NF3 Fellowship Talk at RBCDSAI, IIT Madras

Statistics and Machine Learning as Key Technologies for Social Welfare

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Tanujit Chakraborty, Ph.D. (ISI Kolkata)
Postdoc Fellow at Centre for Data Sciences, IIIT Bangalore.
tanujitisi@gmail.com | www.ctanujit.org

August 10, 2021

---

*Some portion of the Title is a quotation from Ronald A Fisher (1940) and PC Mahalanobis (1965).

Outline

Background

Education:

- 2010 - 2013 : B.Sc. in Statistics from **Bidhannagar College**, Kolkata, India.
- 2014 - 2016 : M.S. from **Indian Statistical Institute** (ISI), Kolkata, India.
  Specialisation: Statistical Quality Control & Operations Research.
- 2016 – 2020 : Ph.D. from **Indian Statistical Institute** (ISI), Kolkata, India.
  Title of the Thesis: "Some Nonparametric Hybrid Predictive Models: Asymptotic Properties and Applications".
  Thesis Advisor: Prof. Ashis K Chakraborty, ISI, Kolkata.

Employment:

- Dec 2020 - Mar 2021 : Postdoctoral Fellow at **IIIT Delhi**, India.
- April 2021 - Present : Postdoctoral Fellow at **IIIT Bangalore**, India.
- June 2021 - Present : Visiting Assistant Professor at **XIM University Bhubaneswar**, India. Teaching **"Causal Inference in Statistics"** course (4 credits) for M.Tech in Data Science Analytics students.
- June 2021 - Present : Guest Lecturer at **Indian Institute of Foreign Trade (IIFT), New Delhi**. Teaching **"Data Analytics"** course (3 credits) for MBA (IB) students.

Consultancy and Recognition

Consultancy & Grant:

- Jan 2016 - June 2016 : Statistical Consultant (GB Mentor) at **ITC Limited**, Tribeni Tissue Division, West Bengal, India.
- Dec 2020 - Feb 2021 : Statistical Consultant (Part-time) at Business Intelligence Unit, **Bajaj Housing Finance Limited**, Pune, India.
- May 2021 - April 2024 : **Mphasis Research Grant** in Cognitive Computing. Role: Co-PI, Funding: 30 lakhs INR ($\sim$ €33K), Duration: 3 years.

Awards:

- B.G. Raghavendra Memorial Award for Best Paper from **Operational Research Society of India** (ORSI) in December 2017.
- Best Student Paper Award (Application Category) at the International Conference held at **IIM Ahmedabad** in December 2019.
- Best Paper Award Winner at ACM International Conference on Data Sciences and Management of Data (**ACM CODS-COMAD**) in January 2021.

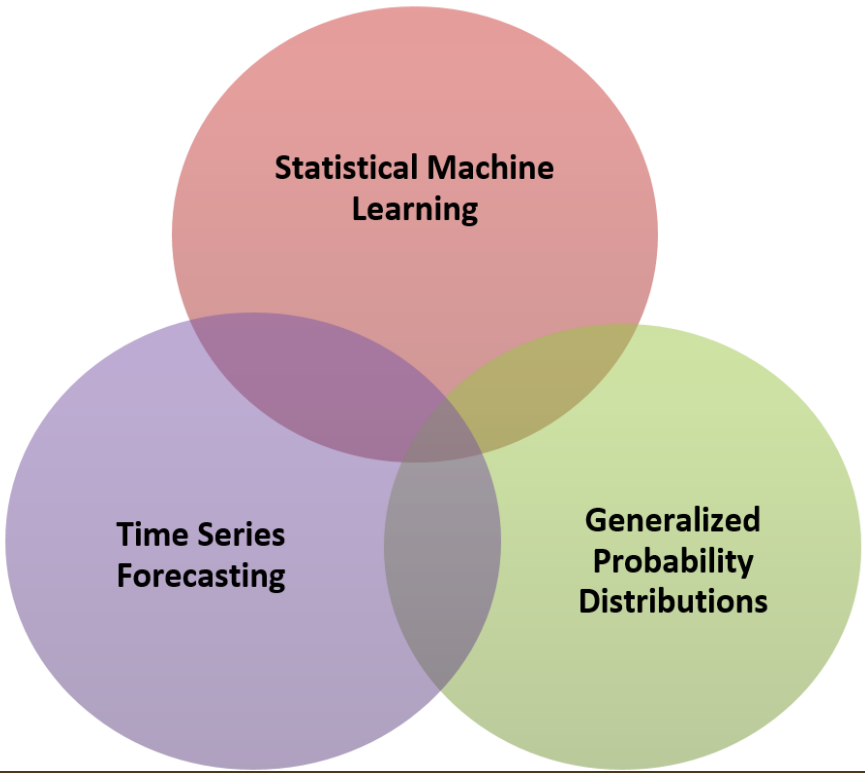
Areas of Research

"Statistics must have a clearly defined purpose, one aspect of which is scientific advance and the other, human welfare and national development" - PC Mahalanobis.

Motivations: *Primary motivation comes from the real-world data sets, with a variety of data types, such as business, process efficiency improvement, water quality control, and software defect prediction, among many others. My research works emphasize on the development of statistical methodologies that are scalable, robust, accurate, statistically sound, and easily interpretable.*

Data-driven research Problems from the areas of:

1. Business
2. Macroeconomics
3. Quality Control
4. Software Reliability
5. Epidemics
6. Engineering
7. Chaotic Systems
8. Network Science
9. Survival Analysis
10. Finance



Area-wise Contributions I

Statistical Machine Learning:

Building Hybrid Learning models for balanced classification, imbalanced classification and nonparametric regression. Some of these works are purely methodological in nature whereas rest are data-driven.

Statistical Learning:

Tanujit Chakraborty, Ashis Kumar Chakraborty, and C A Murthy (2019). "A nonparametric ensemble binary classifier and its statistical properties". In: *Statistics & Probability Letters* 149, pp. 16–23

Tanujit Chakraborty, Ashis Kumar Chakraborty, and Swarup Chattopadhyay (2019). "A novel distribution-free hybrid regression model for manufacturing process efficiency improvement". In: *Journal of Computational and Applied Mathematics* 362, pp. 130–142

Tanujit Chakraborty, Swarup Chattopadhyay, and Ashis Kumar Chakraborty (2020). "Radial basis neural tree model for improving waste recovery process in a paper industry". In: *Applied Stochastic Models in Business and Industry* 36.1, pp. 49–61

Tanujit Chakraborty, Gauri Kamat, and Ashis Kumar Chakraborty (2021+). "Bayesian Neural Tree Models for Nonparametric Regression". In: *Submitted for Publication.*

Tanujit Chakraborty (2021+a). "Near-optimal Sparse Neural Trees". In: *Submitted for Publication.*

Clustering:

Tanujit Chakraborty (2021+b). "Robust Clustering with Optimal Transport". In: *Submitted for Publication.*

Imbalanced Learning:

Tanujit Chakraborty and Ashis Kumar Chakraborty (2020b). "Superensemble classifier for improving predictions in imbalanced datasets". In: *Communications in Statistics: Case Studies, Data Analysis and Applications* 6.2, pp. 123–141

Tanujit Chakraborty and Ashis Kumar Chakraborty (2020a). "Hellinger Net : A Hybrid Imbalance Learning Model to Improve Software Defect Prediction". In: *IEEE Transactions on Reliability*

Area-wise Contributions II

Time Series Forecasting:

Building Forecasting methods for time series data sets from epidemiology, economics and dynamical systems.

Epidemiology:

Tanujit Chakraborty, Swarup Chattopadhyay, and Indrajit Ghosh (2019). "Forecasting dengue epidemics using a hybrid methodology". In: *Physica A: Statistical Mechanics and its Applications* 527, p. 121266

Tanujit Chakraborty and Indrajit Ghosh (2020). "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis". In: *Chaos, Solitons & Fractals* 135, p. 109850

Arinjita Bhattacharyya, Swarup Chattopadhyay, and **Tanujit Chakraborty** (2021). "Theta Autoregressive Neural Network: A Hybrid Time Series Model for Pandemic Forecasting". In: *IEEE International Joint Conference on Neural Networks (IJCNN)*

Arinjita Bhattacharyya, **Tanujit Chakraborty**, and Shesh N Rai (2021+). "Stochastic forecasting of COVID-19 daily new cases across countries with a novel hybrid time series model". In: *Submitted for Publication*

Tanujit Chakraborty, Sk Shahid Nadim, and Indrajit Ghosh (2021+). "Wavelet autoregressive neural network model for phenomenological forecasting of dengue incidence in the presence of exogenous variable". In: *Submitted for Publication*

Macroeconomics:

Tanujit Chakraborty, Ashis Kumar Chakraborty, Munmun Biswas, et al. (2021). "Unemployment Rate Forecasting: A Hybrid Approach". In: *Computational Economics* 57, pp. 183–201

Dynamical Systems:

Arnob Ray, **Tanujit Chakraborty**, and Dibakar Ghosh (2021+). "Optimized ensemble deep learning framework for scalable forecasting of dynamics consisting of extreme events". In: *Submitted for Publication.*

Area-wise Contributions III

Generalized Probability Distributions:

Understanding structural properties of real-world complex networks using generalized probability models.

Networks:

Swarup Chattopadhyay, **Tanujit Chakraborty**, and Kuntal Ghosh (2021b). "Uncovering patterns in heavy-tailed networks: A journey beyond scale-free". In: *8th ACM IKDD CODS and 26th COMAD*, pp. 136–144

Swarup Chattopadhyay, **Tanujit Chakraborty**, and Kuntal Ghosh (2021a). "Modified Lomax Model: A heavy-tailed distribution for fitting large-scale real-world complex networks". In: *Social Network Analysis and Mining*

Tanujit Chakraborty, Suchismita Das, and Swarup Chattopadhyay (2021). "A New Method for Generalizing Burr and Related Distributions". In: *Mathematica Slovaca*

Tanujit Chakraborty, Swarup Chattopadhyay, and Suchismita Das (2021+). "Modeling Heavy-tailed Networks: A Probabilistic Perspective". In: *Submitted for Publication*

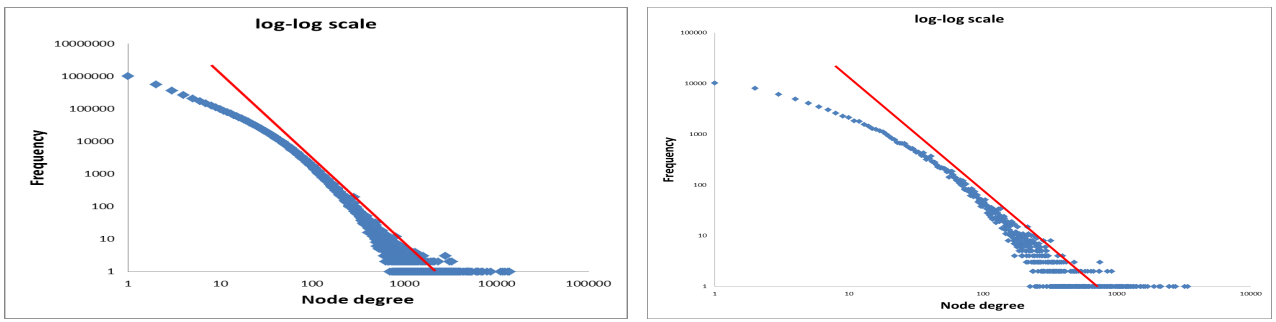


Figure: Twitter and LiveJournal networks displaying non-scale-free behavior.

A Distribution-free Hybrid Model for Manufacturing Process Efficiency Improvement

Related Publications:

1. Tanujit Chakraborty, Ashis Kumar Chakraborty, and C. A. Murthy. "A nonparametric ensemble binary classifier and its statistical properties", **Statistics & Probability Letters**, 149 (2019): 16-23.

2. Tanujit Chakraborty, Swarup Chattopadhyay, and Ashis Kumar Chakraborty. "Radial basis neural tree model for improving waste recovery process in a paper industry", **Applied Stochastic Models in Business and Industry**, 36 (2020): 49-61.

Motivation

- This work is motivated by a particular problem in a modern paper manufacturing industry, in which maximum efficiency of the process fiber-filler recovery equipment, also known as Krofta supracell, is desired.
- As a by-product of the paper manufacturing process, a lot of unwanted materials along with valuable fibers and fillers come out as waste materials.
- The job of an efficient Krofta supracell is to separate the unwanted materials from the valuable ones so that fibers and fillers can be reused in the manufacturing process.

Fig: Pictures of Krofta Supracell

Understanding the Process

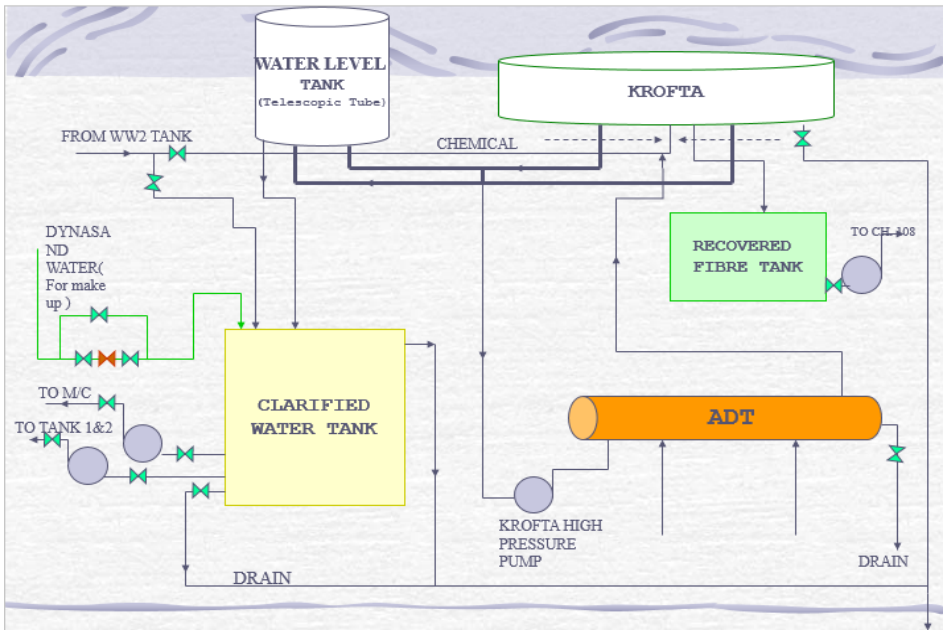


Fig: Process Flow Diagram of Krofta supracell

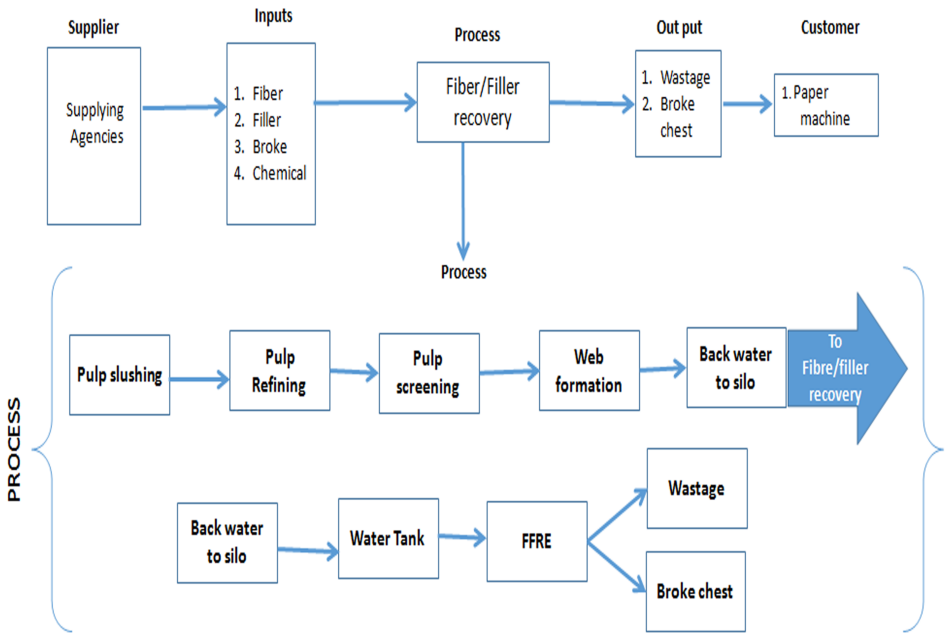


Fig: SIPOC diagram of fiber and filler recovery process

Aims and Objectives of Research

- The Krofta recovery percentage was around 75%. The paper manufacturing company wants to improve the recovery percentage to 90%.
- To identify the important process parameters affecting the Krofta efficiency, SIPOC and failure mode and effect analysis (FMEA) were performed with the help of process experts.
- **Goal**: We would like to come up with a model that can help the manufacturing process industry to achieve an efficiency level of about 90% for the Krofta supracell recovery percentage. **Benefits:** Monetary and environmental benefits.

Fig: Dissolved Air Floatation cum Sedimentation Process of Krofta supracell. Formula for Percentage

$$\textbf{Recovery} = \frac{(\textbf{Inlet PPM} - \textbf{Outlet PPM})}{\textbf{Inlet PPM}} \times 100$$

Process Data Set

- The data set collected for a year from the process on the following causal variables: Inlet Flow, Water Pressure (water inlet pressure to ADT), Air Pressure, Pressure of Air-Left, Pressure of Air-Right, Pressure of ADT-D Left, Pressure of ADT-D Right and Amount of chemical lubricants.
- The response variable (FFRE recovery percentage) lies between 20 to 100.
- This data set will be used for finding crucial process parameters and also finding a prediction model that can help the company for forecasting future recovery percentage of FFRE.

Table: Sample data set

Inlet Flow	**Water** Pressure	**Air** Pressure	**Air-Left**	**Air-Right**	**ADT-D Left**	**ADT-D Right**	**Amount of chemical**	**Recovery Percentage**
1448	6.4	5.8	1.0	2.1	3.2	4.0	2.0	96.80
1794	5.2	5.6	2.4	1.6	3.6	4.0	3.0	97.47
2995	6.0	6.0	1.5	4.5	4.0	4.8	4.0	28.87
1139	6.5	6.0	1.2	1.7	3.0	4.6	2.0	33.05
2899	6.2	5.7	2.0	1.2	3.1	4.0	2.0	97.91
1472	6.6	6.8	3.7	3.1	5.2	4.8	4.0	57.77
1703	6.2	6.0	2.9	1.0	3.0	4.2	2.0	26.94
1514	5.5	5.0	2.0	2.1	3.8	4.7	2.0	67.01
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Proposed Hybrid RBNT Model

- First, apply the RT algorithm to train and build a decision tree and record important features.
- Using important input variables obtained from RT along with an additional input variable (RT output), a RBFN model (with one hidden layer) is generated.
- The optimum number of neurons (k) in the hidden layer of the model to be chosen as $O\left(\sqrt{n/d_m log(n)}\right)$ [to be discussed], where n, d_m are number of training samples and number of input features in RBFN model, respectively.

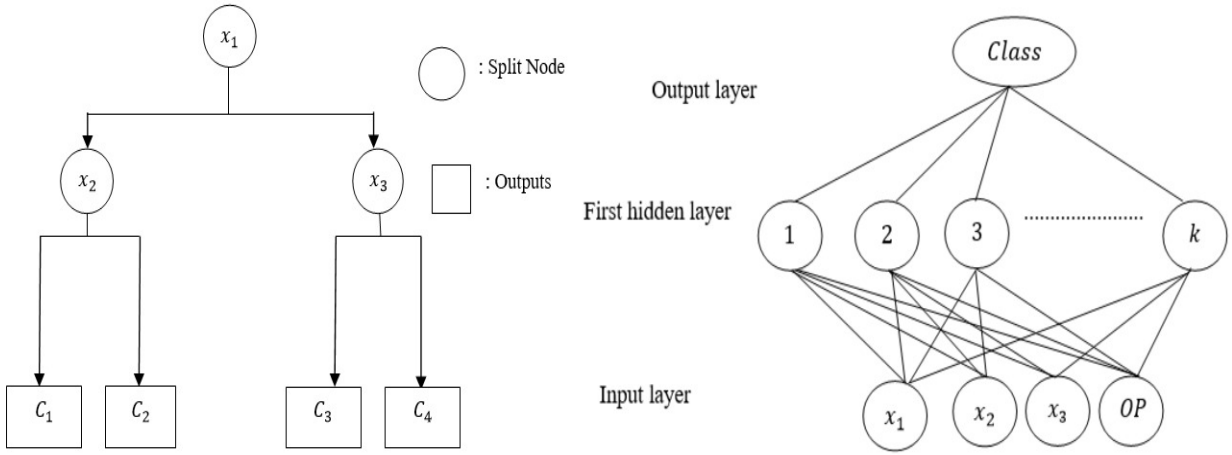


Figure: Graphical Presentation of the RBNT model

On Asymptotic Consistency

Theorem (Chakraborty et al., 2020, Applied Stochastic Models)

Suppose $(\underline{X}, \underline{Y})$ *be a random vector in* $\mathbb{R}^p \times [-K, K]$ *and* L_n *be the training set of n outcomes of* $(\underline{X}, \underline{Y})$. *Finally if for every n and* $w_i \in \tilde{\Omega}_n$, *the induced subset* $(L_n)_{w_i}$ *contains at least* k_n *of the vectors of* $X_1, X_2, \ldots, X_n$, *then empirically optimal regression trees strategy employing axis parallel splits are consistent when the size* k_n *of the tree grows as* $o\left(\frac{n}{log(n)}\right)$.

Theorem (Chakraborty et al., 2020, Applied Stochastic Models)

Consider a RBF network with Gaussian radial basis kernel having one hidden layer with k (> 1) *nodes. If* $k \to \infty$, $b \to \infty$ *and* $\frac{kb^4 log(kb^2)}{n} \to 0$ *as* $n \to \infty$, *then RBFN model is said to be universally consistent for all distribution of* $(\underline{Z}, \underline{Y})$.

Note: The above Theorems states that with certain restrictions imposed on the number k_n of terminal nodes and the parameters β being properly regulated as functions of n, the empirical L_2 risk-minimization provides consistency of the RBNT model.

On the choice of No. of Hidden Neurons

- RBFN is a family of ANNs, consists of only a single hidden layer and uses radial basis function as an activation function, unlike feed forward neural network. RBF network with one hidden layer having k nodes for a fixed Gaussian function is given by the equation:

$$f(z_i) = \sum_{j=1}^{k} w_j \, exp\left(- \frac{\| z_i - c_i \|^2}{2\sigma_i^2} \right) + w_0,$$

where $\sum_{j=0}^{k} |w_j| \leq b \ (> 0)$ and $c_1, c_2, ..., c_k \in \mathbb{R}^{d_m}$.

- For practical use, if the data set is limited, the recommendation is to use $k = \left(\sqrt{n/d_m log(n)}\right)$ for achieving utmost accuracy of the propose model.

Theorem (Chakraborty et al., 2019, Statistics & Probability Letters)

For any fixed d_m and training sequence ξ_n, let $Y \in [-K, K]$, and $m, f \in F_{n,k}$, if the neural network estimate m_n satisfies the above-mentioned regularity conditions of strong universal consistency and f satisfying $\int_{S_r} f^2(z)\mu(dz) < \infty$, where S_r is a ball with radius r centered at 0, then the optimal choice of k is $O\left(\sqrt{\frac{n}{d_m log(n)}}\right)$.

Importance of RT output in neural net

- RT output also plays an important role in further modeling. It actually improves the performance of the model at a significant rate (can be shown using experimental results).
- We can use one hidden layer in Neural Network model due to the incorporation of RT output as an input information in ANN.
- RT predicted results provide some direction for the second stage modelling using ANN.
- Tree output estimates are probabilistic estimates, not from a direct mathematical or parametric model, thus direct correlations with variables can't be estimated.
- It should be noted that one-hidden layer neural networks yield strong universal consistency and there is little theoretical gain in considering two or more hidden layered neural networks (Devroye, IEEE IT, 2013).
- To see the importance of RT given predicted results as a relevant feature, we introduced a non-linear measure of correlation between any feature and the actual values, namely C-correlation (Yu and Liu, 2004, JMLR), shown in Chakraborty et al., 2019, Statistics & Probability Letters.

Experimental Evaluation on Process data

Popularly used performance metric are:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|; RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}; MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right|;$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right]; AdjR^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n - d_m - 1}\right];$$

where, $y_i, \overline{y}, \widehat{y}_i$ denote the actual value, average value and predicted value of the dependent variable, respectively for the i^{th} instant. Here n and d_m denote the number of data points and independent variables used, respectively.

Table: Quantitative measure of performance for different regression models. Results are based on 10 fold cross validations. Mean values of the respective measures are reported with standard deviation within the bracket.

Models	**MAE**	**RMSE**	**MAPE**	R^2	**Adj(R^2)**
RT	11.691 (0.45)	16.927 (0.89)	29.010 (1.02)	59.028 (3.25)	55.304 (1.95)
ANN	12.334 (0.25)	17.073 (0.56)	27.564 (1.85)	58.310 (2.98)	54.529 (2.08)
SVR	12.460 (0.28)	20.362 (1.23)	40.010 (1.81)	40.174 (2.05)	35.325 (2.64)
BART	12.892 (0.59)	16.010 (1.25)	30.038 (1.95)	59.380 (2.50)	56.458 (1.75)
RBFN	13.926 (2.50)	18.757 (3.25)	32.48 (3.45)	49.689 (5.45)	46.335 (3.95)
Tsai Neural tree	10.895 (0.78)	16.012 (0.50)	24.021 (1.85)	65.120 (2.89)	62.946 (1.78)
Proposed Model	**9.226** (0.35)	**14.331** (0.82)	**20.187** (1.45)	**70.632** (2.00)	**68.675** (2.13)

Experiments on UCI Data

Data Sets: The proposed model is evaluated using six publicly available from UCI Machine Learning repository (`https://archive.ics.uci.edu/ml/datasets.html`). These regression data sets have limited number of observations.

Table: Data set characteristics: number of samples and number of features, after removing observations with missing information or nonnumerical input features.

Sl. No.	Data	Number of samples	Number of features
1	Auto MPG	398	7
2	Concrete	1030	8
3	Forest Fires	517	10
4	Housing	506	13
5	Wisconsin	194	32

Table: Average RMSE results for each of the models across the different data sets

Data	RT	ANN	SVR	BART	RBFN	Neural Tree	Our Model
Auto MPG	3.950	4.260	5.720	3.220	4.595	3.300	**3.215**
Concrete	8.700	10.180	11.588	**5.540**	10.210	7.420	7.063
Forest Fires	75.138	90.702	91.985	65.890	82.804	**62.478**	64.411
Housing	4.980	9.054	12.520	3.978	7.871	4.590	**3.077**
Wisconsin	41.059	34.710	41.220	32.054	38.495	40.700	**23.659**

Merits and Deliverables

- Suitable for Feature Selection cum Prediction Problems with limited data sets;
- Simple and Easily interpretable, “white-box-like" model, fast in implementation;
- Based on the model, we further created an experimental design to obtain the optimal level of the tuning parameters;
- Final recommendations based on the results of the design of experiments (Taguchi method) were implemented in the process to monitor the Krofta efficiency.
- The company could achieve 85% recovery percentage at the end of DMAIC stage and making huge monetary benefits.

This work received the Best Case Study Paper Award at 52nd International Conference on Operations Research organized by ORSI India & IIM Ahmedabad.

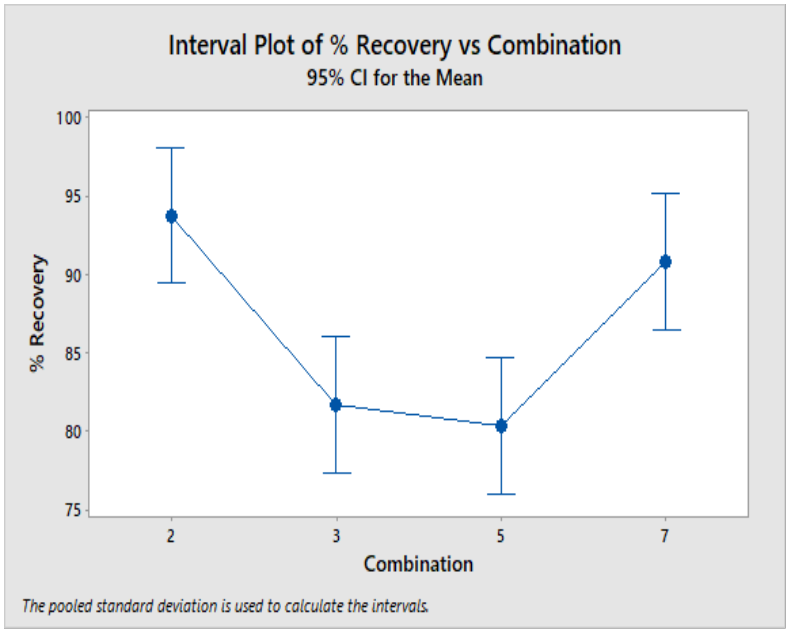


Fig: Interval plot of DOE Combinations

Research Proposal on Socially Relevant AI

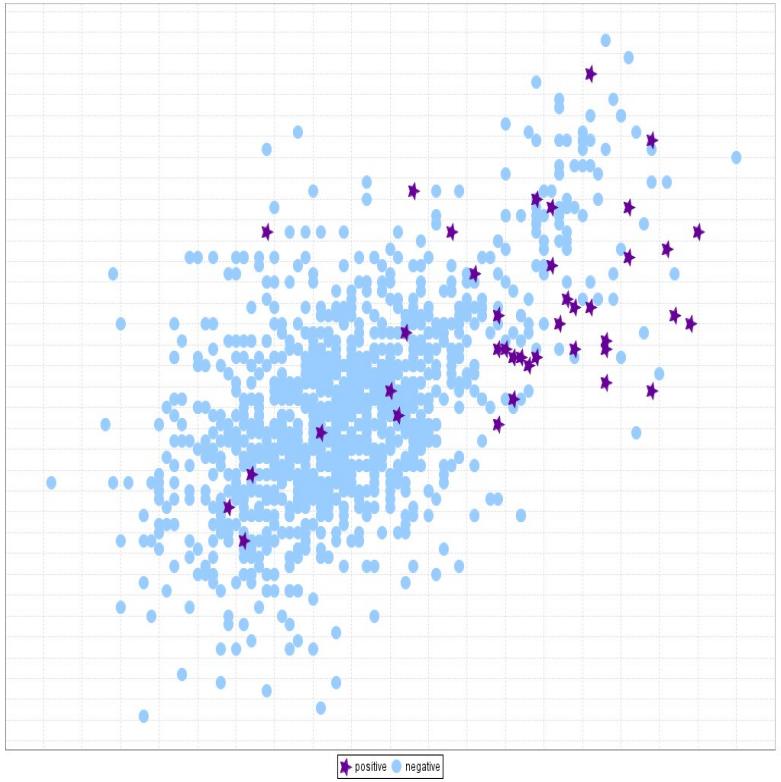
Title: "Generalized Wasserstein Deep Neural Forest - An Imbalanced Learner for Fraud Detection in Social Media"

Problem Statement:

- Online social media platforms sensitize users to involve in different activities. These platforms have shown to play contributory role in several decision making processes such as Presidential election, product purchase and movie selection.
- Recent world events provide ample testimony that fraud activity in social media is a serious challenge to unbiased truth discovery and opinion formation. Fraudsters often take part in online discussions (Facebook, Twitter, Reddit) and manipulate other participants, write fake reviews to the e-commerce products, YouTube videos, etc. to promote/demote online reputation.
- The major limitation of state-of-the-art fraud detection methods is that the collected dataset often tends to be imbalanced, predominantly over-represented by the genuine activities, in order to synthesize the real-world behavior.
- There is also very limited information available to distinguish dynamic fraud from genuine customer behavior in such an extremely sparse and imbalanced data environment, which makes the instant and effective detection more challenging.

Imbalanced Classification Problem

- Real-world data sets are usually skewed, in that many cases belong a larger class and fewer cases belong to a smaller yet usually more exciting class
- For example, consider a binary classification problem with the class distribution of 90 : 10. In this case, a straightforward method of guessing all instances to be positive class would achieve an accuracy of 90%.
- Learning from an imbalanced data set presents a tricky problem in which traditional learning algorithms perform poorly.
- Traditional classifiers usually aim to optimize the overall accuracy without considering the relative distribution of each class.



Crisis and Motivation

- Previous studies handle the issue of class imbalance via three kinds of approaches: (i) data-level approaches (SMOTE, ENN); (ii) Cost-sensitive learning (VCB-SVM, ISDA); (iii) Algorithmic approaches (HDDT, HDRF, CCPDT).
- Despite the progress the aforementioned methods have made, there are still several challenges: (a) Drawbacks of data level approaches and cost-sensitive learning; (b) Model explainability and interpretability; (c) High dimensionality; (d) Classifying extreme samples and unseen categories.
- Deep Learning methods have boosted the capacity of machine learning algorithms and are now being used for non-trivial applications in various applied domains. However, the real-life data sets are extremely imbalanced which severely hampers the neural network's capabilities, reducing the robustness and trust.
- Deep learning methods, like ensemble deep learning model (Yin et al., CMPB-2017), knowledge-shot learning (Chou et al., Neurocomputing-2020) and dynamic curriculum learning (Wang et al., ICCV-2019) for imbalanced data classification incur high time complexity than traditional neural networks and can classify unseen classes only if the knowledge vector of these classes is artificially given.

Moving Forward

- Popularly-used supervised fraud detection methods such as decision forests and cost-sensitive neural networks are either biased towards certain features and dominant class, or incur serious overfitting problem since the size of the datasets is limited.
- The dataset contains small disjuncts when some concepts (disregard their classes) are represented within small clusters. This situation increases the complexity in the search for quality solutions.
- In fraud detection data, the presence of small disjuncts, noisy and imbalance class distribution may lead to enormous challenges and opportunities to the machine learning research community.
- The proposed solution needs to be effective and instant as it is very difficult to recover the damage if a fraud is uncovered during the detection phase.
- In the proposed proposal, we address the curse of imbalanced datasets and the deficiencies of the past literature in the domain of fraud detection, by designing an Generalized Wasserstein Deep Neural Forest (GWDNF) model. This project aims to design an interpretable, scalable deep model, namely GWDNF, that can handle highly imbalanced data sets arising in other applied domains as well.

Related Works

- Perceptron Trees (Utgoff, AAAI, 1988).
- Entropy Nets (Sethi, Proceeding of IEEE,1990).
- Neural trees (Sirat & Nadal, Network, 1990).
- Sparse Perceptron Trees (Jackson, Craven, NIPS, 1996).
- SVM Tree Model (Bennett et al., NIPS, 1998)
- Hybrid DT-ANN Model (Jerez-Aragones et al., 2003, AI in Medicine)
- Flexible Neural Tree (Chen et al., Neurocomputing, 2006)
- Hybrid DT-SVM Model (Sugumaran et al,, Mechanical Systems and Signal Processing, 2007).
- Hybrid CNN–SVM Classifier (Niu et al., PR, 2012).
- Neural Decision Forests (Bulo, Kontschieder, CVPR, 2014).
- Deep Neural Decision Forests (Kontschieder, ICCV, 2015).
- Soft Decision Tree (Frosst, Hinton, Google AI, 2017).
- Deep Neural Decision Trees (Humbird et al., IEEE TNNLS, 2018).
- Adaptive Neural Trees (Tanno et al. ICML, 2019).
- Hellinger Net (Chakraborty et al. IEEE TR, 2020).

Hellinger Distance

Let (Θ, λ) denote a measurable space. Let us suppose that P and Q be two continuous distributions with respect to the parameter λ having the densities p and q in a continuous space Ω, respectively. Define HD as follows:

$$d_H(P,Q) = \sqrt{\int_\Omega (\sqrt{p} - \sqrt{q})^2 d\lambda} = \sqrt{2\left(1 - \int_\Omega \sqrt{pq} d\lambda\right)}$$

where $\int_\Omega \sqrt{pq} d\lambda$ is the Hellinger integral. It is noted that HD doesn't depend on the choice of the parameter λ.

For the application of HD as a decision tree criterion, the final formulation can be written as follows:

$$HD = d_H(X_+, X_-) = \sqrt{\sum_{j=1}^{k} \left(\sqrt{\frac{|X_{+j}|}{|X_+|}} - \sqrt{\frac{|X_{-j}|}{|X_-|}}\right)^2}, \tag{1}$$

where $|X_+|$ indicates the number of examples that belong to the majority class in training set and $|X_{+j}|$ is the subset of training set with the majority class and the value j for the feature X. The bigger the value of HD, the better is the discrimination between the features (Hellinger Distance Decision Tree, Chawla et al. 2008, ECML).

Hellinger Net : Basic idea

- Hellinger Net is composed of three basic steps:

  (a) Converting a DT into rules (HD is used as criterion);
  (b) Constructing a two hidden layered NN from the rules;
  (c) Training the MLP using gradient descent backpropagation (Rumelhart, Hinton (1988).

- In decision trees, the overfitting occurs when the size of the tree is too large compared to the number of training data.

- Instead of using pruning methods (removing child nodes), HN employs a backpropagation NN to give weights to nodes according to their significance.

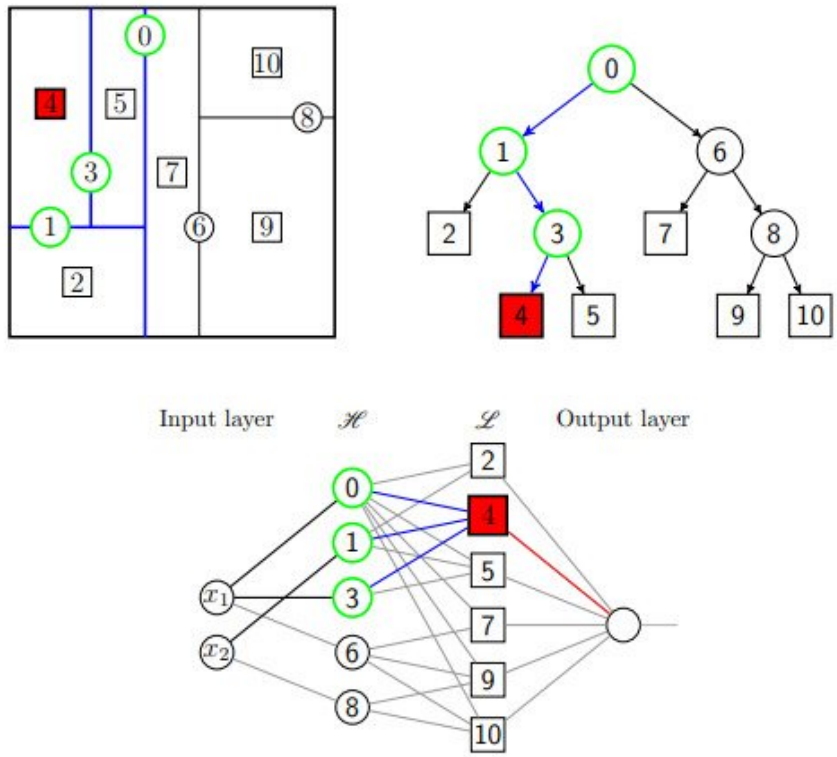


Fig: Graphical Representation of Hellinger Nets

Scope of Improvements

- The arena of research in "learning from imbalanced data" continues to grow, largely driven by challenging problems including land cover classification, fraud detection, face recognition, spam and anomaly detection, medical diagnosis, etc
- The overarching question is “how to push the boundaries of prediction on the underrepresented or minority classes (e.g., fraud activities in social media) while managing the trade-off with false positives?”
- The usefulness and success of Random Forests and Deep Learning methods are evident. Can they be combined together to create a Deep Forest model that can deal with data imbalance problem in social media?
- The proposed solution needs to be effective and instant as it is very difficult to recover the damage if a fraud is uncovered during the detection phase.
- Use of Wasserstein Distance is of much use in the Machine Learning community for the last few decades. Some modification to the Wasserstein distance can be done and incorporated in the DT, RF, and Hellinger net model. This may improve the existing HDDT, HDRF and Hellinger Net models for imbalanced pattern classification.

Optimal Transport and Wasserstein Distance

- Over the last two centuries, the theory of optimal transport has gained a lot of attention as it offers a strong foundation for addressing issues that embed statistical constraints (Villani, 2008 Book).
- The Wasserstein distance (WD) which arises from the idea of optimal transport, is being used more and more in Statistics and Machine Learning literature (e.g., Wasserstein GAN).
- The use of WD (and the optimal transport problem) is ubiquitous in mathematics, notably in fluid mechanics, partial differential equations, optimization, probability theory, and statistics.
- Intuitively, WD is a metric between probability distributions. WD can be used to derive weak convergence and convergence of moments, and can be easily bounded (Bhat et al., NIPS 2019).
- WD is well-adapted to quantify a natural notion of perturbation of a probability distribution. It seamlessly incorporates the geometry of the domain of the distributions in question, thus being useful for contrasting complex objects.
- Due to the success of WD in machine learning problems, here we attempt to propose a generalized version of WD and show its uses in various learning problems.

Generalized Wasserstein Distance

Given two CDFs F_1 and F_2 on $\mathbb{R}$, let F denote the set of all joint distribution on $\mathbb{R}^2$ having F_1 and F_2 as marginals.

Definition

Given two CDFs F_1 and F_2 on $\mathbb{R}$, the Wasserstein distance between them is defined by

$$W_1(F_1, F_2) = \left[\inf_{F \in \Gamma(F_1, F_2)} \int_{\mathbb{R}^2} |x - y| dF(x, y) \right] \tag{2}$$

Our proposed Generalized Wasserstein metric (Chakraborty, 2021+) is a linear combination of the Wasserstein distance metric for discrete probability measures (d_E) and the absolute value of the differences in norms parameterized by a real number μ, and is defined as

$$D_\mu(x, y) = d_E\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right) + \mu \left| \|x\| - \|y\| \right|$$

where, $d_E\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right) = \sum_{i=1}^{p} \left| \left(\frac{x_i}{\|x\|} - \frac{y_i}{\|y\|} \right) \right|$. By using the actual norm information in the WD; we hope that the proposed GWD can deal data imbalance and outliers in the data sets better than Hellinger distance and others.

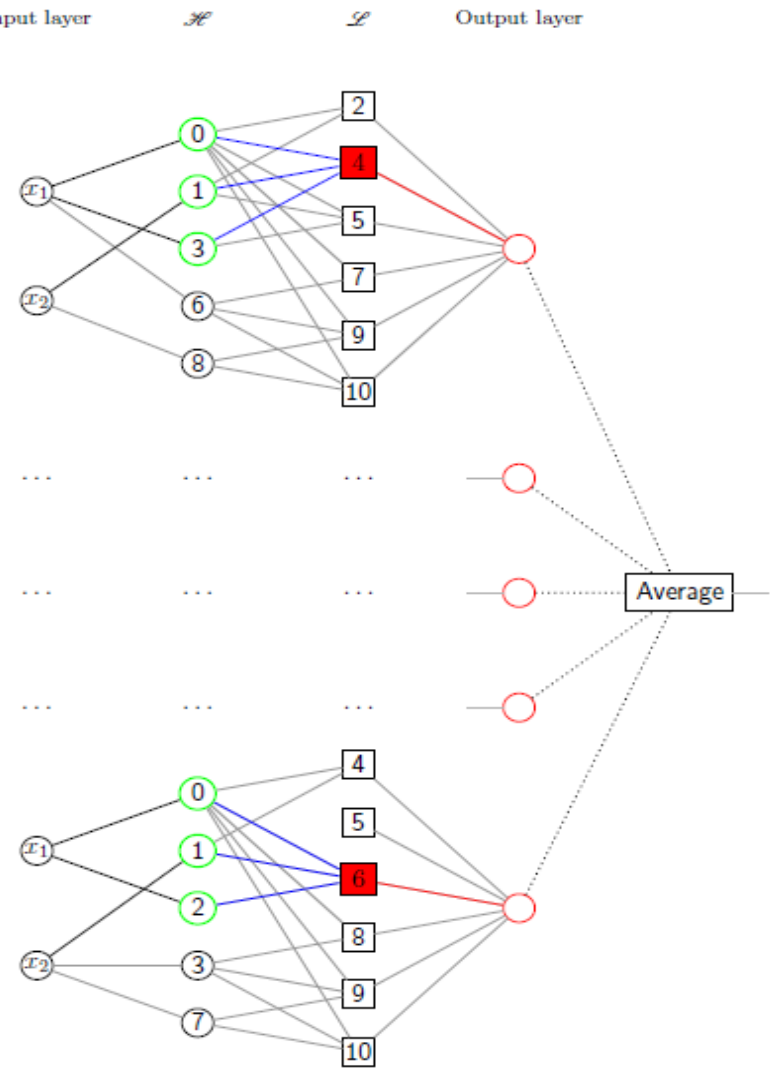
Generalized Wasserstein Distance (contd.)

- It is straightforward to see that the proposed generalized Wasserstein distance is a metric. We have shown its usage for robust clustering problems.
- When x and y with well distinct norm are far away, we get large μ; and for small μ their renormalized versions only matters.
- By using a renormalized version of x and y and adding the norm information with weight parameters, we shall be able to make the WD metric skew-insensitive and useful to handle noisy data in imbalanced social media problems.
- Generalized Wasserstein metric will be able to handle the highly imbalanced data problem within the Deep Forest framework.

Deep Forest - Individual Training

Finally, the project aims to make the proposed GWNDF model:

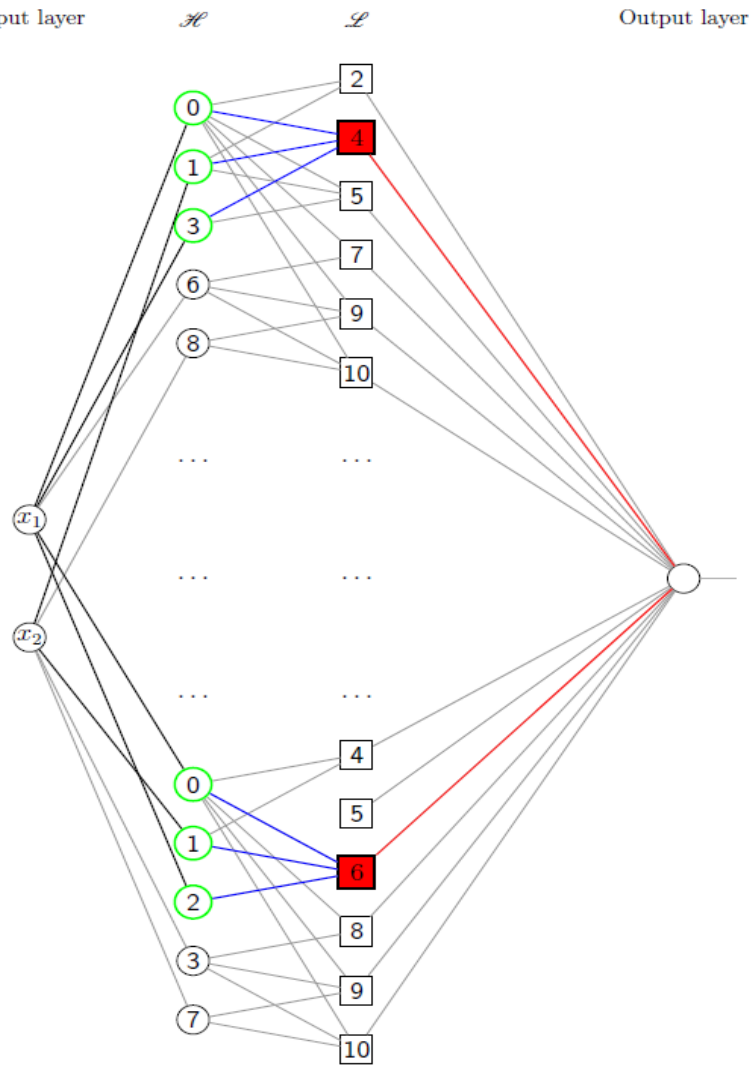
- scalable (the size of the data does not pose a problem),
- robust (work well in a wide variety of problems in the presence of noisy samples),
- accurate (achieve higher predictive accuracy),
- statistically sound (have desired asymptotic properties),
- easily interpretable for its effective implementation in land cover, aerial imagery and physiological data classification.



Deep Forest - Joint Training

We strongly desire that, whilst achieving competitive performance on imbalanced datasets in imbalanced data classification, GWNDF would benefit from

- lightweight inference via conditional computation (sparse connected networks),
- hierarchical separation of features useful to the imbalanced learning task with generalized WD metric as tree splitting criteria,
- a mechanism to adapt the architecture to the size and complexity of the training dataset,



GWNDF : Deliverables

- Generalized Wasserstein Deep Forest is a form of random forests enhanced with deep learned representations.
- Many existing tree-structured models are instantiations of the proposed GWNDF model.
- The outcome of this work will be a suite of novel, scalable, and interpretable deep learner that would solve the imbalanced problem in social media and others.
- We shall further investigate statistical consistency and rate of convergence for theoretical robustness of the proposed Wasserstein Deep Forest.
- We will try to integrate GWNDF with Bayesian Deep Learning to solve adversarial classification problems.
- Apart from the theoretical and computational development of the GWNDF model and its implementation on real-world datasets, we aim to develop an implementation tool (a Toolbox in Python) for public use.


Thank You