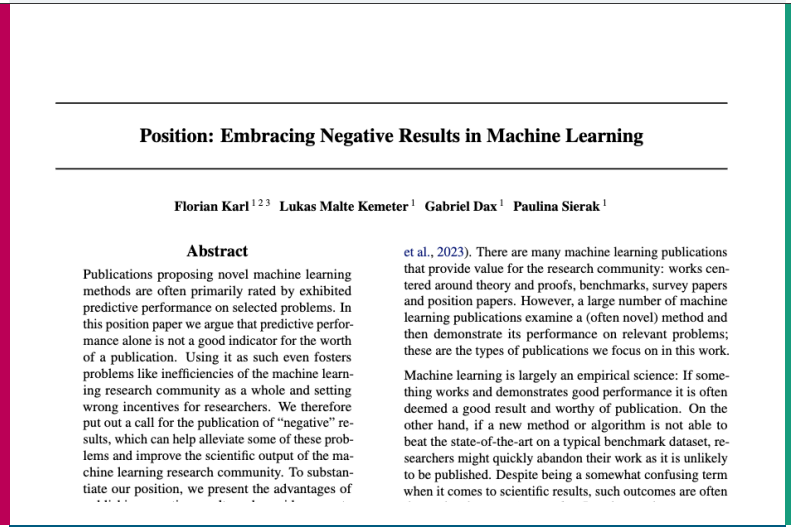
Fraunhofer IIS

Have I heard this before? Why it is worth to listen again.

The struggle to embrace negative results in our community is nothing new – but remains unsolved.

Should we care about negative results?

Do we have a common understanding of the term “negative result”?



Position: Embracing Negative Results in Machine Learning

Florian Karl [1 2 3] Lukas Malte Kemeter [1] Gabriel Dax [1] Paulina Sierak [1]

Abstract

Publications proposing novel machine learning methods are often primarily rated by exhibited predictive performance on selected problems. In this position paper we argue that predictive performance alone is not a good indicator for the worth of a publication. Using it as such even fosters problems like inefficiencies of the machine learning research community as a whole and setting wrong incentives for researchers. We therefore put out a call for the publication of “negative” results, which can help alleviate some of these problems and improve the scientific output of the machine learning research community. To substantiate our position, we present the advantages of

et al., 2023). There are many machine learning publications that provide value for the research community: works centered around theory and proofs, benchmarks, survey papers and position papers. However, a large number of machine learning publications examine a (often novel) method and then demonstrate its performance on relevant problems; these are the types of publications we focus on in this work.

Machine learning is largely an empirical science: If something works and demonstrates good performance it is often deemed a good result and worthy of publication. On the other hand, if a new method or algorithm is not able to beat the state-of-the-art on a typical benchmark dataset, researchers might quickly abandon their work as it is unlikely to be published. Despite being a somewhat confusing term when it comes to scientific results, such outcomes are often

Have we reached an optimal state wrt. dealing with negative results in our community?

What can we do next?

Common understanding of the term

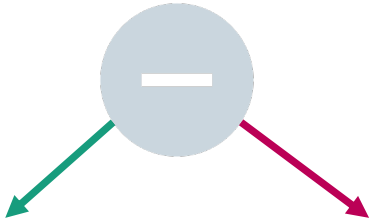
We differentiate between two types of negative results

What is a negative results?

Definition 1: The **usual null hypothesis** of empirical machine learning is that a proposed method does not exhibit significantly better predictive performance than existing methods on a relevant subset of problems.

Definition 2: A **negative result** in empirical machine learning research occurs, when the usual null hypothesis can not be rejected.

Definition 3: A **positive result** in empirical machine learning research occurs when the usual null hypothesis is rejected.



Type 1: "Failure modes of existing methods"
Existing method negative results
(EMNR)

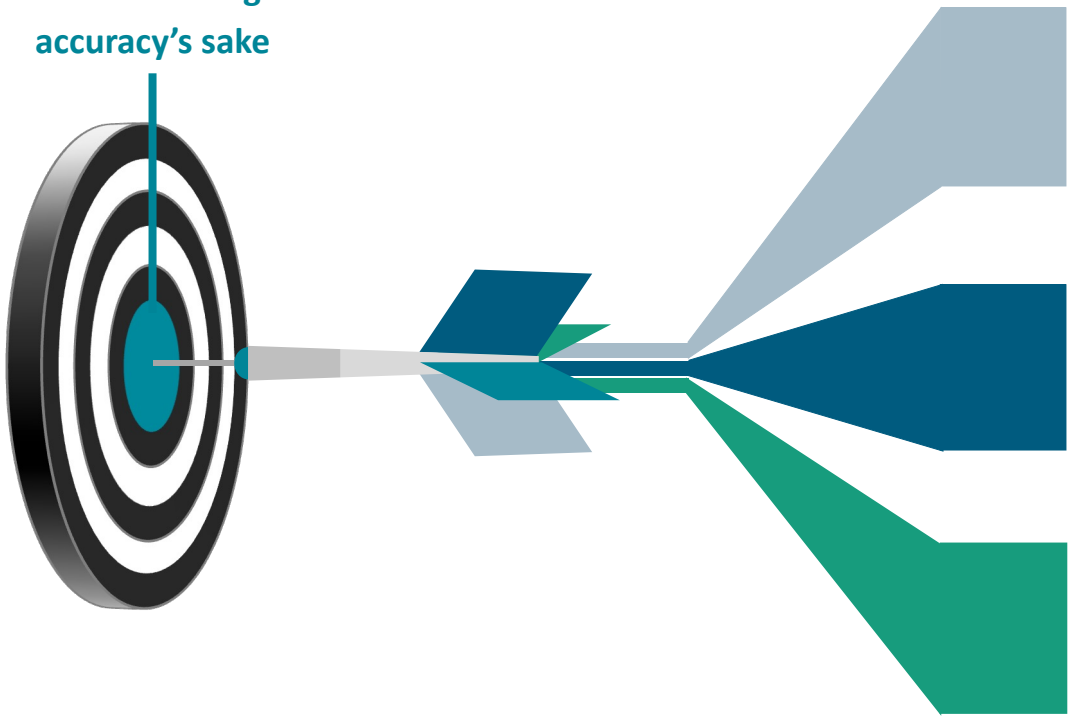
Sometimes published: Vanishing Gradients, Adversarials

Type 2: "New methods that do not beat SotA"
Novel method negative results
(NMNR)

Rarely published!

“Leaderboardism” is a problem in empirical machine learning.

Machine learning for accuracy’s sake



1.) Pure Predictive Performance Is a Faulty Metric for Scientific Progress

- Metrics not necessarily **aligned with impact**
- **Publication bias** and **ε-improvements** lower trust in positive results

Example: [Roberts et al., 2021] examined 2,212 ML models to detect/prognosticate COVID and found none to be of clinical use.

2.) A Hyper-Focus on Predictive Performance Sets Bad Incentives for Researchers

- Researchers are incentivized to submit **only very specific papers**
- Some **Confounding variables** like computing resources are emphasized

3.) Machine Learning Research Has Become Increasingly Inefficient

- Fast-paced environment leads to **parallel works** – negative ones are not revealed
- **Pre-registration** has not taken off in machine learning research

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence, 3(3):199–217, 2021.
Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine, 5(1), 48.

What can we do next:

How to Embrace Negative Results and Success Stories

Special Issues / Conference Tracks / Workshops for Negative Results:

- Showcase important publications with negative results
- Implement this within subfields of machine learning
- Success Stories: “I Can’t Believe It’s Not Better!”-workshop, workshop on “Insights from Negative Results in NLP”

Success Story

“I Can’t Believe It’s Not Better!”-workshop @Neurips

Encourage researchers to discuss negative results:

- Encourage submissions to talk about what didn’t work. Encourage ”challenge papers” to talk about failed attempts
- Success story: iWildCam challenge as part of the workshop on Fine-Grained Visual Categorization at CVPR 2022
- Encourage replication studies, include negative results in teaching.

Make conscious effort to adapt the review process:

- Minimum: Proper guidelines to reviewers (@Journals)
- Re-evaluate certain review criteria with respect to negative results (e.g., “obviousness” of results)? (@Reviewers)
- A structured way for researchers to declare already on submission that their work falls under the category of EMNR or NMNR (@researchers)

Counterfactuals

Opposing positions to better facilitate discussion within the community

1) Publication of negative results lowers the overall quality of research in the field.

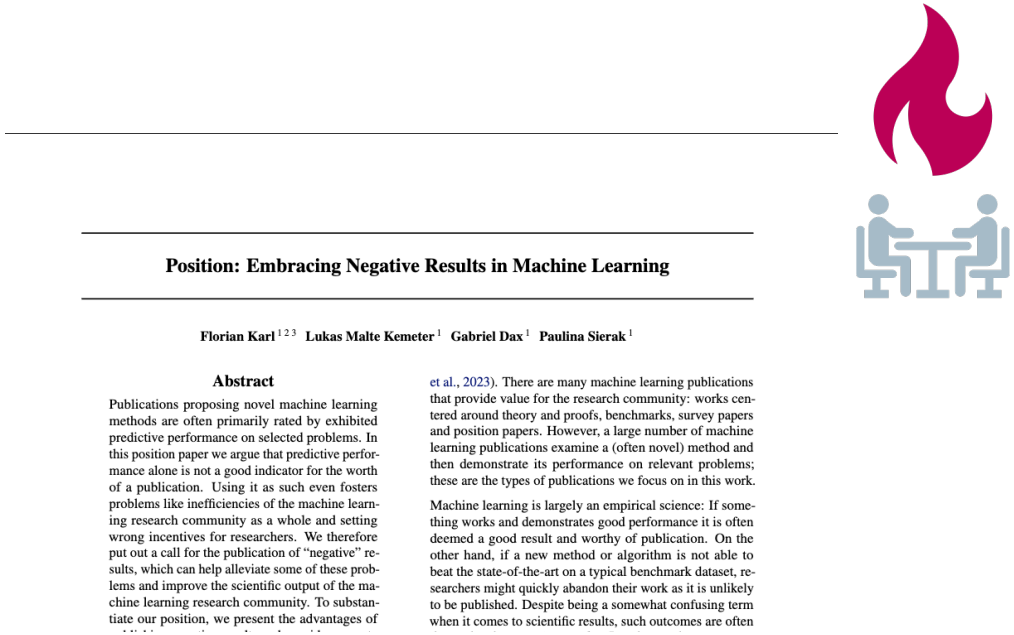
2) Knowing a method does not work in a specific setting has limited value. Knowing it does work in a specific setting is inherently of higher value.

3) New proxies for scientific worth of publications will emerge and a new bias is introduced into what is published.

4) Certain types of negative results are more likely to be published than others.

Our mandate for today: Ignite a discussion

We have an opinion – but everyone here is a stakeholder in this discussion so we appreciate your input!



Position: Embracing Negative Results in Machine Learning

Florian Karl [1 2 3] Lukas Malte Kemeter [1] Gabriel Dax [1] Paulina Sierak [1]

Abstract

Publications proposing novel machine learning methods are often primarily rated by exhibited predictive performance on selected problems. In this position paper we argue that predictive performance alone is not a good indicator for the worth of a publication. Using it as such even fosters problems like inefficiencies of the machine learning research community as a whole and setting wrong incentives for researchers. We therefore put out a call for the publication of "negative" results, which can help alleviate some of these problems and improve the scientific output of the machine learning research community. To substantiate our position, we present the advantages of

et al., 2023). There are many machine learning publications that provide value for the research community: works centered around theory and proofs, benchmarks, survey papers and position papers. However, a large number of machine learning publications examine a (often novel) method and then demonstrate its performance on relevant problems; these are the types of publications we focus on in this work.

Machine learning is largely an empirical science: If something works and demonstrates good performance it is often deemed a good result and worthy of publication. On the other hand, if a new method or algorithm is not able to beat the state-of-the-art on a typical benchmark dataset, researchers might quickly abandon their work as it is unlikely to be published. Despite being a somewhat confusing term when it comes to scientific results, such outcomes are often

! **What is your take on the matter?** Feel encouraged to share your opinion here or later at our poster session. !

Check out our paper.

Thank you for your attention!

Link to Paper

Lukas Malte Kemeter
Machine Learning Researcher at Fraunhofer IIS ADA Center

Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

Thank you for your attention!

Approach us in the break or come to our poster session – we would love to exchange and discuss ideas together.