FALL 2025 VOL. 4

APS Topical Group on Data Science GDS

DATA SCIENCE

APS Topical Group on Data Science Newsletter

Letter from the Newsletter Editor

PHOTO CREDIT: Mikayla Graham

Greetings!

I am excited for this newsletter, which contains career info for those in physics thinking about financial careers in data science as well as info on exciting data science endeavors.

Working on this newsletter has been a nice way to learn more about GDS and data science and physics as well as support others in doing so. Because it is such a great opportunity, I would love to pass this along to someone else in GDS. Are you interested in being the next GDS Newsletter Editor? Please reach out! I am hoping to step down after the 2026 APS Global Physics Summit.

Thanks again for reading this newsletter and supporting GDS.

Best wishes,
Alexis

IN THIS ISSUE

LETTER FROM THE NEWSLETTER EDITOR 1

MESSAGE FROM THE CHAIR 2

INTERVIEW WITH NAGENDRA PANDURANGA, PHYSICIST IN FINANCE 3

KOLMOGOROV-ARNOLD NETWORKS 4

AI MEETS PHYSICS: INSIGHTS FROM THE ML4PS WORKSHOP AT NEURIPS 2024 6

HIGH ENTROPY ALLOYS PHASE STATE PREDICTION AND EXPLORATION WITH MACHINE LEARNING 7

EXECUTIVE COMMITTEE 10

HOW TO CONNECT

There are a variety of ways that you can stay up to date, and informed about GDS. Follow GDS on social media if you haven't done so already!

Email: gds@aps.org
Website: aps.org/units/gds/index.cfm
LinkedIn: linkedin.com/company/apsdatascience
Twitter: twitter.com/apsdatascience
Facebook: facebook.com/APSDataScience

Letter from the Chair

Thank you for the opportunity to serve you as GDS Chair!

With generative AI deeply penetrating people's everyday lives, there is a heightened interest in how data-scientific approaches can transform education and scientific research. The members of GDS had a head start in approaching data through a systematic and principled manner and learning from it. We are now well-positioned to make impactful contributions using our knowledge, experience, and passion. Please spread the word to anyone curious to join the GDS community.

We are also actively preparing for this year's 2026 APS Global Summit! The Global Summit will cover the entire physics spectrum, which was traditionally split between the March and April Meetings in previous years. Please join us at the GDS business meeting to suggest sessions for next year! A special thanks to our program chair Mingda Li and co-chair Aobo Li for organizing the GDS program for the Global Physics Summit. Thank you to all GDS members organizing sessions, sorting abstracts, and contributing to the success of our programs for the Global Summit!

We will be running a short course on "Data Science for Physicists" in partnership with several other units – many thanks to our Past Chair, William Ratcliff, and to our Secretary, Julie Butler, for organizing it! Keep an eye out for future announcements on the GDS IMPACT awards that provide support to students making tremendous contributions to the field and help them attend the Global Summit!

This was a very active year for GDS again! We were very fortunate to have two extremely talented scientists, Frank Noe (Freie Universität Berlin & Microsoft Corporation) and Sergei Gleyzer (University of Alabama), become APS Fellows under the GDS banner! Congratulations, Frank and Sergei!

We also launched several new initiatives to better serve GDS unit members. This year was indeed marked by the establishment of a GDS Industry Advisory Board. We also created a new distinction for early-career researchers and proceeded with the selection of our inaugural GDS Outstanding Dissertation awardees, Biprateep Dey (University of Toronto) and Ouail Kitouni (MIT). Quite an exciting time to be part of GDS! It has been my privilege to serve you and the entire GDS community! Thanks to all Executive Committee members for making GDS a success, with a special thanks to our Past Chair, William Ratcliff, and to our incoming Chair, Eun-Ah Kim. Please encourage your friends and colleagues to join GDS! We are very close to the numbers required to become a division!

Best,

Eun-Ah Kim,
GDS Executive Committee Chair

Interview with Nagendra Panduranga, physicist in finance

By: Amitesh Singh, University of Mississippi

Amitesh Singh (AS): Could you shed some light on your journey from physics to data science? What challenges did you overcome and what were some things which you learned in your Physics PhD to aid you in your data science career?

Nagendra Panduranga (NP): My journey into data science was mainly through finance. After initial years of research in Quantum Condensed Matter, my interests transitioned to complex systems. My PhD thesis was in complex networks and the research group I was part of worked on building physics models to explain stock mar- ket events and other complex systems like cascading failures. My internship with a high frequency trading firm also helped me to get a glimpse into working in Finance professionally. Finally, a senior from my research group referred me to Citi- zens Bank where I joined as Quantitative Analyst and later my position got turned into Sr. Data Scientist to reflect the broad scope of work we do with data.

Data Science is a unique field that requires industry knowledge, being adept with technology and formulating business problems into clear data problems. Working in industry I quickly realized that Industry knowledge is what I lacked and was lost in understanding why certain analyses were performed and why certain problems were worked on. Lack of broader industry knowledge makes information seem scattered without a consistent theme and overwhelming. Asking basic questions without hesitation helped bridge this gap effectively.

A lot of programming background and strong math skills from my theoretical physics PhD helped in picking up the technical aspects of problems faster. Learning R and Python was easier because of past programming knowledge in Fortran and C++. This made up for the gap in domain knowledge. Also, research skills greatly help in starting in a new field.

AS: What projects have you worked on/are working on at your current company? What tools and languages do you use for your daily work, and what is your favorite the projects you are involved in?

NP: I mainly work in credit modeling: developing and re- training loss forecasting models for the Commercial portfolio. The project involves modeling Probability of Default (PD), Credit Rating models and Loss Given Default (LGD) models. These models help the bank to predict the expected losses and set the reserves against these loans. After the 2008-09 Financial Crisis, all banks are required to set aside these reserves (rainy day funds) to make sure they have enough liquidity to avoid a systemic breakdown. There are also quite a few adhoc projects to understand the results of our models. Also, implementing the entire Loss Forecasting machinery (a large group of models) in SAS and R was a project unto itself.

Currently, I am working on stress testing the private credit loans. These are loans given to private hedge funds and structures of these loans are complex. Understanding the finance side and coming up with ways to stress these loans is fun.

For my work, the tools I use vary depending on how in-depth the analysis needs to be. Excel is great for quick analysis. Mostly, I use R, Python, RMarkdown, Jupyter, SAS, SQL, Powerpoint and Latex for my work. In terms of Data Science tools, we stick to supervised learning models because explainability of models is a main requirement.

One of my favorite projects was to understand the model results and come up with exploratory analysis to justify our model usage during Covid 19 pandemic. All credit models use macroeconomic variables as inputs, and artificial spike in unemployment rate during March 2020 made the model outputs not reasonable.

AS: What do you think are some of the most exciting things happening in the data science field right now?

NP: Data Science is a very broad field. At the moment, Artificial Intelligence and specifically, Large Language Models/Transformer technology is in the forefront of most exciting things happening in the modeling side of Data Science. We all know how exciting LLMs are, if you discount the broad existential questions they raise. Currently, we are trying to understand how to best integrate AI into our modeling process and how to satisfy the requirement of explainability in our models since every model needs to be explainable in Banking. There is

a lot of cool stuff happening in development of tools that makes Data science smoother as well. For example, Excel is useful for fast analysis whereas Python/R is useful to modify data. With recent versions of Excel integrating Python, analysis can be even more smooth.

AS: In your experience, what's something everyone assumes is true about data science, but you've found to be completely wrong through your own work?

NP: Everyone (from physics background) assumes that we use a lot of mathematics (e.g., linear algebra) all the time. I have even seen a lot of data science and machine learning courses require linear algebra as a requirement. Most data science positions in industry really don't require use of such heavy math machinery. In reality, a good chunk of the time is spent in locating and collecting the required data in older companies. This might be more streamlined in newer tech savvy companies. After this, a lot of time is spent in cleaning and performing exploratory analysis to understand the data. Also, while understanding the data is important, talking to the Business side that created these datasets provide a lot of valuable feedback and in some cases, even point out if the data is faulty in any way. Only in the last stage, do we build models to predict or explain some variable based on the other variables. As we know, most linear algebra and math is used to come up with ways to fit the models to datasets. These days, most of the models are already implemented as functions in Python,R, or almost any programming language and one rarely goes behind the scenes to the actual code in the model fitting function to modify or build a new model from scratch using Linear Algebra.

AS: Can you recall a project that completely surprised you —either because the data told a very different story than expected, or because something unexpected hap- pened during the process?

NP: During my internship, the trading firm was trying to use News sentiment data to trade profitably. At that time, such trading was supposed to be possible but no trading strategy was known in the company. Just as a background: One assigns Sentiment score to the words in the News Article and therefore, some aggregate sentiment score to the article. The problem is that there is an endless stream of news articles that get published and the goal was to have an automated trading strategy. Luckily, I did come across a possible trading strategy within the limited sample I was working on. The strategy turned out to be very counter-intuitive in the sense that one would have expected to buy stocks/equities when the news sentiment was in one direction but the data was telling us to sell those stocks/equities instead. The pattern actually held up for data spanning over a year. Unfortunately, I had to end my internship and don't know if the pattern would hold up over longer data or could be used as a trading strategy.

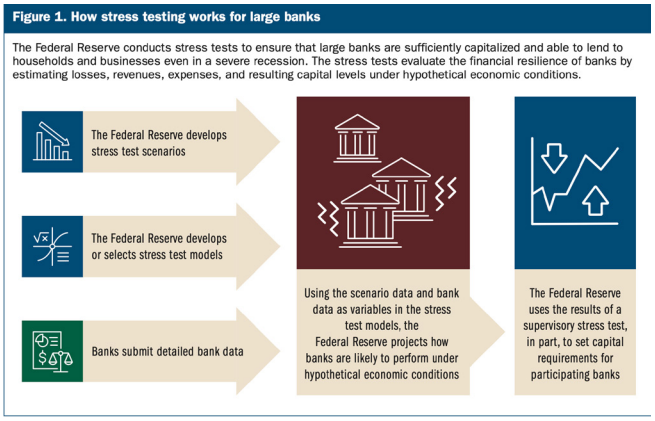


FIG. 1. Nagendra's work mainly focuses on developing the bank's internal version of stress test models that will help the bank pass this stress test performed by Federal Reserve every year and also, help the bank set aside the right amount (not too much or not too little) of reserves for C&I portfolio

AS: Amid the recent advancements in Generative AI, how do you think the field has changed or adapted to this, in terms of the hiring process, and in terms of the work happening in the industry? We have seen that the Gen AI models can code, to varying degrees of accuracy. In this scenario, are pure coding skills still sought after by the hiring companies in Data scientists, or are there other skills which are more important?

NP: Honestly, the answer to the question is evolving rapidly and will change. Gen AI definitely makes the search for informa- tion faster and also coding faster. One does not have to remember syntax for code as much as before. This speeds up the initial exploratory analysis. So far, Gen AI is good at per- forming independent pieces of tasks faster. When it comes to building a full model, a Data scientist still needs to use results of different analysis together in deciding the best modeling approach and a final model. Therefore, Gen AI is still away from doing the job of a full data scientist by itself.

Hallucinations do make it necessary to check everything it does though. At this point, the banking industry is definitely still figuring out internal pilot

projects and a lot of user opinions are being collected. Therefore, Gen AI skills are more looked at as a nice-to-have skill in the hiring process for now. The Hiring process might change after a lot of these test cases play out and the impact of Gen AI on work flows is fully evaluated. Apparently, only 2% of job postings this year explicitly mention AI skills from what I heard in an economics podcast, which is very surprising.

Companies will still expect data scientists to know coding at this time to correct for the hallucinations of Gen AI though. Although, the job scope of data scientists might expand to include heavier coding duties assuming Gen AI is going to help. Thus, data scientist positions will only be merged with roles that do analytical work. Therefore, having strong analytical skills might help differentiate oneself better. Additionally, having good communication skills is very beneficial. Especially, communicating technical results to a non-technical audience can take you far.

AS: Many people think data science is very technical — but what's one non-technical skill you believe made the biggest difference in your career or in your success with projects?

NP: One challenge with data science is that questions useful to the business side of Industry are qualitative and nebulous in a quantitative sense. Oftentimes, business questions are more qualitative like "will this product be successful or is this strategy risky?" Good problem solving skills come in handy in these situations. Formulating these qualitative questions into clear quantitative ones with the view that some quantities can be inferred from data is crucial. For example, "if a strategy is risky" question can be translated into a calculation of probability and all the requisite information needed to calculate this probability needs to be inferred from data. Good communication skills are also needed to engage with Business people to understand the objective. Also, good communication skills enabling you to explain the quantitative analysis back to non-technical business audiences is crucial to show the value of your work. Communicating results effectively has been of great success to me so far.

AS: What is the favorite data science problem you have tackled in your career so far and why? If possible, do you have a picture which summarizes how memorable the problem was?

NP: My internship project is definitely my favorite data project I have worked on. The problem was tackled by many previous interns and other researchers in the company. So, lots of known methods were already tried out and therefore, required quite a bit of first principle thinking to direct the analysis.

Also, I had to use the intuition of investors that trade and how they would behave when positive or negative news would appear. The results of the analysis was that there was a narrow range of windows where the strategy could be used to trade profitably. Unfortunately, I cannot share any plots regarding the project since the results are covered from NDA (Non-Disclosure Agreement). I was able to provide first positive results in this project for the company. The results were very unintuitive. The project taught me that focusing on underlying mechanisms provides valuable insights. The experience also gave me confidence that I could tackle problems outside of academia and the difference between research in academia versus Industry.

AS: What advice would you give to someone just starting their journey in data science?

NP: As someone starting out in Data Science, the number of technical things one has to pick up looks daunting. One does not have to start out as a data scientist, especially if you are graduating from Bachelor's or Masters. The stepping stone positions like data analyst, which requires fewer technical skills and provides opportunity to build on skills required to become a data scientist. Physics education provides a good starting base to work on. Different jobs require differ- ent depths of knowledge in programming, statistics, machine learning and domain knowledge. In this context, I would also suggest not getting too bogged down with the math/theory be- hind different data science techniques and focus on solving a data project using data already available. The biggest bottleneck of real world projects is access to Data and Data is messy in real world projects. Having a portfolio of projects helps showcase the problem solving skills and highlight one's commu- nication skill. In terms of programming languages, I would highly recommend learning to code in Python and some version of SQL.

Amitesh Singh is a PhD Candidate in Physics and Astronomy. He studies black holes and loves analyzing lots of data to model how black holes evolve. In his free time, he travels and learns new languages.

AI meets physics: insights from the ML4PS Workshop at NeurIPS 2024

By: Nicolò Oreste Pinciroli Vago, Information Technology at Politecnico di Milano

Every year, the Conference on Neural Information Processing Systems (NeurIPS) gathers researchers from all over the world working on artificial intelligence. Spanning theory, algorithms, and practical applications, it is one of the most significant venues for artificial intelligence research. Among the workshops, one is particularly interesting for those passionate about physics: the Machine Learning and the Physical Sciences (ML4PS) workshop. This workshop investigates the relationship between physics and machine learning according to two main paradigms: physics-informed machine learning and drawing knowledge from physical data.

Physics-informed learning, on the one hand, incorporates known physical laws, such as differential equations or conservation laws, into the machine learning systems, helping machine learning models to stay physically consistent by using pre-existing knowledge to steer learning.

Conversely, in the absence of a theoretical model, a significant challenge is drawing new knowledge from physical data. In these situations, machine learning can help reveal hidden patterns and infer physical laws. Particularly in complex systems where generating first-principles models is challenging, this data-driven strategy complements conventional physics knowledge.

In the Perspectives track, the workshop also addressed the use of Julia, a programming language that is gaining popularity among the scientific computing community. Used by researchers for constructing and testing machine learning models, Julia is meant for high-performance numerical analysis while maintaining code simplicity. During the workshop, the main advantages and drawbacks of this programming language were explained.

Attending the ML4PS workshop gave us a chance to discover how AI is having an impact on physics research. One emerging topic is multimodal techniques, which use multiple types of data (such as images and time series) to make discoveries. Some of the presented works focused on the applications of large language models (LLMs), which can help analyze, organize, and produce scientific knowledge. Moreover, a significant part of the workshop was dedicated to physics-informed neural networks (PINNs), which were applied to diverse physics applications.

In addition, building interpretable models that not only provide correct predictions but also enable physicists to know why those predictions are made, a major issue when using black-box AI systems in science, was also the focus of papers in diverse fields, including astrophysics and spectroscopy.

To conclude, NeurIPS and ML4PS highlighted how the combination of artificial intelligence and physics is not only a technical challenge but also an emerging field that can help make new scientific breakthroughs. For physics students, this area is a unique chance to be at the crossroads of computation, theory, and experiment.

Nicolò Oreste Pinciroli Vago is a PhD student in Information Technology at Politecnico di Milano, where he also earned his Master's degree in Computer Science and Engineering, and took part in a double degree with NTNU in Simulation and Visualization. His research focuses on multimodal and contrastive learning applied to astrophysics. Recent publications focus on gravitational lensing, anomaly detection and energy efficiency.

High Entropy Alloys Phase state Prediction and Exploration with Machine Learning

By: Omokhuwele Umoru, Yale University

High entropy alloys (HEAs) are a novel class of materials that are composed of random mixtures of five or more pure metallic elements in equiatomic proportions. They set themselves apart from conventional alloys by possessing exceptional physical, thermal and mechanical properties owing to their unique microstructures. Predicting the phases formed in HEAs is crucial for materials design but remains a complex challenge as first principles methods like density functional theory tend to be computationally expensive due to the composition-phase space being vast.

Background

Predicting the phase states of HEAs is critical for understanding their behavior but the complex interactions of multiple elements make it challenging. Phase states are specific atomic arrangements that determine a material's characteristics. Accurate prediction helps choose the right elements and specific metallurgic treatments to achieve desired properties. Traditional methods use thermodynamic properties but those require a lot of data and computational power. Machine learning offers a promise of a faster, more efficient alternative. By analyzing data patterns, machine learning can predict how different elements will behave, helping to develop new HEAs for use in various industries.

Methodology

Data Acquisition and Preparation: Data cleaning addressed missing values and the categorical target variable was label encoded. The input variables were all normalized using MinMaxScaler from Scikit Learn.

Model Development: Four machine learning models were implemented: logistic regression, random forest, support vector classification, and neural network models(Feed Forward Neural Network and Generative Adversarial Network).

Much of the success of machine learning models is based on quality and volume of data, falling short on both criteria leaves room for criticism of the efficacy of such models. Deep learning models also compared to traditional machine learning models require large volumes of data which was inaccessible to the group at this time. The dataset combined an already existing High Entropy Alloy(HEA) dataset with an Multi-Principal Element Alloy(MPEA) dataset and merely had 2700 unique input vectors. It was proportioned into a training and testing set by stratified sampling.

Exploratory Data Analysis

The Pandas and Matplotlib library proved very useful for examining core numerical and categorical features of the dataset. In order to examine correlations between features, the Pearson correlation coefficient was evaluated.

Model Information and Parameters

The logistic regression model used the softmax function to predict the log odds of each class independently, the softmax function also played a role in the feed forward neural network and GAN final output vector creation. The random forest trains an ensemble of decision trees and makes predictions from them, Support Vector Machine (SVM) finds the hyperplane that best separates classes in a high-dimensional space that is typical of HEA compositions.

The Neural Network models synapses in the brain and uses compositions of input functions to model a custom output function(In this case a softmax output). The Feedforward Neural Network (FNN) used four hidden layers, Rectified Linear Unit(ReLU) activation function for each layer, the CrossEntropyLoss Function with mean square error and monotonically decreasing probability dropout regularization, batch normalization, adaptive momentum optimizer and a learning rate of 0.001.

Generative Adversarial Network utilizes two neural networks - A Generator and Discriminator. The Generator learns to create new alloy fractional compositions and the Discriminator attempts to identify predictions made by the generator vs the original dataset. The Generator and Discriminator provide feedback that help both improve and over the course of multiple iterations, the Generator learns to create novel material compositions that more accurately model the probability distribution of our dataset.

The GAN architecture used Binary Cross Entropy with logits loss function for the Generator, Cross Entropy Loss function was used for the Discriminator, batch normalization and leaky ReLU to eliminate the possibility of a vanishing gradient problem.

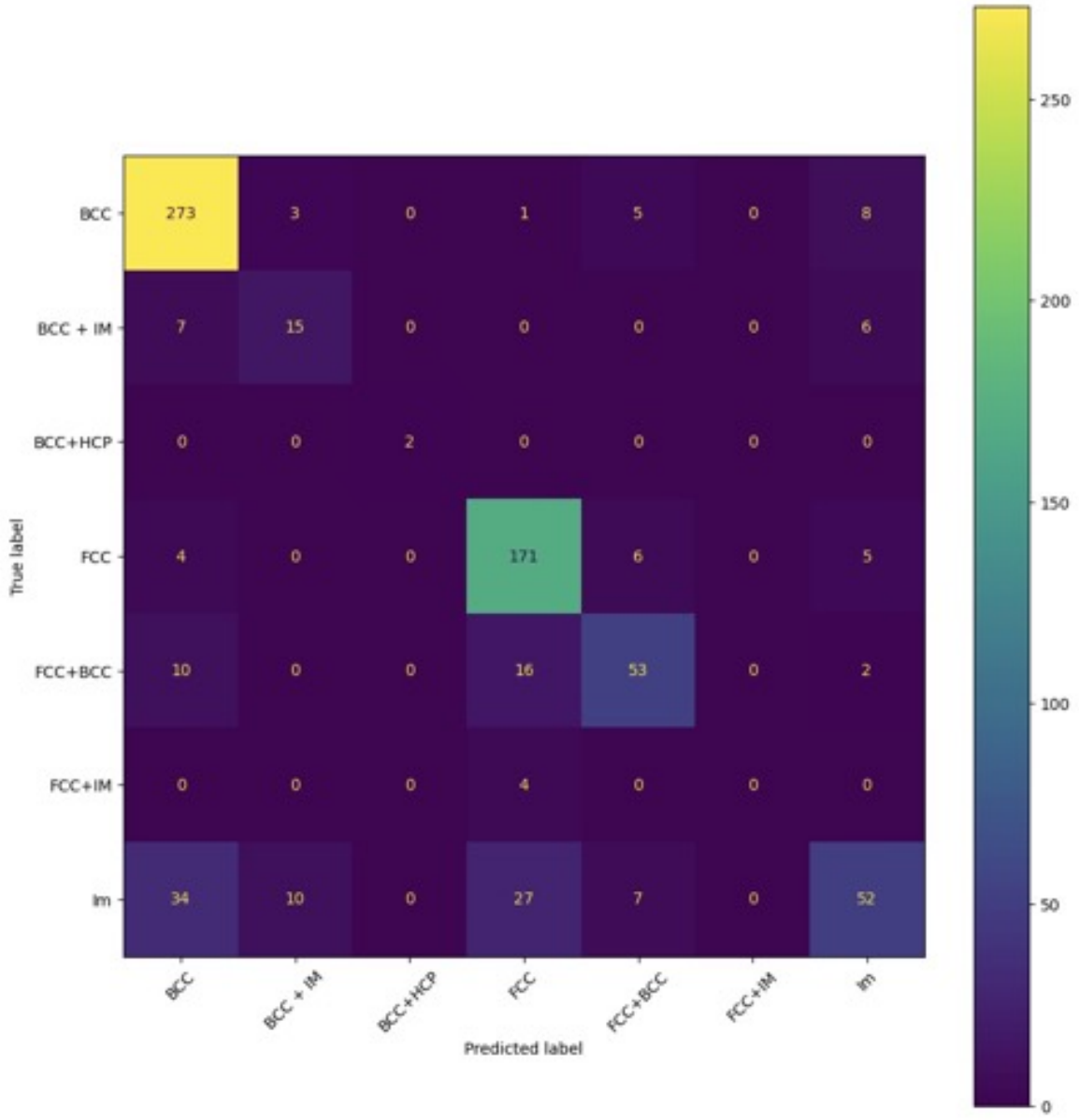
Preliminary Results and Conclusion

The preliminary results showed that traditional machine learning algorithms prove insufficient for the classification problemModel accuracy for tradML models plateaued with random forests at 77%.

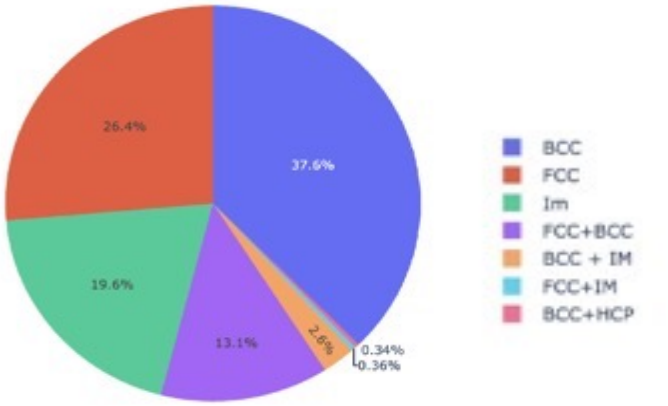
The FNN achieved max accuracy of 80% also showing an inability to model these connections accurately even with deeper networks and attempts to reduce estimation error.

However, results proved promising that machine learning can classify HEA phases, with the neural network and random forest models achieving the highest accuracy.

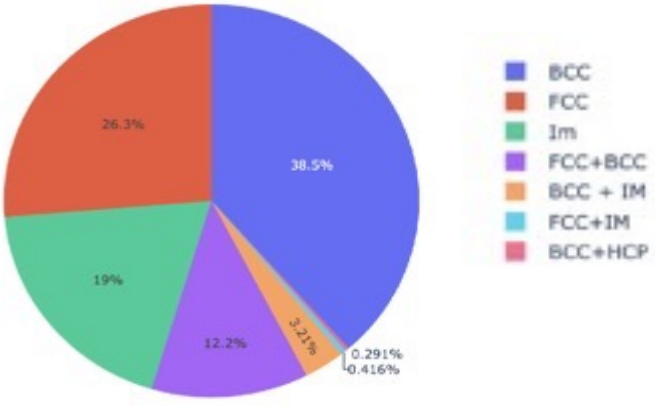
The confusion matrix below visualizes model performance for the neural network.



Original dataset phases chart



Synthetic dataset phases chart



Building a GAN was the first step in exploring the use of generative models, further work will involve diffusion models like variational autoencoders and Flow-based generative models(Normalizing Flows). Flow based models are able to model more complicated distributions than simple gaussians which are typically used for our latent variable of the embedded probability distribution in our generative model. It starts by transforming the simple gaussian into a more complicated probability distribution by applying a sequence of invertible transformation functions(mappings). It flows through a chain of transformations, and repeatedly substitutes the latent variable by applying a change of variables principle and eventually obtains a probability distribution of the final target variable. This architecture will enable the generation of novel phases and atomic compositions of high entropy alloys for experimental validation.

References

- Ye, Y., Wang, Q., Lu, J., Liu, C., & Yang, Y. (2016). High-entropy alloy: challenges and prospects. Materials Today, 19(6), 349–362. https://doi.org/10.1016/j.mattod.2015.11.026
- https://lilianweng.github.io/posts/2018-10-13-flow-models/
- AZoNano. (2023, February 6). What are high entropy alloys? https://www.azonano.com/article.aspx?ArticleID=63
- Peivaste, I., Jossou, E., & Tiamiyu, A. A. (2023). Data-driven analysis and prediction of stable phases for high-entropy alloy design. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-50044-0
- mrzcn. (2021, October 30). High Entropy Alloys EDA and ML. https://www.kaggle.com/code/emrzcn/high-entropy-alloys-eda-and-ml/notebook
- Machaka, R., Motsi, G. T., Raganya, L. M., Radingoana, P. M., & Chikosha, S. (2021). Machine learning-based prediction of phases in high-entropy alloys: A data article. Data in Brief, 38, 107346. https://doi.org/10.1016/j.dib.2021.107346
- Machaka, R. (2021). Machine learning-based prediction of phases in high-entropy alloys.Computational Materials Science, 188, 110244. https://doi.org/10.1016/j.commatsci.2020.110244
- Dieter, G. E. (n.d.). Phase diagram. In Materials Science and Engineering: An Introduction (pp. 7–7). https://uotechnology.edu.iq/dep-laserandoptoelec-eng/branch/lectures/solid%20state/chap ter_7_phase_diagram.pdf

Omokhuwele is an Applied Physics PhD student at Yale. She currently mentors students involved in this research project by the Center for Scientific Machine Learning for Material Science (CSML-MS) - a collaborative research effort with the University of Michigan, Texas A & M University, Texas Southern University and Prairie View Agricultural and Mechanical University funded by the Air Force Office of Scientific Research (AFOSR). This research effort explores the use of machine learning to predict phases in HEAs based on alloy composition and the use of generative artificial intelligence to create new alloy compositions by sampling probability distributions of existing data.

GDS Executive Committee

Chair
Eun-Ah Kim (03/25–03/26)
Cornell University

Chair-Elect
Benjamin Nachman (03/25–03/26)
Stanford University

Vice Chair
Cristiano Fanelli (03/25–03/26)
William & Mary

Past Chair
Jerome P Delhommelle (03/25–03/26)
University of Massachusetts Lowell

Treasurer
Vinicius Mikuni (03/25–03/28)
Lawrence Berkeley National Laboratory

Secretary
Julie L Butler (03/24–03/27)
University of Mount Union

Assigned Council Representative
Amy Y Liu (01/24–12/27)
Georgetown University

Member-at-Large
Jihua Chen (06/23–03/26)
Oak Ridge National Laboratory

Member-at-Large
Casey Berger (06/23–03/26)
Bates College

Member-at-Large
Rebecca K Lindsey (03/25–03/28)
University of Michigan

Member-at-Large
Mark S Neubauer (03/25–03/28)
University of Illinois at Urbana-Champaign

Early Career Member-at-Large
Garrett William Merz (03/24–03/26)
University of Wisconsin - Madison

Early Career Member-at-Large
Nina Andrejevic (03/25–03/27)
Argonne National Laboratory