Project title: Trustworthiness and Robustness of (Deep) Neural Networks for Financial Big Data

Supervisors:

Hao Wang (main supervisor), Associate Professor, hawa@ntnu.no, Dept. of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering

Ibrahim Hameed, Associate Professor, ibib@ntnu.no, Dept. of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering

Richard Glavee-Geo, Associate Professor, rigl@ntnu.no, Dept. of International Business, Faculty of Economics and Management

Aim of the project:

The aim of this project is to study the trustworthiness and robustness of (deep) Neural Networks and develop robust algorithms based on the real financial datasets through TEFT Lab.

Project Description:

The initiative of *Open Banking* (OB) potentially can make the best possible use of existing bank products and assets, and more importantly, build entirely new business models to meet new expectations of customers due to their digital experience and interface offered by companies such as Amazon and Uber[1]. Driven by OB, more data will be collaboratively shared and used to build new applications for all parties including consumers, small businesses, lenders, fraud detectors. Therefore, financial big data processing and analytics are the basic building blocks to enable the development of OB.

The advantage of this project comes from the strong domain expertise and data resources supported by Sparebanken Møre through the TEFT lab and the financial competences from Dept. of International Business, Faculty of Economics and Management. This PhD is expected to work closely with other PhDs, researchers, and domain experts associated in the TEFT Lab.

Very recently, deep learning (DL) has been successfully applied in data analytics for many domains including the financial sector. However on the other hand, high-profile researchers and practitioners (e.g., Gary Marcus in his "Deep Learning: A Critical Appraisal") in the data community are raising critical concerns on the *transparency* of deep learning and neural networks (NN) in general:

> *The transparency issue, as yet unsolved, is a potential liability when using deep learning for problem domains like **financial trades** or medical diagnosis, in which human users might like to understand how a given system made a given decision. As Catherine O'Neill (2016) has pointed out, such opacity can also lead to **serious issues of bias**.*

It has been observed that state- of-the-art NN are highly vulnerable to *adversarial perturbations*, i.e., given a correctly-classified input x, it is possible to find a new input x' that is very similar to x but is assigned a different label. For instance, in image-recognition networks it is possible to add a small amount of noise (undetectable by the human eye) to an image and change how it is classified by the network. **These trustworthy problems are the definite barrier for the adoption of these latest technologies into critical applications such as financial services and justify urgent attention!**

[1] Open Banking: The Art of the Possible. NCR white paper.

Project plan:

Methodology. There are mainly two hypotheses on the weakness of *adversarial perturbations* for NN. The first one was that high complexity and non-linearity of neural networks can assign random labels in areas of the space which are under-explored. But this hypothesis has been refuted because 1) being unable to justify the transferability of adversarial samples from one model to another; 2) Linear models also suffer from this phenomenon. A *linearity hypothesis* has been proposed instead: deep neural networks are highly non-linear with respect to their parameters, but mostly linear with respect to their inputs, and adversarial examples are easy to encounter when exploring a direction orthogonal to a decision boundary. Another conjecture to explain the existence of adversarial examples is the accumulation of errors while propagating the perturbations from layer to layer. A small carefully crafted perturbation in the input layer may result in a much greater difference in the output layer, effect that is only magnified in high dimensional spaces, causing the activation of the wrong units in the upper layers.

Currently there are mainly two approaches to address this problem of NN: 1) casting the problem as to solve an optimization problem; 2) the verification approach that e.g. defines a region of safety around a known input and applies SMT (Satisfiability Modulo Theories) solving for checking robustness of NN.

Assessable objectives. The PhD to be hired in this project is expected to:

AO#1: Collaborate with other researchers, financial professionals, and PhDs in TEFT Lab and acquire the transferrable knowledge and skills on processing and analyze the financial data;
AO#2: Study and identify the adversarial perturbations for NN models for the financial data;
AO#3: Explore both the optimization approach and the verification approach and develop effective methods to identify the robust regions in the NN classification;
AO#4: Improve robustness of NN and develop interpretable machine learning methods for financial data with high trustworthiness.

Work Plan		**Year#1**	**#2**	**#3**	**#4**
AO#1	Transferrable knowledge and skills	■			
AO#2	Adversarial perturbations for NN models	□■	■		
AO#3	Methods to identify robust regions in NN classification		■	■	
AO#4	Interpretable ML/DL methods for financial data			■	■