Minnet Khan

+91-9958240874 khanminnat31@gmail.com linkedin.com/in/minnetkhan minnetkhan.github.io

Education

M.Tech CSE, IIIT-Delhi (CPI - 7.88/10) **Aug 2019 - Jun 2021**

Coursework: *Natural Language Processing, Information Retrieval, Big Data Mining in Healthcare, Statistical Machine Learning, Mobile Computing, Foundations of Parallel Programming*

B.Tech CSE, Jamia Millia Islamia, India (CPI - 7.43/10) **Jul 2013 - Jun 2017**

Coursework: *Data Mining, Artificial Intelligence*

Paper Publication

Mansi G et al. (2022) "Ratatouille: A tool for Novel Recipe Generation." In IEEE International Conference on Data Engineering. DOI: 10.48550/arXiv.2206.08267

Experience

Software Engineer, ServiceNow, India | *Full Time* **Nov 2023 - Present**

- Designing and implementing "Security Incident Response" applications on the cloud-based Software as a Service (SaaS) ServiceNow platform.
- Optimized SIR workspace widgets loading time, reducing it from 2 minutes to within 5 seconds.
- Debugged and fixed a major production issue where observables triggered multiple times, causing delays by calling multiple external Azure Sentinel APIs when creating SIRs.
- Added a feature to attach phishing emails to SIRs for emails sent from Proofpoint.
- Collaborated with teams, customers, and internal stakeholders to resolve and fix production bugs.

Software Engineer, PayU Payments Pvt. Ltd, India | *Performance Awards, Full Time* **Jul 2021 - Nov 2023**

- Led the transition from rule-based to machine learning-based payment gateway selection logic, resulting in a 1.87% improvement in transaction success rates.
- Developed an intelligent reporting system for in-depth analysis, cataloging affected payment modes, and suggesting improvements for success rates.
- Developed a one-click feature to convert credit/debit card transactions into EMI, resulting in a daily GMV of 600K USD.
- Revamped the entire EMI payment mode to comply with the new tokenized flow mandated by the RBI.
- Received a collaboration and ownership award for increasing transaction success rate and GMV of 5 million USD in 3 days.
- Reduced developer's efforts on cards integration and generation of EMI conversion file by half.
- Implemented an EMI tenure-updating feature to enhance system reliability and reduce operational efforts.

Software Engineer, PayU Payments Pvt. Ltd, India | *Intern* **Summer 2020**

- Created a Payment Gateway simulator to perform transactions and test different payments integration.
- This project enabled end-to-end testing and increased potential bug detection.

Software Engineer, Kyk Adventure Pvt. Ltd, India | *Full Time* **Sept 2017 - Jan 2019**

- Developed an admin panel to add and remove tour packages manually; website content auto-updated without developer intervention.
- Created a notification system which sends tour reminders and offers to the users.
- Integrated Razorpay SDK to enable receiving, refunding, and reconciling of payments made through the website.
- Built a system which shows tour packages in trending order.

Projects

Novel Recipe Generation System | https://cosylab.iiitd.edu.in/ratatouille/ **Aug 2020 - May 2021**

- Created a Novel recipe generation system using GPT-2 medium, character-level and word-level LSTM, as well as n-gram models to obtain best performing model.
- Achieved an impressive BLEU score of 0.806 with the GPT-2 medium model.
- Utilized the NLTK library for extensive preprocessing and reduced bias by eliminating similar recipes through cosine similarity within the RecipeDB dataset, which comprises 150K recipes.
- Built a website for realistic novel recipe generation based on user-selected ingredients.

Heart Disease Prediction using Risk factors **Aug 2020 - Dec 2020**

- Developed and applied multiple Deep Learning models on a dataset of 70k patients, predicting cardiovascular disease based on 11 risk factors.
- Conducted rigorous data preprocessing, including outlier removal and min-max normalization, to enhance model performance.
- Achieved a notable accuracy of 73.54% by employing an ensemble approach integrating CNN and LSTM models.

B-Cell epitope prediction **Sept 2020 - Dec 2020**

- Developed a Deep Neural Network and LSTM model for the classification of B-cell epitopes and non-epitopes.
- Conducted thorough data cleaning, which involved removing duplicate sequences and excluding outlier longer sequences.
- Utilized Principal Component Analysis (PCA) for effective dimensionality reduction.
- Achieved a significant AUC-ROC score of 0.79, demonstrating the model's robust ability to distinguish between B-cell epitopes and non-epitopes.

Liver Cancer Stage Detection **Aug 2020 - Oct 2020**

- Developed high-accuracy liver cancer stage detection models using XGBoost on a challenging dataset of 280 samples with 60,484 features.
- Achieved a remarkable 0.665 mean F1-score on a college private leaderboard among 400 students.
- Employed rigorous data cleaning, feature reduction, and normalization techniques to optimize model performance, demonstrating strong data preprocessing skills.

Image Recommendation System **Jan 2020 - May 2020**

- Developed an image recommendation system utilizing the Myntra Fashion product image dataset, encompassing 44,441 images.
- Implemented and trained a range of machine learning models to optimize recommendation performance.
- Successfully attained an impressive 85.7% accuracy rate through the deployment of a CNN model, significantly enhancing product suggestions based on image features.

Almost Deterministic Work Stealing **Jan 2020 - May 2020**

- Implemented the "Almost Deterministic Work Stealing" (ADWS) research paper to improve parallel processing efficiency with optimized multi-threaded applications.
- Improved data locality and maintained efficient load balancing with deterministic task allocation and hierarchical localized work stealing techniques.
- Enhanced memory-bound application performance by up to sixfold compared to traditional work stealing schedulers through ADWS implementation.

Quora questions similarity detector **Apr 2017 - Jun 2017**

- Implemented and fine-tuned machine learning models on a dataset of 404,290 Quora question pairs to determine question similarity.
- Employed text preprocessing techniques, including stemming, lemmatization, and stop word removal, to optimize data quality and model performance.
- Successfully reduced log loss to 0.34 and achieved an impressive 76.28% accuracy using Gradient Boosting Machines (GBM).

Cursive Handwriting Text Recognition System **Jan 2017 - May 2017**

- Developed a system to convert cursive handwritten text into machine-readable format using CNN on the challenging IAM Handwriting dataset comprised of 1500 images.
- Implemented advanced data preprocessing techniques such as binarization, noise reduction, image segmentation, and resizing to enhance model accuracy.
- Produced a comprehensive report that compared multiple classical machine learning models and highlighted the superiority of CNNs for image classification.

Movie Recommendation System **Aug 2016 - Dec 2016**

- Developed a movie recommendation system using hybrid recommender engine by combining Collaborative Filtering and Content-Based Filtering.
- Used Cosine Similarity to get similarity between movies and used Singular Value Decomposition (SVD) for feature reduction.

Technical Skills

Backend: Python, Java, C, JavaScript, NodeJS, ExpressJS, Ruby and MySQL
Frontend: HTML, CSS, Bootstrap and ReactJS
Libraries & Frameworks: PyTorch, Tensorflow, scikit-learn, pandas, NumPy and seaborn

ACTIVITIES & ACHIEVEMENTS

Academic Excellence | *Secured a 99.35 percentile in the GATE Examination and received a two-year scholarship.*
Leadership Roles | *Secretary, Logistics Team in ISTE-JMI, and CSI-JMI Coordinator for LAN Gaming and Chess.*

Statement of Purpose

Applications in biomedical information extraction tasks, such as named entity recognition, motivate me to pursue a doctorate in Data Science and Natural Language Processing (NLP) in Machine Learning. In recent years, breakthroughs have been seen in developing large language models (LLMs) and multimodal learning, and advancements in Machine Learning in biomedical applications have inspired me to explore how these innovations can be applied to solve complex problems. One such problem is to improve medical diagnostics by improving data extraction from unstructured clinical data. Because Columbia University offers exceptional resources, cutting-edge infrastructure, relevant courses, renowned professors, and advanced laboratories, I can perfect my expertise here and collaborate for advanced research to improve the quality of human life. I am confident that the academic environment at Columbia University will provide an ideal platform to contribute to transformative research in Machine Learning within the Natural Language Processing and Data Mining domain.

The world of algorithms and programming was introduced to me in an inspiring seminar titled 'Programming in C' delivered by senior students at Jamia Millia Islamia as I began my bachelor's degree. In light of this revelation, I enrolled in rigorous coursework in Advanced Data Structures and Algorithms. With over 500 problems solved on platforms such as Codeforces and CodeChef, I achieved a 3-star coding certification. Building on this achievement, I became the top coder in my batch of 75 students. Showcasing my skills in the regional ACM-ICPC, I testified to the advanced problem-solving abilities I developed during my undergraduate tenure.

Fascinated by Data Mining and Artificial Intelligence courses, I ventured into the field of Data Science to further practice real-life projects, choosing to address "Cursive Handwriting Text Recognition" for my undergraduate degree major project. Using the renowned IAM Handwriting dataset, I implemented preprocessing methods such as binarization, Gaussian blurring, and feature extraction techniques like static zoning and histogram projection. Throughout my experiments, I applied SVM, KNN, and CNN, showcasing the substantial advantage of CNN over classical ML models. Concurrently, my participation in Kaggle's Quora Question Pair competition involved employing text cleaning, negation handling, stemming, and lemmatization. The result was a notable log loss of 0.34 and an accuracy of 76.28% achieved with the GBM model.

Building upon my earlier accomplishments, I ranked 643rd out of approximately 100,000 participants in a national-level Graduate Aptitude Test in Engineering (GATE) examination, placing me in the 99.35th percentile. This achievement secured my admission to the prestigious Indraprastha Institute of Information Technology, Delhi (IIIT-D), for a master's degree. With a strong interest in data science, I focused on courses such as Natural Language Processing, Data Mining, Information Retrieval, Statistical Machine Learning, and Big Data Mining in Healthcare. These courses deepened my understanding of various machine-learning models and their underlying mechanisms. They also led me to engage in hands-on projects like heart disease prediction, liver cancer stage detection, and a similar image recommendation system. It was through these courses and projects during my master's program that I was able to specialize in Data Engineering.

I played a crucial role alongside my team in our standout project in the Complex Systems Lab (Cosylab), titled 'Ratatouille: A Tool for Novel Recipe Generation,' which led to the publication of a research paper. We preprocessed over 140,000 recipes from the RecipeDB dataset using text cleaning, tokenization, normalization, and spelling correction techniques. My key contribution was using Word2Vec's cosine similarity to reduce bias by excluding similar recipes. My idea to transition from a character-level LSTM model to a word-level LSTM model improved the BLEU score from 0.347 to 0.412. Leading the integration of the GPT-2 medium model

and fine-tuning, our BLEU score significantly increased to 0.806, highlighting my innovative approach and teamwork in this AI-based culinary project. As the dataset was huge, it came with its own challenges of training limitations. To overcome this issue, we implemented data batching. Additionally, my crucial idea and implementation of saving the model at regular checkpoints proved extremely useful as it overcame Google Colab's memory and usage restrictions. Even after graduation, I contributed to the project and research paper development while simultaneously handling my professional duties.

Alongside my academic pursuits, I actively participated in extracurricular activities. During my bachelor's, I also organized Robotics Competitions and LAN gaming events for the Computer Society Of India (CSI-JMI). I also served as a Secretary of the Logistics Team at the Indian Society for Technical Education (ISTE-JMI). Throughout my master's, I served as a teaching assistant and taught courses on Discrete Structures, Advanced Programming, and Fundamentals of Database Systems. As a teaching assistant, I created and graded assignments, quizzes, and exams. I conducted weekly tutorial sessions to help students with course content and projects. I also helped students with their doubts during regular office hours. I created and deployed a website to showcase my research work while participating in various Kaggle competitions. As a result of these experiences, I have demonstrated my dedication, versatility, and commitment to excellence.

During my transition from academia to the professional world, my time at PayU Payments was pivotal. Shortly after joining, I resolved a critical bug, saving $5.2 million for the company with my development efforts. I independently developed an advanced analytical reporting system using Elasticsearch and TimescaleDB. It could process 700 million records and analyze over 10,000 features to provide suggestions and failure reasons within minutes. My engineering and load-testing efforts significantly improved system resilience and scalability. Tasked with improving payment transaction success rates, I implemented a machine learning-driven payment gateway selection system. After experimenting with Logistic Regression and XGBoost, I achieved 92.52% precision and a 76.47% ROC-AUC using Random Forest. In A/B testing, this system outperformed the previous rule-based approach by 1.87%, handling 150,000 transactions daily.

I have an immense alacrity for the opportunity to work with Professor Itsik Pe'er, whose pioneering research in artificial intelligence for biomedical applications profoundly aligns with my academic and research aspirations. His recent work, "Current Status of Artificial Intelligence Methods for Skin Cancer Survival Analysis: A Scoping Review," exemplifies a shared focus on leveraging advanced machine learning techniques for extracting meaningful insights from complex biomedical data. Professor Pe'er's emphasis on integrating multimodal data sources—such as clinical history, genetic information, and pathology—to improve prognostic predictions in skin cancer parallels my interest in using NLP and machine learning to enhance data extraction and predictive modeling for medical diagnostics. I am particularly inspired by his approach to multimodal AI and deep learning, and therefore, I aim to explore similar methodologies to address the challenges of biomedical information extraction from unstructured clinical data. I am confident that under Professor Pe'er's mentorship, I could contribute meaningfully to this vital research area, advancing our shared goals of enhancing AI-driven medical diagnostics and patient outcomes.

My aspiration to pursue a PhD at Columbia University is driven by its academic rigor and the opportunity to engage in cutting-edge Data Science research that advances AI's role in society. It will be a wonderful experience to integrate myself into this dynamic academic community, which will help me grow personally and professionally to become a dedicated researcher. The independence of an academic career and the chance to mentor future researchers particularly excite me. With Columbia University's strong commitment to innovation and excellence, it is the ideal environment to develop my research capabilities.