Duke-NUS Medical School

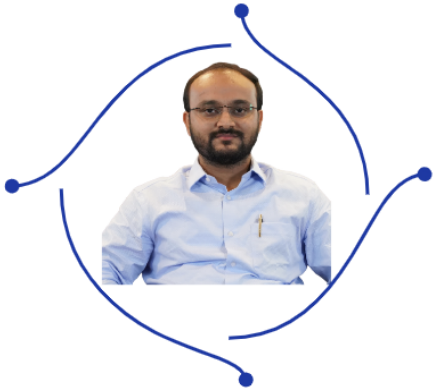
# Hybrid & Ensemble Machine Learning

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Tanujit Chakraborty
Assistant Professor of Statistics at Sorbonne University
ctanujit@gmail.com
www.ctanujit.org

July 12, 2023

# My Profile



## Education

- 2021-2024 — Co-PI MPhasis Research Fellow, IIIT Bangalore
- 2021 — PostDoc IIIT Delhi, India
- 2020 — Ph.D. in Statistical Machine Learning ISI Kolkata, India
- 2016 — MS in Applied Statistics (SQC&OR) ISI Kolkata, India
- 2013 — BS in Statistics Bidhannagar College

## Publications

- 2018 - Now — Published more than 25 research articles in International journals & conference.
- Three Best paper awards from ORSI, India, ACM IKDD, and IIM Ahmedabad.

## Work Experiences

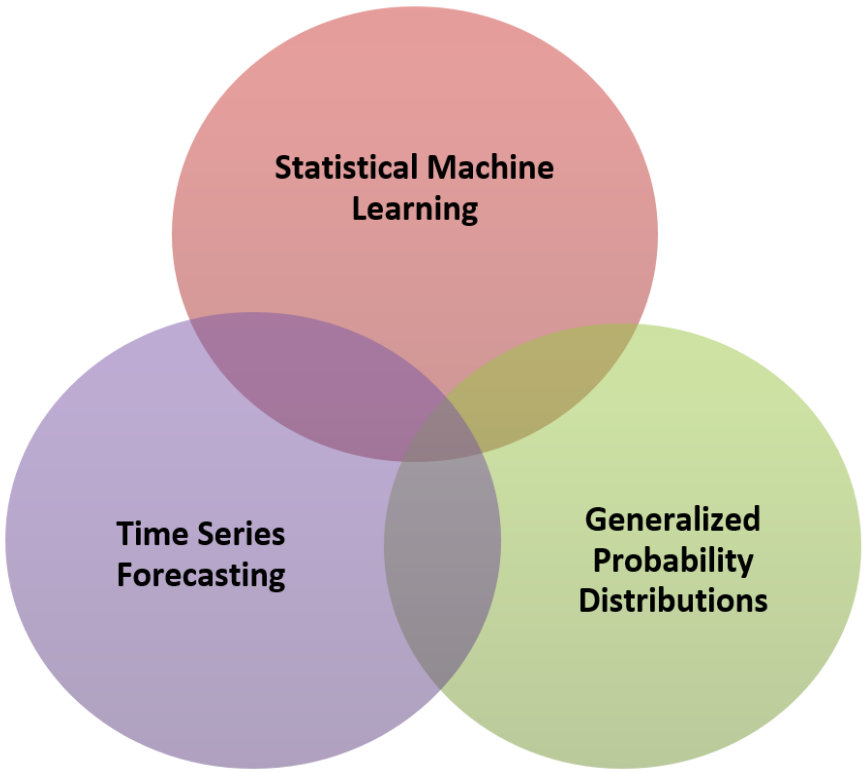
- 2021-Now — Asst. Professor Sorbonne
- 2023 — Consultant - PharmaACE (PharmaPRO)
- 2022-Now — Adj. Professor Woxsen University
- 2021 — Visiting Professor IIFT & XIMB
- 2021 — Consultant - Bajaj Finserv (Imbalanced)
- 2016 — Statistical consultant ITC Ltd (Six-sigma)

# Areas of Research

*Motivations:* *Primary motivation comes from the real-world data sets, with a variety of data types, such as business, process efficiency improvement, epidemics, quality control, and software defect prediction, among many others. My research works emphasize the developments 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



## Hybrid & Ensemble Models in Stat. & ML

# Need for Hybridization of Models

## Representation issue:

- In many learning tasks, the true unknown hypothesis could not be represented by any hypothesis in the hypothesis space.
- By hybridization, it may be possible to expand the space of representable functions.
- Thus the learning algorithm may be able to form a more accurate approximation to the true unknown hypothesis.

## Computational issue:

- Many learning algorithms perform some kind of local search that may get stuck in local optima.
- Even if there are enough training data, it may still be challenging to find the best hypothesis.
- By combining two or more models, the risk of choosing the wrong local minimum can be reduced.

## Statistical issue:

- It is often the case that the model space is too large to explore for limited training data, and that there may be several different models giving the same accuracy on the training data.
- The risk of choosing the wrong model can be reduced by combining two models, like CART and ANN.

# Ensemble Models

## Problem:

- Single models have the drawbacks of sticking to a local minimum or over-fitting the data set, etc.

## Performance:

- Ensemble learning models, where predictions of multiple homogeneous weak models are combined together to build the final model, have been theoretically and empirically shown to provide significantly better performance than single weak learners, especially while dealing with high dimensional, complex regression and classification problems

  Examples: Bagging, Boosting, and Voting Method.

## Caution:

- But ensembles don't always improve the accuracy of the model but tend to increase the error of each individual base classifier.

# Hybrid Models

## Overview:

- Hybrid Methods - takes a set of heterogeneous learners and combines them using new learning techniques.
- It overcomes the limitations of single models and reduces individual variance & bias, thus improving the performance of the model.

## Deliverables:

- Adaptive hybrid systems have become essential in computational intelligence and soft computing, as being able to deal with evolving components, non-stationary environments, and concept drift.
- The integration of the basic technologies into hybrid machine learning solutions facilitates more intelligent search and reasoning methods that match various domain knowledge with empirical data to solve complex problems.

## Caution

- To build a good ensemble classifier the base classifier needs to be simple, as accurate as possible, and distinct from the other classifier used.

# Proposed Hybrid Models (selected)

## Supervised learning:

- Hybrid CT-ANN (Stat. Prob. Let., 2019)
  - Medical data
- Hellinger Net (IEEE TR, 2020)
  - Software defect data
- Radial basis Neural Tree (ASMBI, 2020)
  - Waste recovery
- Bayesian Neural Tree (Aust. N. Z. J. Stat., 2023)
  - UCI dataset

## Epidemic forecasting:

- Hybrid ARIMA-ARNN (Physica A, 2019)
  - Dengue data
- Risk assessment of Covid-19 (ISA) (Chaos, Solitons & Fractals, 2020)
- Theta ARNN (Nonlinear Dyn., 2022)
  - Covid-19 data
- EWNet (Neural Networks, 2023)
  - Epidemic data

## Geophysics & dynamical systems:

- Optimized Ensemble DL (Chaos, 2021)
  - ENSO data
- Knowledge-based DL (ICMLA, 2022)
  - Climate data
- RidgeGAN (Submitted)
  - Urban planning
- ForePINN (In progress)
  - Geoscience data

# Proposed Method - 1

## HYBRID MODELS FOR IMBALANCED LEARNING

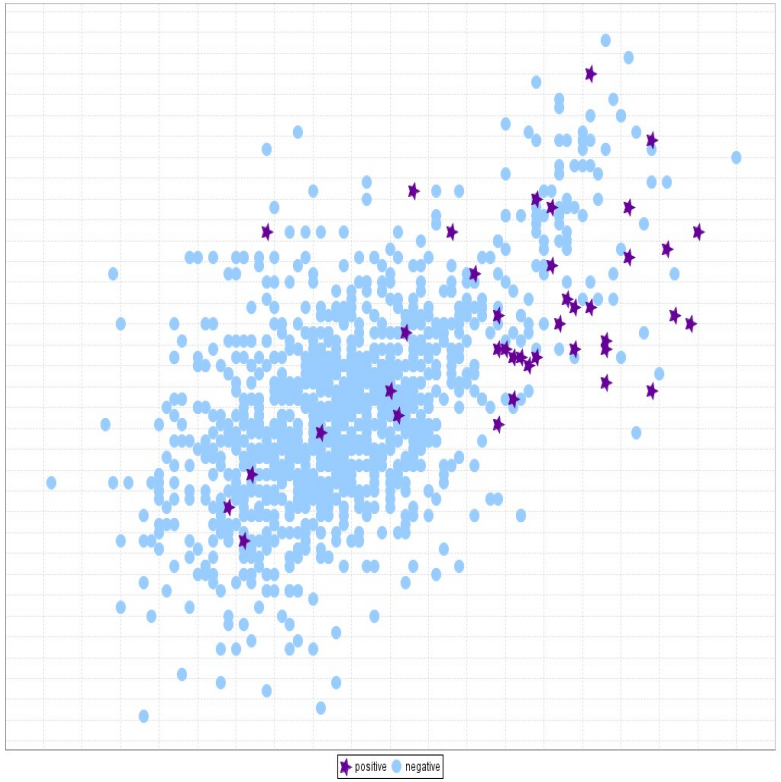
### Related Publications:

1. Tanujit Chakraborty and Ashis Kumar Chakraborty. "Hellinger net: A hybrid imbalance learning model to improve software defect prediction", **IEEE Transactions on Reliability**, 70 (2020): 481-494.

2. Tanujit Chakraborty and Ashis Kumar Chakraborty. "Superensemble classifier for improving predictions in imbalanced datasets", **Communications in Statistics: Case Studies, Data Analysis, and Applications**, 6 (2020): 123-141.

# Imbalanced Classification Problem

- Real-world data sets are usually skewed, in that many cases belong to 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.

# Hellinger Distance

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

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

where $|X_+|$ indicates the number of examples that belong to the majority class in training set and $|X_{+j}|$ is the subset of the 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).

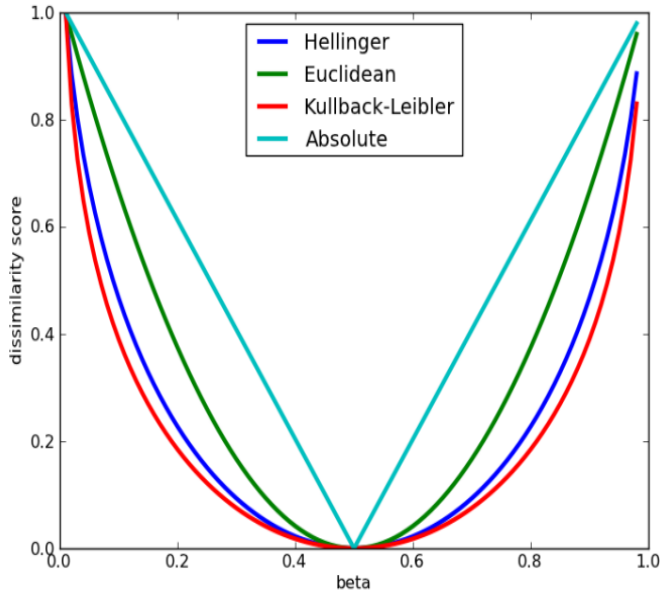



**Fig: An illustration of the behavior of the Squared Hellinger distance, Euclidean distance, absolute distance, and the symmetric Kullback-Leibler divergence.**

# Hellinger Net : Basic idea

- Hellinger Net is composed of three basic steps:

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

- In decision trees, 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.

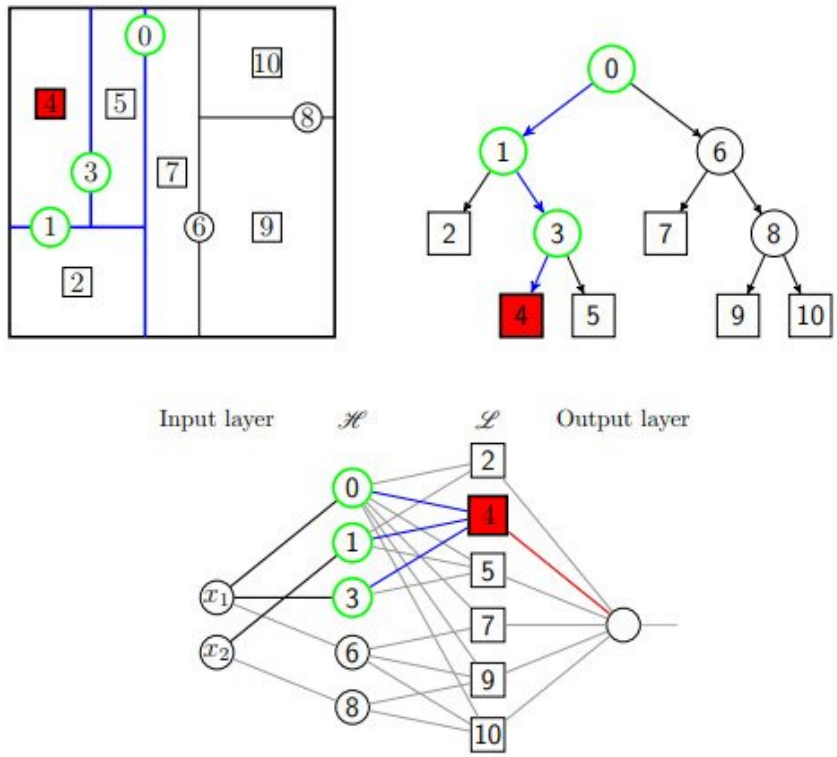



**Fig: Graphical Representation of Hellinger Nets**

# Individual vs Joint Training

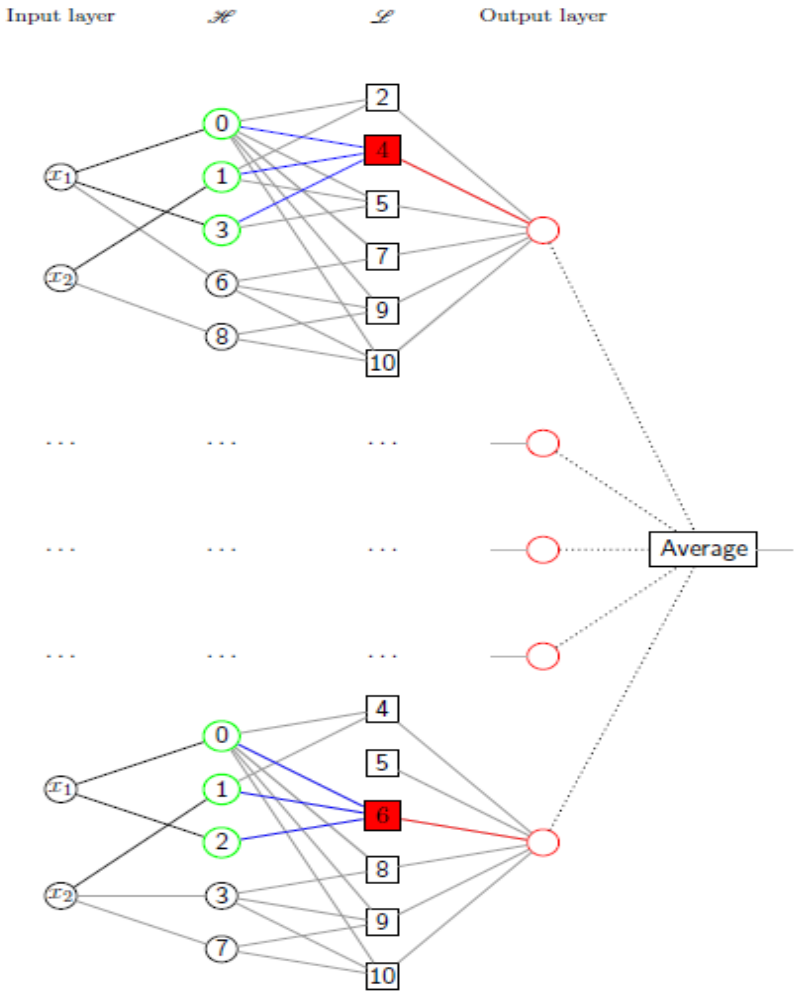



Fig: Individual training in Hellinger Net

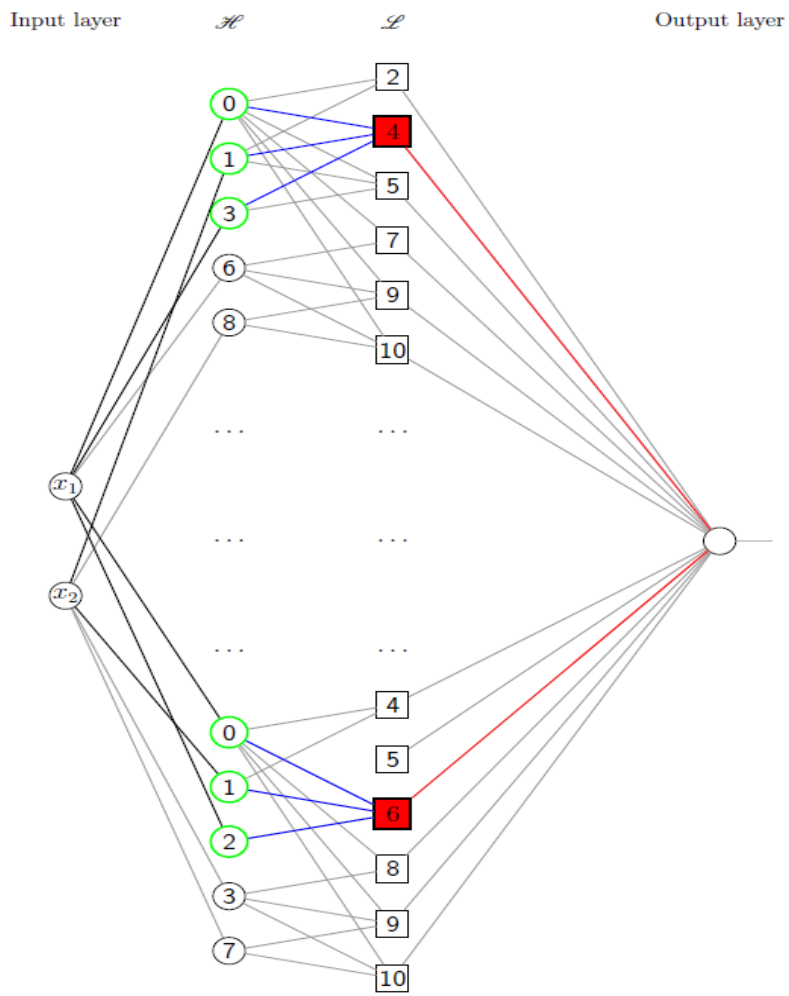
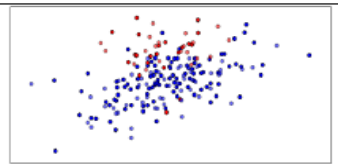
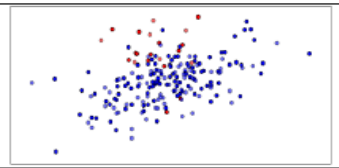
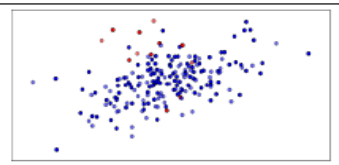
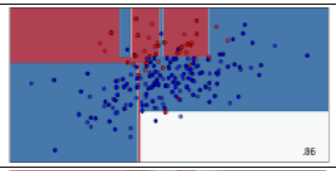
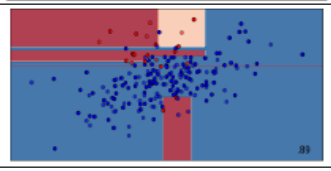
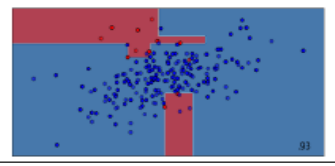
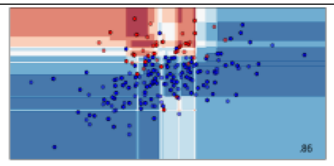
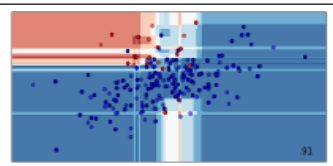
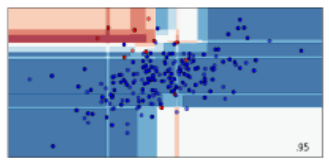
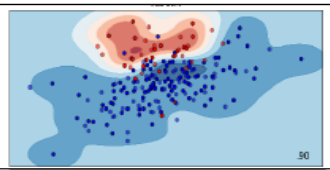
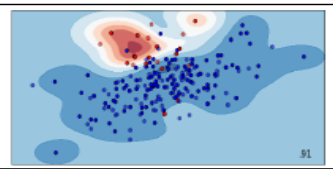
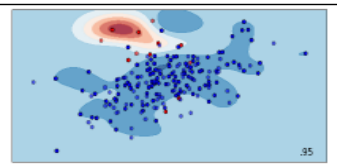
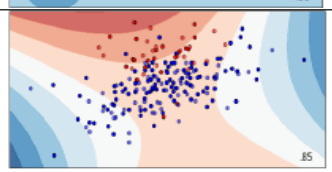
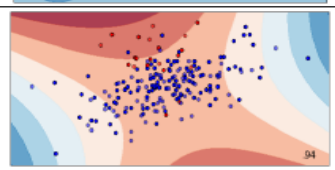
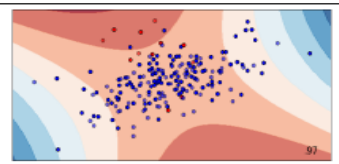
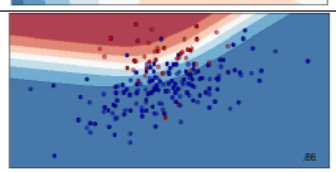
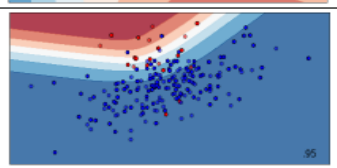
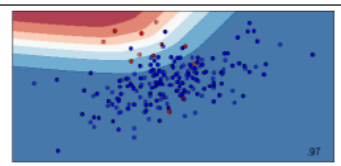



Fig: Joint training in Hellinger Net

# Simulation Study

*A comparison of several imbalanced classifiers on synthetic data sets. The plots show training points in solid colors and testing points semi-transparent.*

| Synthetic data | Imbalance = 20% rate | Imbalance = 10% rate | Imbalance = 5% rate |
|---|---|---|---|
| Input Data |  |  |  |
| HDDT |  |  |  |
| HDRF |  |  |  |
| VCB-SVM |  |  |  |
| ISDA |  |  |  |
| Hellinger net |  |  |  |

# Proposed Method - 2

## An Ensemble Wavelet Neural Network approach for epidemic forecasting

### Related Publications:

1. Madhurima Panja, Tanujit Chakraborty, Uttam Kumar, and Nan Liu "Epicasting: An Ensemble Wavelet Neural Network for forecasting epidemics", **Neural Networks**, (2023).

2. Madhurima Panja, Tanujit Chakraborty, Sk Shahid Nadim, Indrajit Ghosh, Uttam Kumar, and Nan Liu. "An ensemble neural network approach to forecast Dengue outbreak based on climatic condition", **Chaos, Solitons & Fractals**, 167 (2023): 113-124.

# Epicasting

## "Epicasting: An Ensemble Wavelet Neural Network for Forecasting Epidemics"

- Epidemic waves of infectious diseases are one of the top contributors to human illness and deaths worldwide.
- The unavailability of drugs and ready-to-use vaccines to prevent most of these epidemics makes the situation worse.
- To design countermeasures for such epidemic outbreaks and reduce the potential impact, policymakers, and health officials rely on early warning systems.
- Generating accurate and reliable forecasts for these epidemic outbreaks (epicasting) is of primary importance to assist stakeholders in developing situational awareness and formulating control response strategies well ahead of time to handle large-scale emergencies.

# Crisis and Motivation

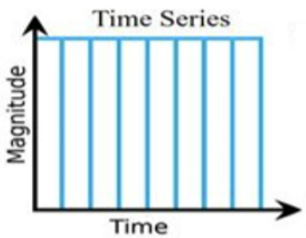
Epidemiological forecasting is a centuries-old field and it encapsulates a wide range of techniques namely:

- Deterministic methods (e.g. SIR model) are suitable for studying the changes in the characteristics of the population owing to infectious disease outbreaks however, they lack the ability to generate reliable forecasts.
- Phenomenological methods encompass a wide variety of statistical, machine learning, and deep learning frameworks. Although this methodology tries to learn temporal disease dynamics in a purely data-driven approach, they impose certain restrictions on the data characteristics or requires a huge volume of data in the training phase.

# Role of Wavelet Transformation

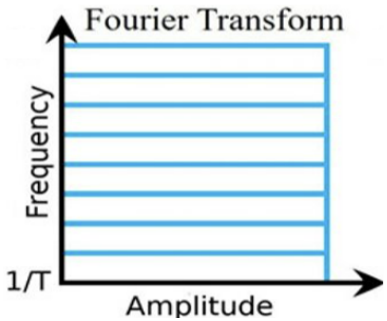
## Log transform

- Reduces the variability of skewed datasets.
- Transformed datasets conform more closely to normal dist.
- Highly impacted by outliers.
- Errors are symmetric on the original scale but asymmetric on the log scale.



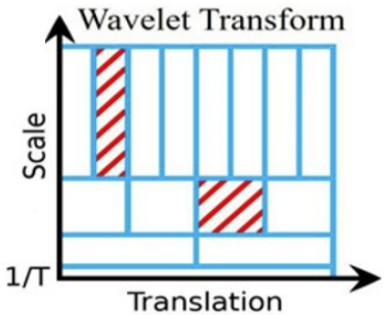

## Fourier transform

- Ideal for periodic signals.
- Represents a signal only in frequency domain
- For non-periodic signals with time-varying features, it gives averaged data, hence unsatisfactory.




## Wavelet transform

- Generalization of Fourier transform.
- It allows the independent choice of time and frequency resolution at different times and frequencies.



# EWNet Model: Basic idea

- We proposed an Ensemble Wavelet Neural Network (EWNet) that possesses the capabilities to handle the complex characteristics of epidemic datasets through its stable learning structure.
- Proposed EWNet architecture decomposes the epidemic data into a pre-specified number of "details" and "smooth" series by applying a maximal overlapping version of discrete wavelet transformation (MODWT). Subsequently, the "details" and "smooth" series are modeled using an autoregressive neural network with a pre-defined architecture in an ensemble setup.
- EWNet$(p, k)$ model is a non-stationary and non-linear model which can be written as follows:

$$y_t = \sum_{j=1}^{J} f_j(D_{j,t}) + f_0(S_{J,t}),$$

where $J+1$ ($\lfloor log_e(\text{length of training data}) \rfloor$) is the number of wavelet levels, $D_{j,t}$ $(j = 1, 2, \ldots, J)$ and $S_{J,t}$ is the corresponding details and smooth coefficients of the time series $y_t$, $f_i$ $(i = 0, 1, \ldots, J)$ is the one-hidden layered feedforward neural network with $p$ input nodes and $k$ hidden nodes ($\lfloor \frac{p+1}{2} \rfloor$).

# Architecture of EWNet Model

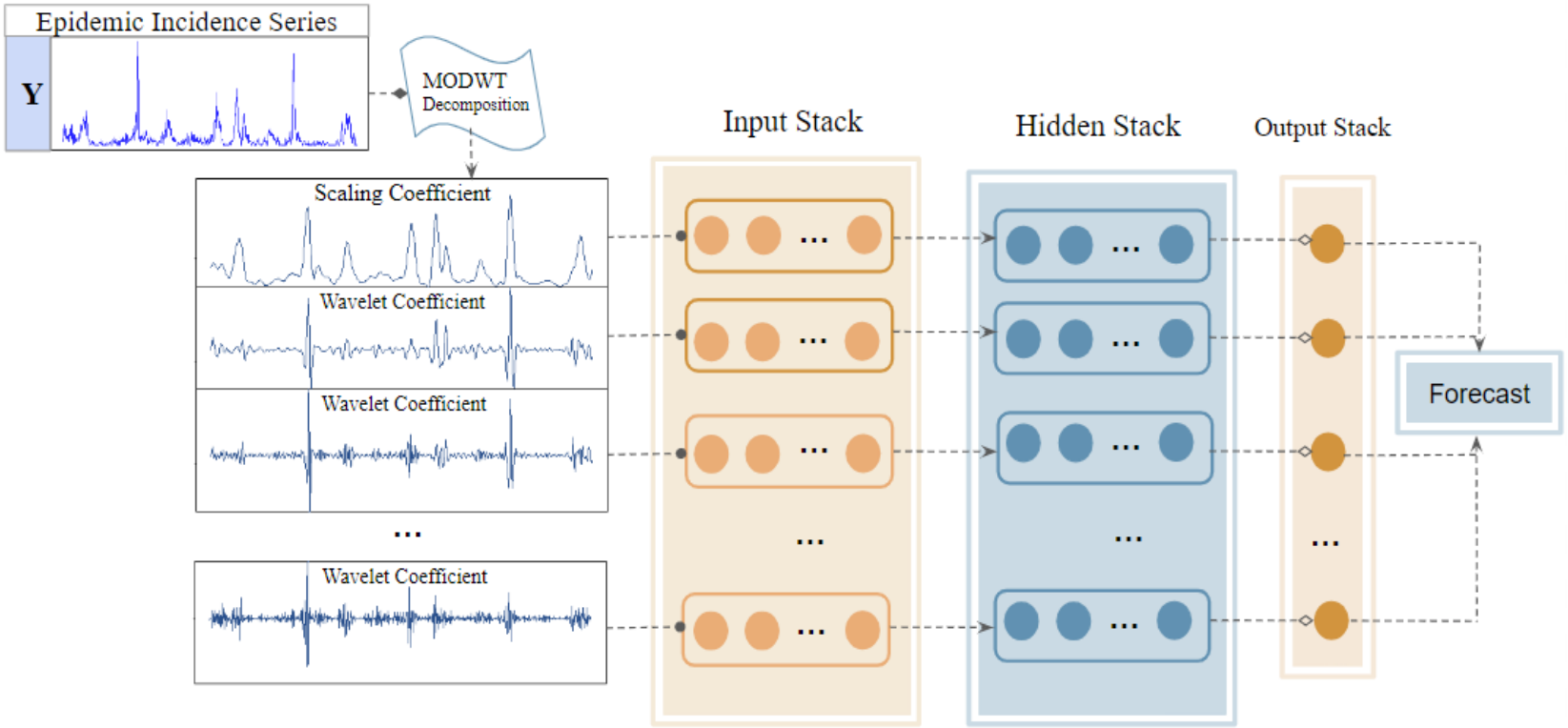



Figure:

Flow diagram of the proposed EWNet model

# Remarks

- Suitable for non-stationary, seasonal, and non-linear forecasting problems with limited historical data.
- Theoretical properties (learning stability, geometric ergodicity, and asymptotic stationarity) ensure the stability of model output.
- Simple and easily interpretable model, fast in implementation due to pre-defined architecture (multivariate set-up is yet to be explored).
- Experimental results suggest a significant improvement in long-range forecast accuracy owing to the wavelet decomposition.
- Epidemic dataset repository:
  https://github.com/mad-stat/Epicasting/tree/main/Datasets
- R package for implementation:
  https://cran.r-project.org/web/packages/epicasting/index.html

# Proposed Method - 3

## FORECASTING NONLINEAR CHAOTIC SYSTEMS USING PHYSICAL LAWS AND DL APPROACHES.

### Related Publications:


1. Zakaria Elabid, Tanujit Chakraborty, and Abdenour Hadid. "Knowledge-based Deep Learning for Modeling Chaotic Systems", **IEEE International Conference on Machine Learning and Applications (ICMLA)**, (2022): 1203–1209.

2. Arnob Ray, Tanujit Chakraborty, and Dibakar Ghosh. "Optimized ensemble deep learning framework for scalable forecasting of dynamics containing extreme events", **Chaos: An Interdisciplinary Journal of Nonlinear Science**, 31 (2021): 11.

# Extreme Climatic Events Prediction Problem

- Extreme climatic events pose a significant threat to the existence of mankind.
- Although these geophysical systems are governed by complex physical laws, their timely predictions are difficult but crucial for reducing the devastating impacts.
- Previous works in predicting chaotic systems using physical laws have always complemented the data-driven ML and DL approaches.
- In recent years there is a convergence between the approaches resulting in Physics Informed ML approaches.

# Proposed Knowledge-based Deep Learning (KDL)

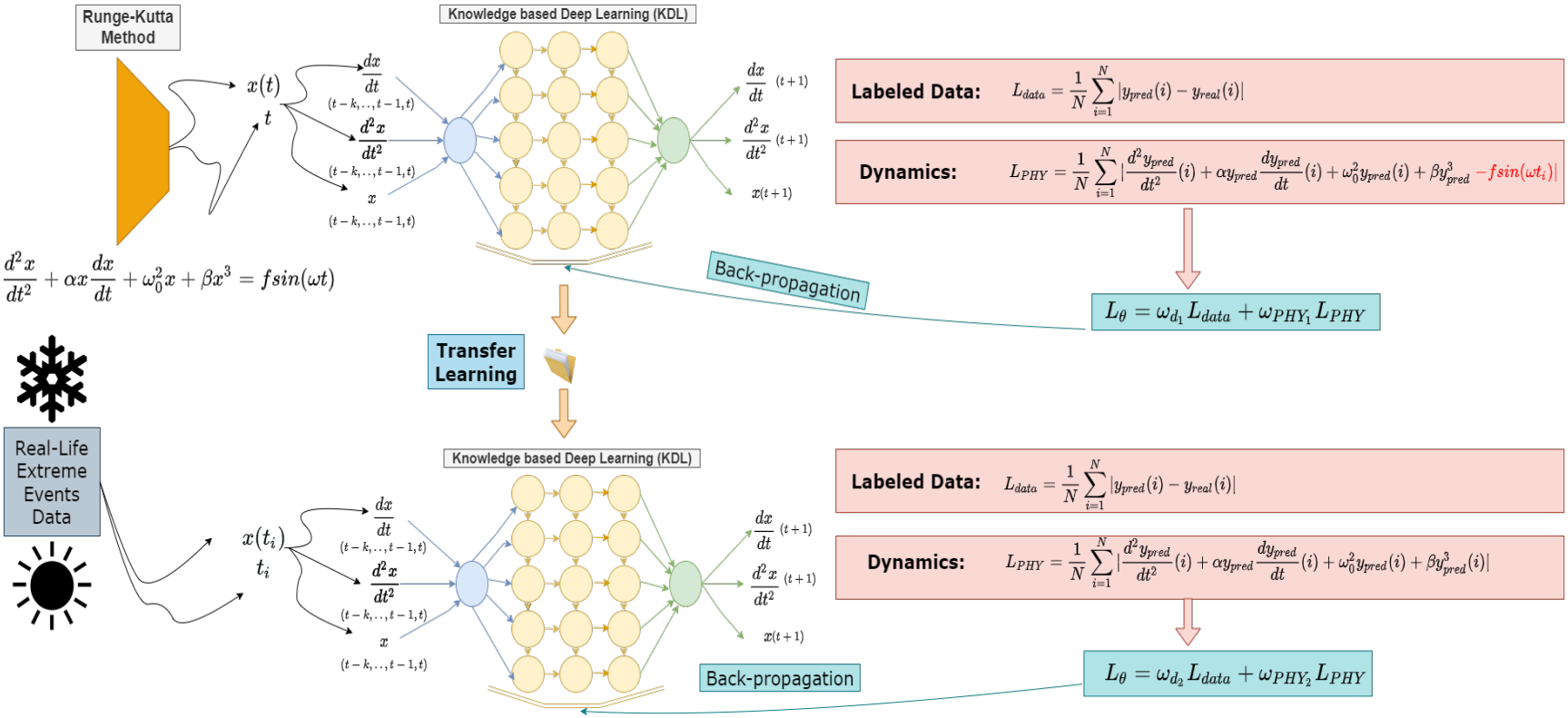



Fig: Proposed KDL architecture. A differential equation is simulated from a Lineard system into a time series $t, x(t)$. The first and second-order derivatives of $x$ are computed and fed along with $x$ to our model that forecasts the next instances through a backpropagation mechanism of the losses: 1) Data loss 2) Physics loss: Predicted data should satisfy the dynamics. The same process is used on real-world datasets with the addition of transferred knowledge from the pre-training.

## Some Possible Directions

# Research Directions on Forecasting+GNN

*" With the recent advancements in graph neural networks (GNNs), there has been a surge in GNN-based approaches for time series analysis that can explicitly model inter-temporal and inter-variable relationships, which traditional ML and other deep neural network-based methods struggle to do" (Ming Jin et al. 2023)*

## Prediction Problems:

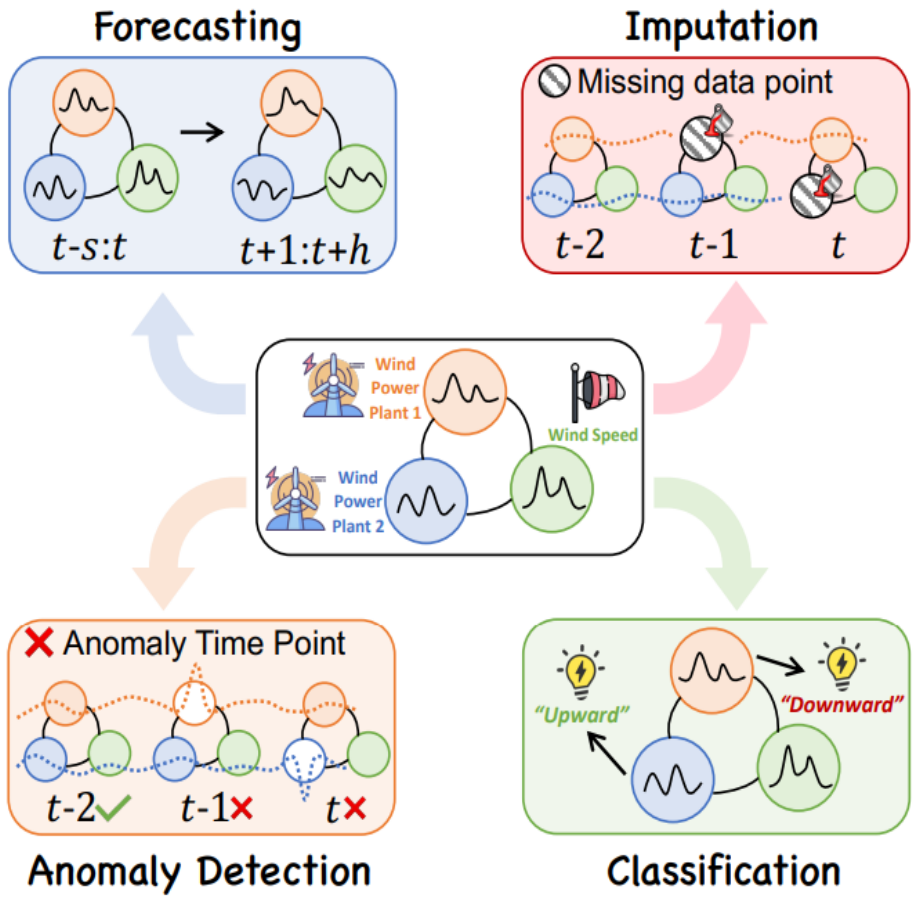
- Time series are the primary data type used to record dynamic system measurements and generated in great volume by both physical sensors and online processes (virtual sensors).
- Four fundamental dimensions: Forecasting, classification, anomaly detection, and imputation.

# Graph neural networks for time series

- This spatial-temporal characteristic is a common feature of many dynamic systems, including an example of a wind farm is presented here.

- In Fig., underlying time series data displays a range of correlations and heterogeneities that contribute to the formation of complex and intricate patterns, posing significant challenges for effective modeling.

- In this example of a wind farm, different analytical tasks can be categorized into time series forecasting, classification, anomaly detection, and imputation.



# Future Directions

## Problem:

- Traditional SVR, GBDT, VAR, and ARIMA struggle to handle complex time series relations (e.g., nonlinearities and inter-series relationships), resulting in less accurate prediction results.
- CNN, RNN, and Transformers have shown significant advantages in modeling real-world time series data. However, one of the biggest limitations of the above methods is that they do not explicitly model the spatial relations existing between time series in non-Euclidean space which limits their expressiveness.
- Similar situation exists in Biomedical data. Potential Remedy: GNN and PINN-based approaches.

*All models are wrong*
*but some are useful*

George E.P. Box