Yukti Makhija

Google DeepMind
🏠 Homepage
✉ makhijayukti@gmail.com
🎓 Google Scholar

Education

Indian Institute of Technology Delhi *2019 - 2023*
B.Tech in Biochemical Engineering with minor in Computer Science **GPA: 8.32/10**
Advisors: Prof. Rahul G. Krishnan (UofT) **Maintained Rank 7 out of 60+**

Work Experience

Google DeepMind (India) *Sept 2023 - Present*
Pre-Doctoral Research Associate, Advisors: Dr. Rishi Saket
Working on ML algorithms surrounding the setup of aggregated data, learning theory and differential privacy

Vector Institute for AI and University of Toronto *May 2022 - May 2023*
Research Intern, Advisor: Prof. Rahul Krishnan
Designed hybrid methods for learning interpretable checklists which combine discrete algorithms with deep learning.

Publications

1. **Weak to Strong Learning from Aggregate Labels** **[P.1]**
   Yukti Makhija, Rishi Saket
   Full Paper Under Review @ ALT 2025 | *Accepted @ OPT-ML Workshop NeurIPS 2024*
2. **Learning Predictive Checklists with Probabilistic Logic Programming** **[P.2]**
   Yukti Makhija, Edward De Brouwer, Rahul Krishnan.
   Full Paper Under Review @ TMLR - Completed 2 positive reviews
3. **Aggregating Data for Optimal and Private Learning** **[P.3]**
   Sushant Agarwal, **Yukti Makhija**, Rishi Saket, Aravindan Raghuveer.
   Full Paper Under Review @ AISTATS 2025 | *Accepted @ OPT-ML Workshop NeurIPS 2024*
4. **Modularity Aided Consistent Attributed Graph Clustering via Coarsening** **[P.4]**
   Samarth Bhatia*, **Yukti Makhija***, Manoj Kumar, Sandeep Kumar.
   Full Paper Under Review @ TMLR - Completed 1 positive review | *Accepted @ OPT-ML Workshop, NeurIPS 2024*
5. **FRACTAL: Fine-Grained Scoring from Aggregate Text Labels** **[P.5]**
   Yukti Makhija, Priyanka Agrawal, Rishi Saket, Aravind Raghuveer.
   Full Paper Under Review @ AAAI 2025 | *Accepted @ FITML Workshop NeurIPS 2024*
6. **Learning predictive checklists from continuous medical data** **[P.6]**
   Yukti Makhija, Edward De Brouwer and Rahul Krishnan.
   Accepted @ ML4H Symposium NeurIPS 2022
7. **Label Differential Privacy via Aggregation** **[P.7]**
   Anand Brahmbhatt, Rishi Saket, Shreyas Havaldar, Anshul Nasery, **Yukti Makhija**, Aravindan Raghuveer.
   Full Paper Under Review @ AISTATS 2025
8. **Severity and mortality prediction models to triage Indian COVID-19 patients** **[P.8]**
   Samarth Bhatia*, **Yukti Makhija***, Sneha Jayaswal, Shalendra Singh, Prabhat Singh Malik, S K Venigalla, Pallavi Gupta, Shreyas N. Samaga, Rabi Narayan Hota, and Ishaan Gupta.
   Accepted at PLOS Digital Health 2022

Awards and Honors

- FRACTAL **[P.5]** was chosen as one of 6 projects for a lightning talk at Google Research Week, standing out from over 30 projects lead by predoctoral researchers due to its significant impact.
- Among 9 from 10,000+ applicants selected for Google DeepMind's prestigious Pre-Doctoral Research Program 2023.
- Awarded a travel grant worth $800 to present my paper at the ML4H Symposium 2022 (held with NeurIPS).
- Awarded the Vector Institute Research Grant 2022 worth CAD $7500 provided through the Pan-Canadian AI Strategy to pursue my internship at Vector Institute.
- Awarded MITACS Globalink Research Internship Scholarship 2022, declined it to pursue an internship at Vector Institute.
- Awarded the Summer Undergraduate Research Award 2021 and a grant of INR 50,000 by IRD Unit, IITD to work on Breast Cancer Risk Prediction using ML.
- Discover and Learn Research Program: Awarded grants of INR 400,000 from IRD Unit, IITD for research projects:
  2020: Digitisation in Healthcare
  2019: Developing a molecular sponge as a novel therapy technique against glioblastoma (brain cancer).

- Secured a **Department Rank of 4 out of 60+** students in semesters II and III (academic year 2020-2021).
- Among the **top 0.49%** of the students in IIT-JEE 2019 exam (out of more than 10,00,000 students).

Key Research Projects

Algorithms for Aggregated Data *Google DeepMind*
Advisors: Dr. Rishi Saket

❖ Weak to Strong Learning from Aggregate Labels *Apr 2024 - Jun 2024*

- Provided the **first algorithmic result for boosting** in the context of Learning from Label Proportions (**LLP**) and Multiple Instance Learning (**MIL**).
- Proved an **impossibility result for LLP** on 2-sized bags for *any weak classifier* by constructing a set of bags and demonstrating that no strong classifier can be derived from any weight assignment across this collection.
- Showed that boosting is **impossible** on 2-sized bags in the **MIL** setting for weak classifiers with *accuracy below 2/3*.
- For LLP, showed that a weak learner on large enough bags can be used to obtain a strong learner for small bags in polynomial time. Additionally, proposed a more efficient sampling-based approach which provides the same guarantees with high probability. **[P.1]**

❖ Aggregating Data for Optimal and Private Learning *Jul 2024 - Oct 2024*

- Developed algorithms to optimally partition instances into bags in LLP and Multiple Instance Regression (MIR), with theoretical guarantees on utility (upper bounds) for both instance and bag-level loss functions in linear regression.
- Proved that in MIR, the optimal bagging strategy reduces to **k-means clustering on instance labels**.
- For bag-loss LLP, we derived an upper bound on MSE, showing that maximizing the minimum eigenvalue of the covariance matrix of aggregated featuresle ads to improved utility. Proposed a **random bagging algorithm** to lower bound the minimum eigenvalue with high probability.
- Extended these results to Generalized Linear Models. Experimental results on synthetic datasets validated k-means feature clustering as an effective **label-agnostic bagging heuristic**. **[P.3]**

❖ FRACTAL: Fine-Grained Scoring from Aggregate Text Labels *Sept 2023 - Feb 2024*

- Developed a novel framework that disaggregates response-level labels into sentence-level pseudo-labels using MIL and LLP, enabling **fine-grained RLHF** without requiring sentence annotations (reducing labeling effort to 10%). Improved model performance by integrating sentence-level **priors into the loss function**.
- Proposed a maximum likelihood **pseudo-labeling** algorithm, **proving its correctness** for multiclass classification.
- Achieved up to **93% performance** of a model trained directly on sentence labels. Demonstrated a **6% improvement in text generation (ROUGE) over Preference RLHF**.
- **Featured** during **Google Research Week 2024** for its significant impact. **[P.5]**

Learning interpretable and predictive checklists *Research Internship + Final Year Project, Vector Institute*
Advisor: Prof. Rahul Krishnan *May 2022 - Aug 2023*

❖ Learning Predictive Checklists with Probabilistic Logic Programming *Sep 2022 - Aug 2023*

- Studied hybrid architectures which integrate combinatorial solvers/algorithms into deep learning models as black boxes and approximate gradients for these layers to complete backpropagation. [**Presentation**]
- Proposed the **first-ever hybrid architecture for learning checklists for decision-making** which uses diverse data modalities as input. Significant improvement over existing methods which only operate over tabular datasets. Achieved this by providing an innovative method for learning discrete structures using **probabilistic logic programming**.
- Introduced a novel regularization to enhance interpretability of the learned concepts and to enforce fairness.
- Developed probabilistic logic programs for learning decision trees and checklist of decision trees. Proposed entropy-based regularization for learning balanced trees to enhance interpretability. **[P.2]**

❖ Learning predictive checklists from continuous tabular data *May 2022 - Sep 2022*

- Developed a **Mixed Integer Program** (MIP) for learning checklists from continuous tabular data.
- Incorporated constraints for learning thresholds to binarize continuous features and for optimizing the checklist structure. **Paper accepted at ML4H Symposium, NeurIPS 2022**. [**Poster**]**[P.6]**

Modularity Aided Consistent Attributed Graph Clustering via Coarsening *IIT Delhi*
Advisors: Prof. Sandeep Kumar *Dec 2022 - Sept 2023*

- Formulated unsupervised graph clustering as a **non-convex optimization problem**, combining modularity maximization with graph coarsening and graph regularization techniques. Developed a **block majorization-minimization** algorithm to optimize the objective, with provable **convergence guarantees** and KKT optimality.
- Proved weak and strong **consistency** of our objective under **Degree-Corrected Stochastic Block Model** (DC-SBM).

- Integrated the objective with GCNs/VGAEs, achieving SOTA results with significantly reduced computation time. **[P.4]**

Visualization Algorithm Design *USC*
Advisors: Prof. Jiapeng Zhang *Apr 2024 - Present*

- Investigated the dynamics of algorithms that optimize 2D point positions based on high-dimensional similarity, derived the **effective loss** function for algorithms which use node sampling techniques, and analyzed how **stable** the distances between pairs of points are during optimization and their **convergence rates**.
- Proposed two novel algorithmic improvements: **momentum** optimization with variable learning rates for different pairs of points and **batch-wise** optimization using node centrality scores.
- Conducted rigorous **statistical evaluations** to validate the improvements over existing methods.

Other Research

Versatile Representations for spectral-temporal data *Summer Research Project (remote)*
Advisor: Dr. Edward De Brouwer *Jun 2023 - Aug 2023*

- Integrated interpretable, non-learnable (mathematical) decomposition techniques for spatial (e.g., graph Fourier and scattering transforms) and temporal (orthogonal polynomial projections) components of the data.
- Performed theoretical analysis to ensure representation quality by deriving error bounds for signal reconstruction using graph filter inversion and polynomial approximations. Established a connection between the smoothness properties of the original and reconstructed graph signals.
- Provided empirical validation of the results and compared performance with learnable methods. [**Report**]

Contrastive learning for semi-supervised segmentation using graphs
Advisor: Prof. Prathosh AP *Jan 2022 - May 2022*

- Developed novel methods for few-shot segmentation in both fully and semi-supervised settings.
- Designed a graph architecture with contrastive learning (CL) to learn node embeddings and structure from support pixels.
- Compared against SOTA methods - vision transformers (ViT), Dense Prediction Transformers (DPT), and cycle-consistent transformers (CyCTR), and with graph CL models - Deep Graph Infomax, GraphCL, and InfoGCL. [**Code**]

Understanding Word Order using Information Theory and ML *Minor Project Semesters V & VI*
Advisor: Prof. Mausam, Prof. Sumeet Agarwal *July 2021 - Apr 2022*

- Proposed four novel scoring functions using Pointwise Mutual Information (PMI) and the Information Locality Hypothesis to predict word order in English. Trained models combining these scores with surprisal and memory-based features.
- Devised novel ways to compute PMI by fine-tuning BERT or training maximum entropy models to derive word embeddings, followed by clustering. Analyzed placement of POS tags in dependency trees of sentences and their construction types to find ordering patterns. [**Code**][**Report**][**Presentation**]

Machine Learning for Healthcare *IIT Delhi*
Advisors: Prof. Ishaan Gupta

❖ **Breast Cancer Risk Prediction for Indian Population** *May 2021 - June 2021*

- Trained machine learning models to predict genetic mutation carriers in patients with a family history of breast cancer, targeting pathogenic mutations like BRCA1/2 and variants of unknown significance. Applied dimensionality reduction techniques and UMAP for visualization. Used SHAP and LIME for model interpretation. Awarded **Summer Undergraduate Research Award 2021** for this work. [**Code**][**Report**][**Presentation**]

❖ **COVID-19 Patient Triaging and Survival Analysis for Indian Patients** *Oct 2020 - Jan 2021*

- Applied survival analysis (Cox Regression) and machine learning (XGBoost) on imbalanced clinical datasets to develop models for predicting disease severity and mortality, achieving an AUC-ROC > 0.9. **[P.8]**

❖ **Psychometric Analysis of bioRxiv papers** *June 2020 - Sept 2020*

- Curated a database of COVID-19 papers on bioRxiv along with their associated tweets. Performed sentiment analysis on the tweets using pre-trained models and mapped out a graph/network of users tweeting and retweeting each paper.
- Developed ranking metrics for the papers based on the Twitter network data.
- Applied clustering algorithms to reveal trends such as prominent keywords and the demographics of the Twitter users.

Course Projects

- **PageRank using Map Reduce programming paradigm over MPI**, *Parallel Programming* [**Code**][**Report**]

- **NAFLD progression in liver biopsies and clinical data using deep learning**, *Machine Learning* [**Code**][**Report**]

Relevant Courses

Mathematics	Convex Optimization (Graduate), Probability Theory, Stochastic Processes, Linear Algebra, Differential Equations, Calculus
Computer Science	Discrete Mathematical Structures, Analysis & Design of Algorithms, Machine Learning (Graduate), Parallel Programming, Data Structures & Algorithms, Bioinformatics Stochastic Control and Reinforcement Learning (Graduate)
Engineering	Process Dynamics & Control, Dynamics of Microbial Systems, Signal Processing, Mass Transfer, Fluid-Solid Dynamics, Fluid Mechanics, Heat Transfer, Transport Phenomena, Material Science and Engineering, Mass and Energy Balances