SURYABHAN SINGH HADA

suryabhan90@gmail.com | Google scholar
+1 (209) 777-2068 | 500 Beale St., San Francisco, CA 94105

SHORT BIO

I am an AI and ML researcher with an extensive track record of peer-reviewed publications, specializing in **Interpretable Machine Learning**. My industry experience is centered on **Large Language Models** and **Interpretable Machine Learning**. Additionally, I have worked on various projects involving **Computer Vision (CV)**, **Financial Analytics**, and more.

EDUCATION

University of California, Merced, Merced, California, USA 08/2016 - 07/2022

- **Ph.D. Candidate** in Electrical Engineering and Computer Science (Machine Learning).
- *Thesis:* Approaches to Interpret Deep Neural Networks
- *Advisor:* Miguel Á. Carreira-Perpiñán.

Indian Institute of Technology (BHU), Varanasi, India 06/2009 - 05/2014

- **Integrated Masters of Technology** (B. Tech & M. Tech) in Mathematics and Computing.
- *Advisor:* Santwana Mukhopadhyay.

PROFESSIONAL EXPERIENCE

LinkedIn (AI Engineer) 06/2022 - present

- **Large Language Models:** Presently involved in the design of Large Language Models as a core modeling team member, focusing on the development of user embeddings for various query searches across the entire LinkedIn platform.
- **Machine learning framework:** Contributed to the creation of a machine learning framework for predicting the lifetime value of users or businesses (monetary contribution by the user) subscribing to various LinkedIn services.
- **Model Interpretability:** Established a Model Interpretability Framework to explain the decisions made by Go-to-Market AI models, enhancing transparency and understanding.

University of California, Merced (Graduate Research Assistant) 08/2016 - 07/2022

- **Class-specific neurons in a deep net:** Created an algorithm to find a small subset of neurons in a deep neural network associated with a specific class and visualization of the distribution of classes in the latent space.
  - By controlling the activation of these neurons, we can make the net only to predict or never predict a given class.
  - These neurons can find the input features essential to a given class.
  - Our approach also allows us to extract if-else type rules from a deep net.
- **Inverse-set of a neuron in a deep net:** Created an algorithm to characterize the behavior of neurons in a deep neural network in terms of what concept in a given class changes the activation of a neuron.
- **Exact counterfactual explanations:** Created exact and efficient algorithms for the non-differentiable problem of counterfactual explanations to interpret large decision trees for datasets with both continuous and categorical variables. Our approach allows us to answer various questions, like finding the closest class to a given input or finding the critical feature to flip the model decision. The algorithms are fast enough for real-time use.
- **Counterfactual Explanations for Ensembles:** Developed an algorithm to efficiently generate counterfactual explanations for interpreting large tree-based ensembles both in regression and classification problems. This algorithm rapidly (within seconds even for large ensembles) produces realistic counterfactual explanations, offering a highly accurate approximation of the NP-Hard problem associated with generating counterfactual explanations in tree-based ensembles.

Cvent, Inc. (Software Engineer) 07/2014 - 06/2016

- Developed a new platform for event management and customized data integration for third parties using Microsoft .net as web tier for UI, Drop Wizard for scalable rest layer, RabbitMq as message broker, Couchbase as NoSql datastore.

PUBLICATIONS

- M. Á. Carreira-Perpiñán and **S. S. Hada**: "*Very fast, approximate counterfactual explanations for decision forests*" in AAAI conference on Artificial Intelligence (AAAI) 2023.
- **S. S. Hada**, M. Á. Carreira-Perpiñán and A. Zharmagambetov: "*Sparse Oblique Decision Trees: A Tool to Understand and Manipulate Neural Net Features.*" Data Mining and Knowledge Discovery (DMKD) 2023.
- **S. S. Hada** and M. Á. Carreira-Perpiñán: "*Interpretable Image Classification using Sparse Oblique Decision Trees*" in proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022.
- **S. S. Hada**, M. Á. Carreira-Perpiñán and A. Zharmagambetov: "*Understanding and Manipulating Neural Net Features Using Sparse Oblique Classification Trees*" in proceedings of IEEE International Conference on Image Processing (ICIP) 2021.
- **S. S. Hada** and M. Á. Carreira-Perpiñán: "*Sampling the "Inverse Set" of a Neuron: An Approach to Understanding Neural Nets*" in proceedings of IEEE International Conference on Image Processing (ICIP) 2021.
- M. Á. Carreira-Perpiñán and **S. S. Hada**: "*Counterfactual Explanations for Oblique Decision Trees: Exact, Efficient Algorithms*" in AAAI conference on Artificial Intelligence (AAAI) 2021.
- **S. S. Hada** and M. Á. Carreira-Perpiñán: "*Style Transfer by Rigid Alignment in Neural Net Feature Space*" in proceedings of IEEE conference on Winter Conference of Applications on Computer Vision (WACV) 2021.
- A. Zharmagambetov, **S. S. Hada** and M. Á. Carreira-Perpiñán, M. Gabidolla: "*Non-Greedy Algorithms for Decision Tree Optimization: An Experimental Comparison*" in proceedings of IEEE International Joint Conference on Neural Networks (IJCNN) 2021.
- M. Á. Carreira-Perpiñán and **S. S. Hada**: "*Inverse classification with logistic and softmax classifiers: efficient optimization*" arXiv:2309.08945.
- **S. S. Hada** and M. Á. Carreira-Perpiñán: "*Sparse Oblique Decision Trees: A Tool to Interpret Natural Language Processing Datasets*" under review in IEEE World Congress on Computational Intelligence (WCCI) 2022.

WORKSHOPS

- **S. S. Hada** and M. Á. Carreira-Perpiñán: "*Exploring counterfactual explanations for classification and regression trees.*" ECML 2021.

TECHNICAL SKILLS

- **Programming Languages:** Python, MATLAB, Java, C++, C#
- **Deep learning tools:** Pytorch, Tensorflow, MatConvNet
- **Databases:** SQL, Couchbase, Cassandra

OTHER

- Reviewer for AAAI 2024.
- Reviewer for ICLR 2024.
- Reviewer for AISTATS 2021–2024.
- Reviewer for ICML 2023.
- Reviewer for NeurIPS 2022 and 2023.
- Reviewer for Explainable Artificial Intelligence Approaches for Debugging and Diagnosis Workshop, NEURIPS 2021.

AWARDS

- UC Merced Bobcat Fellowship, 2018 and 2019
- UC Merced Travel Fellowship, 2019 and 2020