ARUN SHARMA

https://www.linkedin.com/in/arun08sharma/

132 Englewood Avenue	+1 716 495 8491
Buffalo, NY	arunshar@buffalo.edu

EDUCATION:

State University of New York at Buffalo, Computer Science — **New York, U.S.A**
Master of Science — *August 2016-June 2018*

Gautam Buddha University, Computer Science — **Uttar Pradesh, INDIA**
Bachelor/Masters of Technology — *July 2011-July 2016*

EXPERIENCE:

NASA Europa World-Wind Challenge 2017, NASA Ames Research Centre — **Mountain View, CA**
Project Assistant — *June 2017-August 2017*

- Contributed to an open source ongoing project webGlobe, a powerful tool for visualizing and processing spatial temporal weather dat. Implemented statistical analysis techniques such as correlation relative to the location using Spark on Hadoop Distributed File System and improving throughput of the application and optimizing it while deploying cluster on Amazon S3 (Team of 2).
- Participated in NASA's annual World-Wind Challenge and submitted the results under Academic Track under the guidance of Dr. Varun Chandola (UB).
  ***Platform**: Scala, Java, JQuery/Ajax, JSP/Servlets, JavaScript, HTML5, CSS3, Bootstrap, Spark, MySQL.*

Defence Research and Development Organization — **New Delhi, INDIA**
Defence Terrain and Research Laboratory — *Civil Lines, New Delhi, INDIA*
Research Intern — *May 2015 - July 2015*

- Built a scalable interactive system used for real time features extraction of sky and classification of clouds which further helps to predict the weather conditions.
- Used Image/Video Processing techniques and Machine Learning methods for extracting both tonal and textural features and classification of clouds for further predictive analytics.
  Platform: MATLAB

Solid State Physics Laboratory — *Trimarpur, New Delhi, INDIA*
Software Intern — *May 2013 - July 2014*

- Designed a front-end application of ticketing system indigenous to the lab which is being used for day to day user complaint resolution task in the lab today. (Team of 5).
- Worked as backend developer for building a virtual marketplace to sell and buy car models which can be searched with many attributes defined in the user query. (Team of 4).
  Platform: HTML5, CSS3, JavaScript, JSP/Servlet, PL/SQL, Forms4.5, Report2.5, Oracle 7g Developer Suite.

TECHNICAL SKILLS:

- **Skills**: Python, Java, Scala, R, JavaScript, Bootstrap, JSP/Servlets, JQuery/Ajax, HTML5, CSS3, SQL, PL/SQL, C/C++
- **Tools:** Spark, Keras, Tensorflow, Hadoop, SOLR, Lucene, EC2, S3, Jupyter, Tableau, OpenCV, Scipy, Android Studio.
- **Platform:** Windows, Ubuntu, Mac OSX.

PUBLICATIONS:

A novel approach of adaptive thresholding for image segmentation on GPU, *Fourth International Conference on Parallel, Distributed and Grid Computing (IEEE), Dec-2016.* (link)
Authors: Pawan Kumar Upadhyay, Satish Chandra, Arun Sharma
Graph based technique for Hindi Text Summarization, *International Conference on Information Systems Design and Intelligent Applications (IEEE), Jan 2015.* (link)
Authors: K Vimal Kumar, Divakar Yadav, Arun Sharma**.**

ACIEVENMENTS:

Breaking New Ground
HackHarvard, Major League Hackathon 2018 — ***Harvard University, Cambridge, MA***

- Won Breaking the Ground in Hardware or Best Hardware Hack in one of the prestigious Ivy League 36 hours hackathon of MLH HackHarvard organized by Harvard University (Team of 4).
- Build a hardware tool named Handy from scratch which takes input from the user as speech and convert it into American Sign Language which is coordinated by a robotic mechanical arm.
- Contributed by building an interactive system for fetching data through Audio-Visual channels using various Google Platform API's and pass the extracted features to Arduino instructions which directs the movement of robotic arms.

RESEARCH PROJECTS:

webGlobe: Bringing NetCDF Data to WebWorldWind (Independent Research) *Fall 2017 - Spring 2018*

Supervisor: *Varun Chandola*

- Deploying proposed statistical analysis algorithm on a Multi-Node Cluster using Hadoop Map-Reduce on a Spark based framework and comparing the results with other state of the art climate analysis tools.
- webGlobe is to be tested on climate streaming data for real-time analytics using Apache Kafka and Cassandra. Results to be communicated on IEEE Conference on BigData, 2018.

Medical Image Segmentation using Deep Learning (Independent Research) *Fall 2017 - Spring 2018*

Supervisor: *Mingchen Gao*

- Exploring and Implementing various kinds of state of the art Deep Learning techniques used in the field of Medical Image segmentation of lungs, brain etc.
- Preprocessing complex ultrasound, MRI and DICOM images for further segmentation tasks and exploring different Deep Learning frameworks and libraries.

GRADUATE COURSES:

Artificial Intelligence: Machine Learning (Sargur Srihari), Deep Learning (Sargur Srihari), Computer Vision and Image Processing (Chan Wen Chen)

Information Systems: Information Retrieval (Rohini Srihari), Data Intensive Computing (Bina Ramamurthy), Data Mining and Bioinformatics (Jing Gao)

Networking: Distributed System (Steve Ko)

Algorithms: Parallel and Sequential Algorithms (Russ Miller)

Miscellaneous: Data Quality (Jan Chomicki), Software Engineering Concept (Matthew Hertz)

GRADUATE ACADAMIC PROJECTS (github):

Anomaly Detection using Generative Adversarial Network *Deep Learning*

Supervisor: *Sargur Srihari*

- Implemented unsupervised learning for Anomaly Detection on credit card transaction for detection fraud in any form by comparing models which are commonly used for anomaly detection like Autoencoders and Restricted Boltzmann Machine using AUC score, loss vs accuracy curve etc. and varying their respective parameters.
- Proposing Generative Adversarial Network for anomaly detection and comparing with above models and planning to communicate in International Conference on Machine Learning 2018.

Sentiment Analysis using Recurrent Neural Network with LSTMs *Deep Learning*

- Implemented Recurrent Neural Network using Long-Short Term Dependencies (LSTMs) for classifying the sentiment of the Movie Review as positive or negative. Further planned to implement Bi-directional LSTMs and CNN hybrid model for humor detection on complex dataset such as political debates and speeches.
- The model was implemented in tensorflow and further tuned with optimal parameters such as batch-size, LSTM units, iterations etc. for better training and testing accuracy.

Data Visualization and Association Rule Mining *Data Mining and Bioinformatics*

- Visualizing three biomedical datasets with Principal Component Analysis without using any libraries. Also visualizing and comparing with other visualizing techniques such as t-SNE and SVD.
- Implementing Apriori Algorithm on gene expression dataset without using any external libraries and analyze the changes of the support and confidence and their effects on the algorithm.

Clustering Algorithms using Hadoop MapReduce *Data Mining and Bioinformatics*

- Implemented K-means, Hierarchical Agglomerative clustering with Single Link and DBSCAN on two gene datasets and comparing them through external indexes such as Rand Index or Jaccard Coefficient and further depicting the result through visualization of the data using Principal Component Analysis.
- Implemented K-means using Map Reduce framework on Hadoop Distributed File System by exploiting the functionality of driver, mapper, reducer and combiner for efficient clustering. Also compared performance and evaluation parameters form original non-parallel K-means clustering.

Classification Algorithms using Ensemble Methods *Data Mining and Bioinformatics*

- Implemented state of the art classification algorithms from scratch such as Nearest Neighbor, Decision Trees and Naïve Bayes by changing the parameters and using 10-fold cross validation for evaluation measures such as accuracy, precision, recall and F-measure.
- Further implemented Random Forrest based on the above Decision Tree and applying ensemble method boosting by identifying and boosting weak features on the given dataset and compare the results through above evaluation measures.

Simple Amazon Dynamo with chain replication and Failure Handling *Distributed Systems*

- Implemented simplified version of Amazon Dynamo using Android Studio, a key-value storage replication that handled with chain replication with failure handling.
- Numerous test cases of the application were passed including concurrent read and write operation along with random failure detection of the emulators.

Simple Distributed Hash Table using Chord *Distributed Systems*

- Implemented from scratch the fundamental concept of chord or distributed hash table using Android Studio and performing simple transaction operation such as insert, delete, update and query.

- The application was tested on various operation on each of the emulators and their content provider and checking the destination of each message by comparing their hash values which gives rise to basic architecture of Amazon Dynamo.

Simple Group Messenger Application with TOTAL and FIFO ordering *Distributed Systems*

- Developed a simple group messenger application using Android Studio and implement fundamental concepts of TOTAL and FIFO ordering while transmitting messenger to other emulators.
- The application was testing under numerous test cases including concurrency control of messages which inserting and preserving their TOTAL and FIFO ordering.

Document Analytics using Spark *Data Intensive Computing*

- Implemented word co-occurrence with n-grams on Latin Text along with lemmatization and scaling it over 100 documents on Spark cluster. The results shown significant improvement in time complexity due to its lazy evaluation with the help of Direct Acyclic Graphs operations and it's in-memory computation.

Twitter and Document Analytics with MapReduce using Hadoop *Data Intensive Computing*

- Implement word count using MapReduce using Hadoop framework on over 2000 collected tweets. Also, applied word strips, word pairs and word co-occurrences (n-grams) on Latin Text and scaled it up to 70 documents.

Exploratory Data Analytics on various datasets using R *Data Intensive Computing*

- Analyzed data in various forms and performing various tasks such as finding trending of topic w.r.t location of tweets, executing various queries on Kaggle's European Soccer dataset using dplyr library, performing regression analysis and visualization of the result according to each player's performance through the season.

Communicating Healthcare Data Analytics using Tableau *Data Intensive Computing*

- Made an interactive Dashboard for country wise healthcare analysis involving predictive analysis of next 3 years. Dataset includes TB and HIV incidents, mortality rates and many other attributes with around 3500 instances.

Topic Summarization based on Twitter Data *Information Retrieval*

- Built an application for creating a topic summary based on real time data ingestion and indexing of relevant tweets and divide them into subtopics using Latent Dirichlet Allocation with was further used for creating summary using SMMRY API.

Inverted Index and Boolean Query *Information Retrieval*

- Build inverted index from the information extracted from RCV2 multilingual news corpus using Lucene API. Boolean query processing was done on the inverted index previously build by generating postings list and using AND/OR query with Term at a Time and Document at a time.

Evaluation of Information Retrieval Models *Information Retrieval*

- Configured and improved query search engine with tuning parameters based on various IR models like Vector Space, Okapi BM25 and DFR through comparison based on their Mean Average Precision value on a given set of queries.

Data Ingestion and Indexing *Information Retrieval*

- Over 50000 tweets were collected on various trending topics using hashtags and later indexed as per dates, topic, language, text, location etc. using various tokenizers and filters using Apache SOLR.

Classification of Handwritten Digits with Convolutional Neural Network *Machine Learning*

- Classified Handwritten Digits on USPS image dataset using Logistic Regression and Neural Network with backpropagation without using any external library and were compared with CNN using Tensorflow library giving almost perfect accuracy.

Optimal Parameter Tuning using Regression Models *Machine Learning*

- Implemented Gaussian Radial Basis function on Microsoft Learning to Rank (LeToR) data-set with 46 features and tuned hyper-parameters for achieving optimality with both Closed Form and Stochastic Gradient Descent without using any external library. The results were compared and analyzed with change in parameters.

Probabilistic Graphical Model for Maximizing the Likelihood *Machine Learning*

- Designed a simple probabilistic model on a US News university dataset taking into account of different features and parameters which may or may not depend on each other and finding its maximum likelihood.

Stereo Vision Correspondence using Dynamic Programming *Computer Vision and Image Processing*

- Applied block based matching approach for finding disparity map and compared with Dynamic Programming technique using Longest Common Subsequence based on their MSE value.

UNDERGRADUATE ACADEMIC PROJECTS:

Comparative Analysis of Supervised Classification Techniques on Multispectral Dataset using Ensemble Learning

- Compared different types of classification methods (both supervised and unsupervised) for classifying types of terrain from multispectral dataset. Comparison was based on evaluation measures such as Accuracy, Precision, Recall and F-measure and parameter tuning for improving these results. (link)
- Further ensemble methods such as bagging, boosting and random subspace method was applied prior to classification techniques such as Decision Trees and Random Forrest which results highest accuracy while comparing them with other classifiers. (link)

Intuitive K means method for Renal Calculi Detection in Ultrasound Images

- Implemented an efficient k-means algorithm for detecting renal calculi or kidney stone from ultrasound images. The image was first preprocessed and compared time complexity with original k means and other segmentation algorithms like Fuzzy C means etc. The segmented image was generated with lesser time complexity and with similar accuracy when compared with above segmentation method.

Stock Market Prediction on financial dataset using Support Vector Machine

- Compare variants of support vector machines such as libSVM, nuSVM etc. on financial dataset using R and evaluate them on the basis of Accuracy, Precision, Recall and F-measure.