

Deep-Stress: A deep learning approach for dynamic balance sheet stress testing

Anastasios Petropoulos¹, Vasilis Siakoulis¹, Nikolaos Vlachogiannakis, Evaggelos Stavroulakis

Abstract

The recent financial crises amplified the need for rigorous stress testing in order to assess the resilience of the banking systems under an adverse macro scenario by regulatory authorities. In this paper, we present a dynamic balance sheet simulation engine for stress testing, called Deep-Stress, which constitutes a new approach for emulating bank's key financial variables in a holistic way by utilizing deep learning algorithms. To evaluate the performance of the proposed model we compare its forecasting accuracy with other accepted stress testing frameworks: constant balance sheet approach and dynamic balance sheet approach with satellite modelling. The prediction error of the Capital Adequacy Ratio drops significantly under the deep learning approach, due to its better performance in simulating the one year ahead P&L evolution of the financial institutions. The proposed methodological framework can become a powerful tool for macro prudential stress testing and could strongly increase the signalling power of an early warning system in order to predict future financial crises and individual bank's failures.

Keywords: Stress Testing, Deep Learning, Bayesian Model Averaging, Capital adequacy Ratio, Forecasting, Neural Networks, Dynamic balance sheet, Constant balance sheet assumption.

JEL: G01, G21, C53

This version: March 2019

The views expressed in this paper are those of the authors and not necessarily those of Bank of Greece.

1. Introduction – Motivation

Financial Stability is a core component for economic prosperity of countries and individuals. The recent financial crises of 2007 had a measurable effect on the life of many individuals across the globe through the realization of significant income reduction, the increasing unemployment rate as well as an overall economic slowdown [3]. The methods of risk management that have been employed before the crises proved to be inadequate to provide early warning signs to central governments and banks in order to proactively intervene and prevent adverse financial events. Regulatory authorities, and international organizations such as IMF performed stress testing exercises long before the financial crisis of 2007 to assess the resilience of the banking system but failed to predict the unprecedented economic turmoil after Lehman's default.

Previous Stress testing frameworks disregarded the propagation channels of a default event through the whole micro macro dynamics of the global interconnected financial system. Additionally, the non-linear relationship that materialize between the macro economy and the financial balance sheets was not adequately captured due to the broadly use of linear regression models. Moreover, weaknesses in the validation function of stress testing frameworks also decreased the confidence in the quantification of the impact of an adverse macro scenario in the banking system. Subsequently, market participants and regulators have performed rigorous stress testing exercises enhancing statistical , using more granular data, in order to assess and predict the risks and in some cases attempting to quantify second round effects stemming from a liquidity shock or from the default of a counterparty.

Another aspect in the current regime, i.e. post crisis, of supervisory regulation is the collection of a significant amount of granular information as a response to a more proactive supervision. Although, the integration of *big data* in the banking supervision, regulatory authorities has not yet explored statistical techniques such as machine learning, in order to extract more information regarding the risks in their banking systems. Segmentation, classification and data mining functionalities are important tools for regulators to identify weaknesses in the supervised financial entities that can be further enhanced by using machine learning techniques.

Machine learning algorithms have drastically improved the capabilities of performing pattern recognition, signal analysis and forecasting in various scientific fields such as biomedical, engineering and social sciences. The structure of machine learning method offers the ability to adjust in streaming sequences using continuous learning algorithms as well as offer state-of-the-art performance in the recognition of and evolving patterns in time series data. In addition, deep learning has proven effective to deal with high dimensional data. Recent studies demonstrate that machine learning techniques could lead to better predictive performance in financial time series modelling problems due to their multidimensional and non-parametric structure [17,18,19]. This can be attributed to their capacity to learn and adapt to new data. Thus, improve their performance over time, offer increased capabilities to capture non-linear relationships, and decompose the noise that often exist in financial data. Furthermore, this new generation of statistical algorithms offer the necessary flexibility in modelling multivariate time series, as its structure includes a cascade of many layers with non-linear processing agents. The functionality of deep learning networks consists of the interaction of layers that simulate the abstraction and composition of similar functions in the human brain. Therefore, via capturing the full spectrum of information contained in financial datasets, deep learning networks are capable of exploring in depth the inherent complexity of the underlying dynamics and dealing with *high-dimensional time series data*.

This empirical study introduces a new statistical technique for stress testing using deep learning algorithms to model banks financial data in a holistic way. In particular, financial or macro

shocks are propagated to banks' balance sheets by simultaneously training deep neural networks with macro and financial variables, thus, taking advantage of their capabilities to capture more information hidden in big datasets. We develop inference algorithms for our networks, suitable for learning financial time series data on a multivariate forecasting setup.

The main propose of this study is to illustrate a holistic framework for balance sheet stress testing, which overcomes the limitations of the currently approaches and yields more robust results by loosening the static balance sheet assumption. Our research analysis based on the intersection of computational finance and statistical machine learning, leverages the unique properties and capabilities of deep learning networks in order to increase the prediction efficacy of the capital adequacy and minimize the modelling error. Under the proposed approach, forecasting of balance sheet items can be heavily supported by artificial intelligence algorithms simulating better the propagation channels of the macro economy into the financial institutions business models. Our vision is to provide a stress testing framework that leads to an implementation of *an* early warning system for financial shocks on individual banks' balance sheets.

The structure of this study is organized as follows. In section 2, we are focused on the related literature review regarding financial institutions stress testing. Section 3 describes the data collection and processing. In section 4, we provide details regarding the estimation process of the various stress testing frameworks examined in this study. In section 5, we compare various methodologies and provide our experimental results using a test dataset of financial balance sheet sequences. With our methodology, we assess the applicability of the proposed approach as well as the generalization in its forecasting capacity. Finally, in the concluding section 6, we summarize the performance superiority of the proposed methodology and identify any potential weaknesses and limitations of this study, while we also underline the need for further research.

2. Literature review

The architecture of current stress testing frameworks is usually a feed forward shock engine not capturing the nexus of relationships of the highly interconnected financial system and the accompanied feedback loops in the macro environment. The following graph describes a typical macro stress testing framework used mainly by regulatory authorities currently.

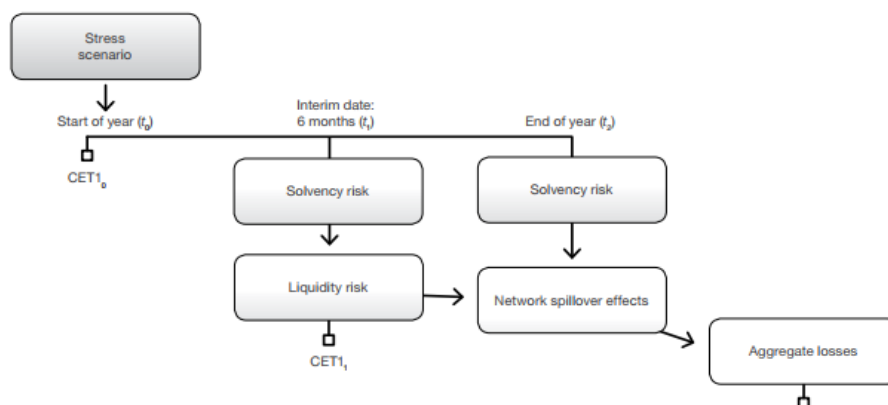


Figure 1: Current feed forward architecture of currently established stress testing frameworks

A typical stress testing engine is composed by four elements: the perimeter of risks subjected to stress, the scenario design, the calculation engine that transforms the shocks into an outcome in Banks balance sheet, and a measure of the outcome [1]. In particular the most well-known stress testing exercises currently publicly available are: EBA[4], CCAR (FED) [8], PRA - Bank of England [7] , ECB (top down) [2] [6], Bank of Canada[28], Central Bank of Austria (ARNIE)[27], IMF[29] and Bank of Greece (Diagnostic Exercise)[5]. The structure of all these stress testing exercises follows a left to right flow to estimate the impact of an adverse shock in the economy. One of the basic components of all these exercises is the time horizon which span from 2 to 5 years to estimate future losses for the participating banks. During this period the macro economic scenario is given. This set of macro scenarios is passed through to financial institutions to project their P&L and RWA and eventually estimate capital using regulatory hurdle rates. Some of these exercises include in their structure a second round effects mechanism for the banking system to account for contagion risk. Macroeconomic feedback effects i.e for example the impact of a significant institution becoming insolvent in the macro economy, usually are not considered in these frameworks. Stress tests under this structure can mainly serve as a tool to challenge the recovery plans of banks and to assess their viability. But their role as an early warning system is questionable.

As Drehmann [11] pointed, systemic banking crises are reflected in the performance of credit and property prices and usually they appear at the high point of the medium-term financial cycle. Therefore, crisis starts before it's depicted in macro scenarios. According to Borio [1] a financial system is not fragile when a large financial shock materializes but when even a small negative change in financial and macro variables is amplified through the different dynamic system relationships and can lead to a systemic shock. For example after the default of Lehman the financial market crashed and the US GDP exhibited a sharp decrease causing a structural break in the macro data time series. Current versions of stress tests possess a macro scenario over time in a static way without modelling or tracking in a path dependent nature the multistep decision process and financial behaviour that in reality takes place from all economic participants. [30] Furthermore, non-linearity is not modelled adequately in the statistical techniques currently employed. Risks under the current globalized market tend to be amplified when a stress event occurs. Non-linear relationship kicks in through channels of amplification leading to a chain of unpredictable events from the static nature of stress test. Under stressed conditions the relationships between modelled variables are non-linear [31] [9] and exhibit structural breaks [10]. Stress testing frameworks are composed by standalone models usually combined in a qualitatively manner. A small single-step prediction error at the beginning could accumulate and propagate when combined without taking the correlation of the financial variables, often resulting in poor prediction accuracy. Furthermore standalone models can lead to double counting effects or overestimation of the impact stemming from the changes of the predefined macro variables. Finally univariate setups are not able to model adequately complex distributed variables with non-linear behaviour.

Current stress testing frameworks exhibit simplification assumptions that may affect the reliability of the final estimation. EBA EU's wide stress testing is a bottom up exercise covering only specific risk on banks individuals balance sheet based on a macro scenario usually based on simplified assumptions. One of the weaknesses in EBA methodology is the static balance sheet assumption i.e assets and liabilities remain constant over the horizon without acknowledging for management actions and new generation of loans. In addition mitigation actions are taken into account after the stress testing are finalized through a strong qualitative overlay and not in a dynamic way. [4]

System wide stress testing exercises on a micro prudential level are heavily relied upon on the interaction with individual banks with respect to data analytics and propagating the macro scenarios to their balance sheet. Thus estimations are not performed in a uniform statistical process but inherit the model deficiencies and forecasts errors embedded in banks individual

models. The heterogeneity in the results increases the estimation errors significantly and there is no robust process for regulators to account for it. Thus the need for independent central modelling for simulating the financial system is of great importance. [13] Furthermore the stress testing process involves the disclosure of the methodological framework to all market participants which in turn there are evidence of second round effects regarding the accounting treatment of banks. Specifically based on the study [14] banks that participate in regulatory exercises tend to manipulate their provisions for credit risk as well as to absorb the impact of the upcoming stress test.

Finally Stress Testing outcomes in the current regulatory exercises heavily rely on regulatory ratios like capital adequacy ratio which in turn is highly dependable on the estimation of RWA. Evidence in the literature [32] indicates that relaying in the risk weights applied internally by the financial institutions under the Basel Framework can lead to underestimation in capital needs. This is driven by the significant variability stemming from internal models of the banks when applying internal model methods. Furthermore the regulatory framework currently employed for assessing the RWA cannot capture the hidden risk in banks complex portfolio structure. In the current literature [33] there evidence regarding especially more sophisticated banks (A-IRB) that they may perform regulatory arbitrage and manipulates their true risks to lower their capital requirements. Thus robust macro modelling of the RWA using an independent top down model is important to account for these cases.

Although a significance progress in designing stress testing has been implemented in recent years, there are concerns that this type of exercises cannot be used as early warning systems for financial distress [3]. By analysing the publications regarding stress testing exercises either performed by regulators or individuals banks we outlined a series of weaknesses and inefficiencies to provide a clear and concise view on the nature and on the way how the proposed approach in this study, DeepStress, attempts to address part of the aforementioned weaknesses.

Deep neural networks architecture is one of the main innovations in our proposed approach for dynamic balance sheet stress testing. DeepStress is putting all the components together in a multivariate structure. We identify the main channels of risk propagation in a recurrent form to account of all the existing evidence of feedback effects in a financial institutions' balance sheet. The current architectures is constrained by the use classical econometric techniques which offer limited capabilities for simulating complex systems. DeepStress accounting for temporal patterns in banks' balance sheets provides a dynamic modelling approach. This is achieved through the multivariate training of deep neural networks taking account the dynamic nature of banks metrics and the whole structure of the bank's balance sheet. DeepStress is composed by multivariate input and output layers able to capture the cross correlation between balance sheet items and the macro economy. Training is performed as one big complex network minimizing estimation errors and double counting effects among various financial variables.

To account for non-linear relationships that materialize under adverse macroeconomic conditions machine learning techniques like deep learning can provide more efficient estimations. Deep Neural networks based on academic literature are capable of simulating real life phenomena where relationships are complex. Therefore, our proposed framework using multilayer deep networks envisages in capturing the dynamics inherent in a financial distress. In addition the architecture of DeepStress aims to capture the amplifications channels leading to structural breaks.

DeepStress is a proposed micro macro prudential stress testing framework independently assessing the system without relying on banks for performing its estimations. Methodology applied relies only on publicly available data and models are developed in a uniform way thus making the process of validation and error correction more feasible to be performed centrally.

In addition offers the opportunity to experiment on advanced statistical machine learning techniques a need recognized also in the academic literature [12].

To sum up, our modelling approach is balanced between capturing the determinants that strongly affect the health of a financial institution, while at the same time developing a dynamic balance sheet simulator engine for establishing an early warning system to predict bank failures under an adverse scenario. The modelling framework that we implement captures temporal dependencies in a bank’s financial indicators and the macro economy. At the same time, it explores up to 3 years of lagged observations, which are assumed to carry all the necessary information to describe and predict the financial soundness of a bank, and combines their evolution with the relevant macroeconomic indicators. Overall, we envisage that DeepStress offers a dynamic simulation engine projecting the whole status of a financial institution one year ahead in an efficient way.

3. Data collection and processing

We have collected information of non-failed, failed and assisted entities from the database of the Federal Deposit Insurance Corporation (FDIC), an independent agency created by the US Congress in order to maintain the stability and the public confidence in the financial system. The collected data are related to all US banks, while the adopted definition of a default event in this dataset includes all bank failures and assistance transactions of all FDIC-insured institutions. Under the proposed framework, each entity is categorized either as solvent or as insolvent based on the indicators provided by FDIC. Observations regarding failed banks are excluded from the analysis since stress testing is performed on healthy financial entities.

The dataset covers the 2007-2015 period; a 9 years’ period with quarterly information resulting in dataset with more than 175,000 records. The selected time period, seems to approximate a full economic cycle, in terms of the Default Rate evolution. Figure 1, shows the number of records included in each observation quarter and the corresponding default rate. From a supervisory perspective, most of the financial institutions in the sample apply the standardized approach for measuring the Credit risk weights assets based on the United States adaptation of the Basel regulatory framework [20].

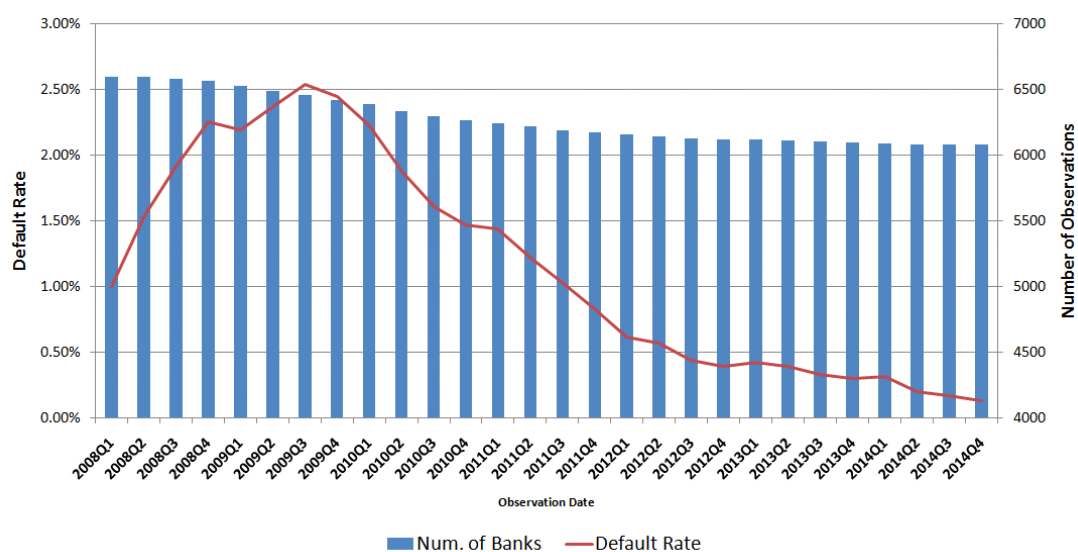


Figure 2: USA financial institutions in the sample. Historical overview for the period 2008-2014 of the failed entities (source: FDIC)

The dataset was split into three parts (Figure 2). An in-sample dataset (Full in sample) that is comprised of the data pertaining to the 80% of the examined companies over the observation period 2008-2013 amounting to 101.641 observations. For performing hyper parameter tuning of deep neural networks we define an out-of-sample dataset (validation sample), including the rest 20% of the observations for the period 2008-2013 amounting to 25.252 observations. This is useful for deep learning models, in which the training sample is used to train various candidate models with different architectures and specifications, while the validation set is used for selecting the best parameter setup and avoid overfitting in the training dataset. This way the generalization capabilities in other datasets of the final selected model increases substantially. Finally performance evaluation is investigated on an out-of-time dataset (test sample) that spans over the 2013-2015 observation period reaching 48.756 observations. In all cases, the dependent (target) variable is the CAR ratio of each bank in the end of the one year forecast horizon. To summarize, we performed model fitting using exclusively the available training sample prescribed above. To perform model selection, we employed five-fold cross-validation, using predictive accuracy as our model selection criterion (CAR ratio prediction error). Performance evaluation results are assessed on the available test sample, to allow for evaluating the generalization capacity of the developed models.

In developing our model specifications, we examine an extended set of variables that fully describe the financial status of each bank in the sample. The complete list of financial variables examined is included in Annex II. In addition to the above-mentioned variables, we have also included in the dataset quarterly observations of the most commonly used macro-economic variables. Macro variables are the main input in the models developed since they are important for scenario analysis under a stress testing framework. The current model setup includes contemporaneous macro variables along with 3 year lags. The intuition for this approach is to build models for scenario prediction which is the main methodology for Stress Testing modelling. The macro variables included in the development are:

- GDP: Gross Domestic Product growth
- EXPORT: US Total Exports growth
- GOVCREDIT: Government Credit to GDP
- DEBT: US public debt to GDP
- GOVEXP: US government expenditure to GDP
- INFLAT: US inflation
- RRE: House Price Index growth
- UNR: Unemployment Rate
- YIELD10Y: 10Y US sovereign bonds yields
- STOCKS: US Stock index – S&P 500 returns

The relevant stress financial variables for simulating the profitability and the risk weighted assets of each financial institution are:

- NLOAN: Net loans exposure
- DEP: Total Deposits
- DDEP: Total domestic deposits
- ASSET: Average Total Assets
- EASSET: Average Total Earning Assets
- EQUITY: Average Total Equity
- LOAN: Average total loans

- CFD: Deposits Cost of funding
- YEA: Yield on earning assets
- NFIA: Noninterest income to average assets
- RW: Risk Weight Density
- LOSS_LOAN: Loss allowance to loans
- RWA: Total risk weighted assets
- CAR%: Total risk based capital ratio

Modelling for the evolution of the balance sheet is performed on the growth rate of four key financial items: Deposits, Total Earning Assets, Total Loans and Total Assets.

In order to capture the idiosyncratic characteristics of each financial entity, 3 year lags are included in the training process for each financial variable. In the final model setup the use of multiple years financial and macroeconomic variables allows for capturing internal trends of key items of a bank' balance sheet and also the degree each entity is affected by the status of US economy.

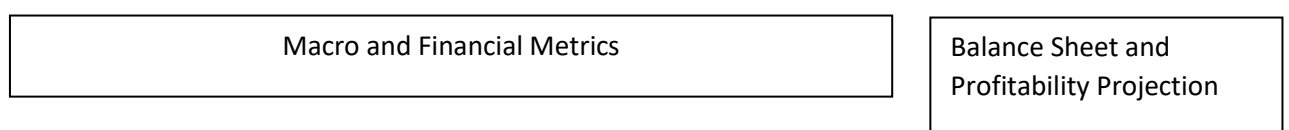
4. Model Development

The success of the stress testing exercises performed in the past by regulatory authorities was put under scrutiny by all market participants and the research community. In order to investigate the capabilities of the proposed DeepStress approach for stress testing against broadly used technical frameworks we simulate two additional methods for balance sheet forecasting to benchmark its performance. Specifically we developed a constant balance sheet approach following the framework adapted by EBA to perform EU wide stress tests [4] and a dynamic balance sheet approach support by a group of satellite models to forecast individual financial variables used by other regulatory authorities like ECB for macro prudential stress testing. In this section we provide an overview of the overall setup of the study and technical details of the three individual approaches employed.

General Setup of the Study

The main component of a micro prudential solvency stress testing framework is the projection of a financial institution capital adequacy ratio or recently the CET 1 ratio (Core Equity Tier I ratio). In this study we develop a Deep Neural Network structure which receives as input the Macro variables and Balance sheet components mentioned in chapter 3 and provides as output the balance sheet and profitability structure of the bank on one year horizon as measured by 9 core variables namely Net loans, Deposits, Assets, Earning Assets, Deposits, Cost of funding, Yield on earning assets, Noninterest income to assets, Risk Weight Density and Cost of Risk {Loss allowance to loans}

Figure 3: Stress Test Deep Neural Network



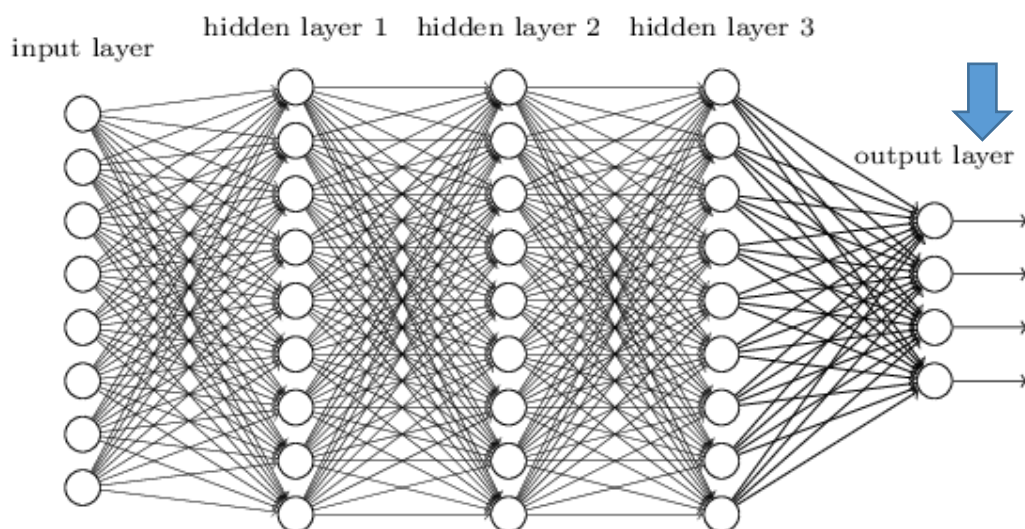


Figure 3: Stress Test Deep Neural Network architecture

We focus on the forecasting of the CAR ratio since CET-1 ratio was introduced under Basel III and is not available throughout our dataset. Specifically, our aim is to project the in a one-year-ahead the CAR ratio of each financial institution in the sample. CAR ratio by definition is the ratio of a bank's capital over the risk weighted assets in each time point t . In order to simulate the core mechanics of a stress testing framework we simulate the evolution of the key financial variables of a financial institutions balance sheet. The main setup is that we project one year ahead the evolution of the capital and the risk weighted assets in order to forecast the one year ahead CAR. The approach followed to adjust the capital in time t is given by the formula:

$$\text{Capital}_t = \text{Earnings from Assets}_t - \text{loans loss provisions}_t + \text{Net fees and commissions}_t - \text{cost of funding from deposits}_t + \text{Capital}_{t-1}$$

(1)

In order to adjust the capital of each entity we model 8 key financial variables. The first four variables refer to the dynamic evolution of the balance sheet i.e the growth of the asset and liability side: the growth rate of Deposits, Total loans, Total Assets, Total Earning Assets. The remaining 4 variables refer to the yield in the next year of each item from the asset or liability side: cost of risk of loans, yield on earning assets, yield on deposits and yield of net fees and commissions of total assets.

The RWA are adjusted in 3 different ways depending on the ST methodology. Specifically in terms of the deep learning technique we project the growth of the RWA, for satellite modelling a dedicated model is trained to project the RW density of each financial institution in the sample, while for the constant balance sheet approach we assume RWA remain constant for one year.

Before developing the relevant statistical models we remove and linearly interpolate the outliers utilizing the R package DescTools¹. The algorithm is encompassed in the Winsorize

¹ <https://cran.r-project.org/web/packages/DescTools/index.html>

function in R , and attempts to clean the data by means of winsorization, i.e., by shrinking outlying observations to the border of the main part of the data.

Deep Learning MXNET

We implement a Deep Neural Network (henceforth DNN) to address the issue of dynamic balance sheet forecasting. Deep learning has been an active field of research in the recent years, as it has achieved significant breakthroughs in the fields of computer vision and language understanding. In particular it has been extremely successful in as diverse time-series modelling tasks as machine translation [21, 22] machine summarization and recommendation engines [23]. However, its application in the field of finance is rather limited. Specifically, our paper constitutes one of the first works presented in the literature that considers application of deep learning to address the challenging financial modelling task of financial balance sheet stress testing.

Deep Neural Networks differ from Shallow Neural Networks (one layer) on the multiple internal layers employed between the input values and the predicted result (Figure 2). Constructing a DNN without nonlinear activation functions is impossible, as without these the deep architecture collapses to an equivalent shallow one. Typical choices are logistic sigmoid, hyperbolic tangent and rectified linear unit (ReLU). The logistic sigmoid and hyperbolic tangent activation functions are closely related; both belong to the sigmoid family. A disadvantage of the sigmoid activation function is that it should be kept small due to their tendency to saturate with large positive or negative values. To alleviate this problem, researchers have derived piecewise linear units like the popular ReLU, which are now the standard choice in deep learning research ReLU. The activation layers increase the ability and flexibility of a DNN to capture non-linear relationships in the training dataset.

On a different perspective, since DNNs comprise a huge number of trainable parameters, it is key that appropriate techniques be employed to prevent them from overfitting. Indeed, it is now widely understood that one of the main reasons behind the explosive success and popularity of DNNs consists in the availability of simple, effective, and efficient regularization techniques, developed in the last few years. Dropout has been the first, and, expectably enough, the most popular regularization technique for DNNs [24]. In essence, it consists in randomly dropping different units of the network on each iteration of the training algorithm. This way, only the parameters related to a subset of the network units are trained during each iteration. This ameliorates the associated network overfitting tendency, and it does so in a way that ensures that all network parameters are effectively trained.

Inspired from these merits, we employ Dropout DNNs with ReLU activations to train and deploy feed forward deep neural networks. We employ the Apache MXNET toolbox of R^2 for implementing the deep learning algorithm. We postulated deep networks that are up to five hidden layers deep and comprise various numbers of neurons. Model selection using cross-validation was performed by maximizing the RMSE metric on the projected CAR.

In our setup multivariate deep learning networks will learn the balance sheet of financial institutions separately generating yearly forecasts by the interactions of layered neurons after receiving historical values of banks previous economic states. This hierarchical transmission of observed data between cascading layers of abstraction can decompose the structure of a bank balance sheet and foster the multivariate representation of the financial variables for capturing the correlations between various assets and liabilities. This provides the functionality of simultaneously modelling the balance sheet as a whole instead of using satellite models of

² <https://mxnet.incubator.apache.org/api/r/index.htm>

regular stress testing frameworks. This is feasible based on the fact that DNN are composed of multiple features for input and output complex representations. Deep learning can facilitate the dynamic balance sheet projection approach through the non-linear relationships representations of each layer offering a more realistic approach for stress testing. Information flows through the system as a vector of macro and financial variables describing the state of both the bank and the macro economy at any time stamp during the forecast period. Specifically the input vector contains around 60 variables and the output vector is composed of 9 variables. The DNN architecture employed is capable of modelling the lead lag relationships between macro variables banks variables financial variables and sovereign variables. Finally in the DeepStress engine using the aforementioned multivariate forecasting setup on individual balance sheet we model simultaneously the RWA evolution of the bank and connect it to the macro environment.

Satellite Modelling – Bayesian Model Averaging

Satellite models are used for univariate estimations of the impacts of standalone balance sheet items in current stress testing frameworks [2]. A usual statistical technique employed by regulators and the banking industry is the Bayesian Model averaging. The main intuition behind the use of BMA econometric technique is to account for the uncertainty surrounding the main determinants of risk dynamics especially in a period of recession. This approach is able to handle a short time series of balance sheet realizations which is usually the case for stress testing. Thus BMA offers the possibility to perform multivariate modelling including all potential predictors with different weight while the output of each trained model remains univariate.

Using BMA, a pool of equations is generated using a random selection subgroup of determinants. Subsequently a weight is assigned to each model that reflects their relative forecasting performance. Aggregating all equations using the corresponding weights produces a posterior model probability. The number of equations estimated in the first step is large enough to capture all possible combinations of a predetermined number of independent variables. Thus Bayesian model averaging addresses model uncertainty and misspecification in selected explanatory variables in a simple linear regression problem.

To further illustrate BMA, suppose a linear model structure, with Y_t being the dependent variable, X the explanatory variables, α constant, β the coefficients and ε_t a normal error term with variance σ .

$$Y_t = \alpha_\gamma + \beta_\gamma X_{\gamma,t} + \varepsilon_t \quad (2) \quad \varepsilon_t \sim N(0, \sigma^2 I) \quad (2)$$

A problem arises when there are many potential explanatory variables in a matrix X_t which transforms the task of selecting the correct combination quite burdensome. The direct approach to inference in a single linear model that includes all variables is inefficient or even infeasible with a limited number of observations. It can lead to overfitting, multicollinearity and increased manual re-estimations to account for non-significant determinants. BMA tackles the problem by estimating models for all possible combinations of $\{X\}$ and constructing a weighted average over all of them.

Under the assumption that X contains K potential explanatory variables, BMA estimates 2^K combinations and thus 2^K models. Applying Bayes' theorem (6), model averaging is based on the posterior model probabilities.

$$p(M_\gamma | Y, X) = \frac{p(Y|M_\gamma, X)p(M_\gamma)}{p(Y|X)} = \frac{p(Y|M_\gamma, X)p(M_\gamma)}{\sum_{s=1}^{2^K} p(Y|M_s, X)p(M_s)} \quad (3)$$

In equation (3), $p(Y, X)$ denotes the integrated likelihood which is constant over all models and is thus simply a multiplicative term. Therefore, the posterior model probability (PMP) is proportional to the integrated likelihood $p(Y|M, X)$ which reflects the probability of the data given model M . Thus the corresponding weight assigned to each model is measured using $p(M_\gamma|Y, X)$ in equation (3).

In equation (3), $p(M)$ denotes the prior belief of how probable model M is before analyzing the data. Furthermore, to estimate $p(Y, X)$ integration is performed across all models in the model space and to estimate the probability $p(Y|M, X)$ integration is performed given model M across all parameter space. By performing renormalization of the product in equation (3), PMPs can be inferred and subsequently the model's weighted posterior distribution for estimator β is given by

$$p(\beta|Y, X) = \sum_{\gamma=1}^{2^K} p(\beta|M_\gamma, Y, X)p(M_\gamma|X, Y)(4)$$

The priors, posteriors and the marginal likelihood employed in the estimation are described analytically in Appendix.

For model development, the same train set used for DNN is employed. Before applying the Bayesian Averaging algorithm we remove and linearly interpolate the outliers. In Bayesian Model Averaging estimation we employ unit information prior (UIP), which sets $g=N$ commonly for all models. We use also a birth/death MCMC algorithm (20000 draws) due to the large number of covariates included since using the entire model space would lead to a large number of iterations. We fix the number of burn-in draws for the MCMC sampler to 10000. Finally the models prior employed is the 'random theta' prior by Ley and Steel [25], who suggest a binomial-beta hyper prior on the a priori inclusion probability. This has the advantage that is less tight around prior expected model size (i.e. the average number of included regressors) so it reflects prior uncertainty about model size more efficiently. For robustness purposes we varied the used prior employing the Fernandez [26] propositions but the results were not substantially different.

In order to develop all the satellite models for this approach we employ the utilities of BMS R package³. After the training process 9 BMS models are developed: 4 for the growth of balance sheet items, 4 models are forecasting the yields of a various assets and liabilities and one model for forecasting the RW assets density.

Constant Balance sheet modelling setup

For the constant balance all balance sheet items are assumed constant along with the RWA metric for one year. Thus we combine the respective univariate satellite models BMA to project yields of assets and liabilities while assume zero growth in the balance sheet in order to project the CAR ratio one year ahead.

5. Model Validation - Experimental Evaluation

No thorough and consistent framework exists for validating the results of a stress testing exercise since the adverse scenario used in their design never materialize. Back testing methods is an important process to recognize modelling inefficiencies and fine tune the estimations taking into account specificities in the time series data that were not capture in the initial calibration and development phase. Thus in order to improve the quality of stress testing

³ <https://cran.r-project.org/web/packages/BMS/index.html>

rigorous validation procedures of actual vs predicted financial variables are important to be introduced [35]. Furthermore, according to previous studies the success of the stress testing exercises after the financial crises maybe be circumstantial [34] since no robust methods are applied to quantify their estimation error.

Following a different venue in this study we perform a thorough validation procedure in order to assess the robustness of our approach. In this section we summarize the results of the three approaches. More precisely, we report the performance results obtained from the experimental evaluation of our method, in terms of in-sample fit (train dataset) and out-of-time performance (test sample). Finally we report in terms of evaluating the model's predictive ability separately on the 2011 year to investigate the models behaviour during the European Sovereign crises. To sum up, after developing our Stress testing frameworks in the "train" dataset, we assess their performance results under three different time period samples. The first, being the "in sample", is used to in sample error of the each approach for evaluating overfitting. The second is the time period sample of 2011 which is included in the train dataset but is separately reported to investigate the performance of each model during a period of financial turbulence. While, under the third "Out-of-time" dataset the performance of each model is evaluated during a future time period for evaluating their generalization capacity. More precisely, we report performance results obtained by evaluating our method over a two year (8 quarters) out of sample time-period comprising, spanning from 2014 – 2015. Validation is performed with respect to the one year ahead forecast of the CAR ratio. Note that the last 2 two years of the dataset were not used for model development.

Prediction accuracy of the CAR ratio, as measured by the deviation between the forecast of each framework against the actual CAR ratio of each financial institution, is the main criterion to assess the efficacy of each method and to select the most robust one. In this section, we present a series of metrics that are broadly used for quantitatively estimating the forecasting accuracy on continuous outcomes. We evaluate the stress testing methods with the usual forecast metrics of Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE). These metrics are used so as to derive a full-spectrum conclusion regarding the relative forecasting power of each framework.

As we observe in Table 2, in the out of sample horizon the MXNET algorithm provides the best empirical performance. This is followed by the dynamic balance approach utilizing standalone satellite models methodology. Hence, MXNET deep neural networks offer significantly superior predictive accuracy under the CAR ratio forecasting setup on the test sample. Another remark based on the experimental results is that, by moving from simple neural networks to deep networks, we are able to infer richer and subtler dynamics from the data, thus increasing our capacity in modelling nonlinearities and cross-correlations among balance sheet P&L items. Deep learning offers a more efficient way to simulate the CAR ratio under a specific set of macro scenarios of key macroeconomic variables. This is also evident from the graph below where in the out of sample performance constant balance sheet and satellite modelling diverge significantly from the actual evolution of the CAR ratio in the dataset. In addition the two benchmark approaches exhibit higher volatility in their forecast based on the figure 4. The same holds also in figure 5 where the projected CAR is graphed only for large banks (more 200bl in assets).

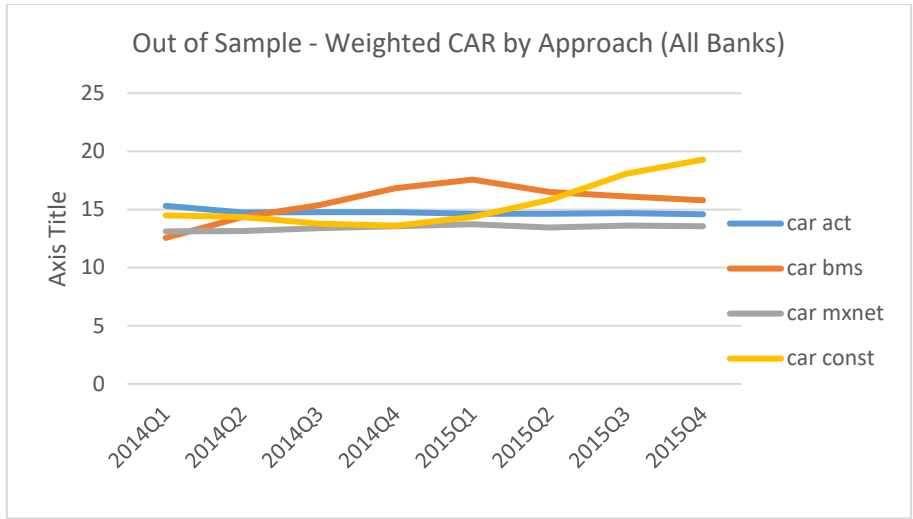


Figure 4: Out of sample back testing results of CAR ratio of the three balance sheet approaches (Whole Sample)

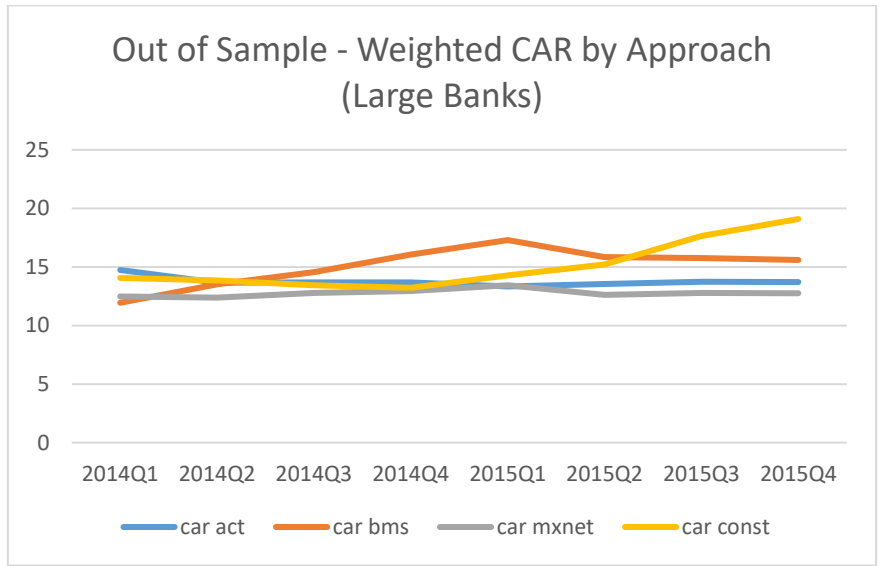


Figure 5: Out of sample back testing results of CAR ratio of the three balance sheet approaches (Large Banks in the out of Sample)

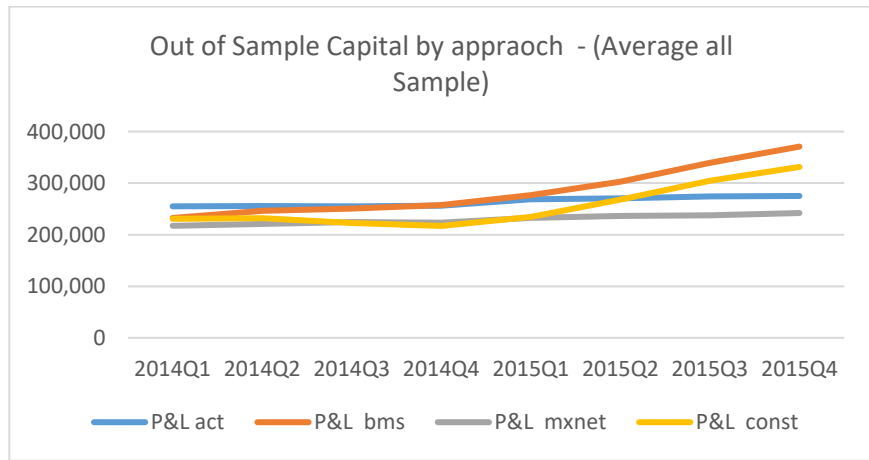


Figure 6: Out of sample back testing results of the Capital of the three balance sheet approaches compared with the actual figures (Whole Sample)

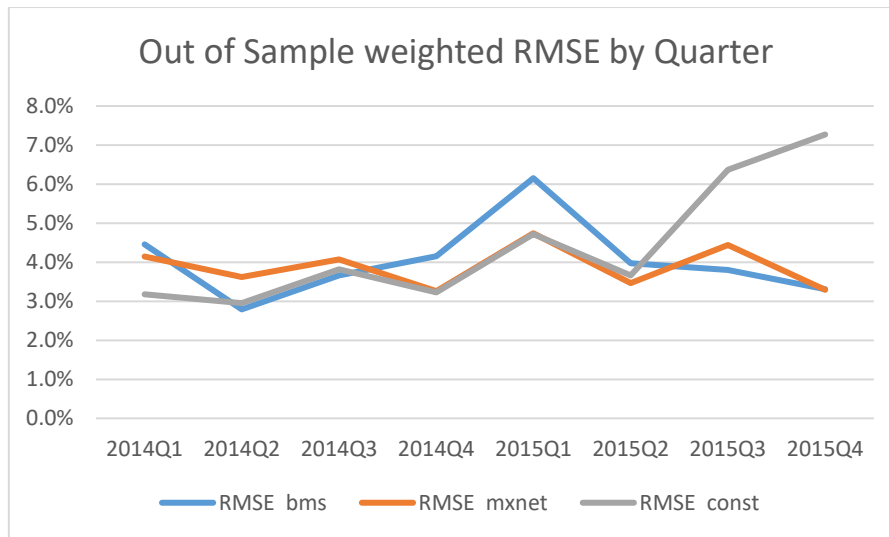


Figure 7: Out of sample back testing results of the RMSE of CAR by quarter of the three balance sheet approaches compared with the actual figures (Whole Sample)

All banks in the dataset	Out of Sample CAR	In Sample CAR
Satellite Modelling(BMA)	20.61	17.07
Deep Learning (MXNET)	18.01	17.89
Constant Balance Sheet	20.03	17.49
Actual	19.33	18.73
Large Banks (>200bl)		
Satellite Modelling(BMS)	15.07	11.04
Deep Learning (MXNET)	12.77	11.12
Constant Balance Sheet	15.11	11.48
Actual	13.75	14.16

Table 1: Comparison of the predicted one year ahead CAR by ST approach for all banks and only for Large financial institutions (more than 200billions in assets)

Table 1 summarizes the results of all aforementioned samples with respect the CAR ratio and the prediction error validation metrics (RMSE, MAE, MAPE). Based on the figures reported in the test sample MXNET provide more accurate estimation of the CAR ratio exhibiting a significant decrease in the forecasting error. Figure 7 provides a more detail evolution of the RMSE in the out of time dataset for the three approaches. The volatility and instability in the forecasts of the satellite modelling and the constant balance sheet approach are evident against the MXNET performance. To investigate further the superior performance of DeepStress we have graphed the evolution of the P&L against the other two approaches in figure 6.

All Banks (Assets Weighted)	Out of Sample (2014Q1 - 2015Q4)		
	RMSE	MAPE	MAE

Satelite Modelling(BMS)	8.28	15.15%	2.88
Deep Learning (MXNET)	7.23	11.93%	2.36
Constant Balance Sheet	7.88	15.25%	2.85
	In sample (2011Q1 - 2011Q4)		
	RMSE	MAPE	MAE
Satelite Modelling(BMS)	9.10	17.86%	2.63
Deep Learning (MXNET)	7.70	16.66%	2.55
Constant Balance Sheet	16.44	18.39%	2.79
	In sample (2010Q1 - 2013Q4)		
	RMSE	MAPE	MAE
Satelite Modelling(BMS)	9.16	15.79%	2.58
Deep Learning (MXNET)	9.70	15.00%	2.55
Constant Balance Sheet	11.15	15.40%	2.55

Table 2: Validation Measures – CAR for all financial institutions in the dataset

To further investigate the performance of Deep Stress approach we narrow down the results on a subset of large financial institutions where performance of a robust stress testing methodology is more important due to their size and social-economic impact. Big financial institutions are defined as entities with more than 200 billion in assets for the purpose of this study. Table 3 outlines the experimental results on all datasets. Additionally, in this group of financial institutions the superiority of deep neural network is confirmed with significant drops in the forecasting error in the test sample. Another worth mentioning results is the fact that although satellite univariate modelling in the sample dataset was expected to provide a better fitting against the DNN this is not the case. DNN is trained in a multivariate setup attempting to model 9 variables at the same time and still exhibits a rather comparable in sample error against the other two methods.

Large Banks (>200bl in assets)	Out of Sample (2014Q1 - 2015Q4)		
	RMSE	MAPE	MAE
Satelite Modelling(BMS)	3.04	17.05%	2.31
Deep Learning (MXNET)	2.23	14.66%	1.97
Constant Balance Sheet	3.25	19.05%	2.58
	In sample (2011Q1 - 2011Q4)		
	RMSE	MAPE	MAE
Satelite Modelling(BMS)	3.75	25.67%	3.59
Deep Learning (MXNET)	3.72	25.16%	3.52
Constant Balance Sheet	3.61	24.24%	3.41
	In sample (2010Q1 - 2013Q4)		
	RMSE	MAPE	MAE
Satelite Modelling(BMS)	3.48	23.39%	3.24
Deep Learning (MXNET)	3.50	23.25%	3.23
Constant Balance Sheet	3.28	22.19%	3.09

Table 3: Validation Measures – CAR only for large institutions in the dataset (more 200billinos in Assets)

Summarizing the results across all metrics in the test sample, it is evident that the MXNET system exhibits higher predicting power compared to all the considered benchmark approaches. Among the other two approaches it is evident that the constant balance assumption although easier to implement exhibits the highest error. Hence, it is crucial for supervisory authorities to rethink current stress testing exercise that are based on the constant balance sheet assumption and move towards a dynamic balance sheet approach. .

6. Conclusions and future work

In this study we propose a new approach to be utilized in regulatory stress testing exercises called Deep Stress leveraging on the properties of deep learning. The main novel contribution of this empirical research to the literature of forecasting economic and financial crisis events is that we explore this new statistical technique to tackle the problem of dynamic balance stress testing. Deep learning is utilized to provide a holistic modelling approach for a bank's key financial items. We perform thorough testing and validation of the proposed technique. Experimental results provide strong evidence to further be explored in the future by regulators and financial institutions in order to produce a new generation of stress testing. Deep stress is compared with two broadly accepted stress testing frameworks: constant balance sheet and satellite dynamic modelling.

Summarizing our experimental results, we have found that Deep Neural Networks built using the MXNET library consistently outperform the benchmark approaches. Our results provide strong evidence of increased forecasting accuracy with respect to the CAR ratio and performance consistency, which implies a much stronger generalization capacity compared to alternative benchmark frameworks. Specifically, Deep stress is compared with two broadly accepted stress testing frameworks: constant balance sheet and satellite dynamic modelling. Validation measures RMSE, MAE and MAPE significantly decrease in the test sample using DeepStress providing better simulation of the CAR ratio. Hence, these findings render our approach much more attractive to researchers and practitioners working in real-world financial institutions. The main driver for achieving this higher forecasting accuracy is the potential to model the balance sheet inter-correlation of P&L items providing better simulation of the banks one-year-ahead activities. Deep Stress offers a better dynamic balance sheet simulator which is a major component in any stress testing framework by better capturing that small macro and financial changes that can be amplified exponentially under a crisis event. The dynamic nature of our framework leads to significant decrease in the forecasting error by modelling better the feedback loops and the interdependence of various items of a financial institution balance sheet with the macro economy.

The aforementioned cascading layers structure of deep learning algorithms will open up new horizons for financial system simulation combining brain inspired computation and statistical machine learning. However, our initial endeavour is concentrated on the banking system which the backbone of the global economy but is scalable to other entities such as large corporate, insurances and shadow banking. The system can be used by policy makers to test various measures and to monitor the system in a forward looking manner and increase awareness for possible future financial shocks. DeepStress can finally be used to measure the social impact of a possible financial or systemic shock through the adjusted projections of various key macro variables like unemployment, wealth, credit expansion etc.

An aspect this work has not considered concerns developing deep learning models that can be continuously retrained in a moving window (online learning) setup. Another possible

way forward is the exploration of deep neural networks under broader dataset referring to multiple jurisdictions. Finally, it is evident that the postulated Deep Learning networks can effectively capture nonlinearities in the relationship between the input variable and the output variable. Although, the validation framework implemented in this study, cannot fully capture the estimation error in a Stress testing exercise due to the fact the dataset does not include crisis years. The results though provide evidence for the forecasting efficacy of DeepStress for several years simulating a baseline scenario of a Stress Testing exercise. To enhance the validation framework of our approach we will intensify the data collection process to gather information referring to several years before the financial crises in order to use DeepStress to simulate and predict failed entities that took place during this period. The value of such novel developments remains to be examined in our future research endeavours.

References

1. Borio, Claudio, Mathias Drehmann, and Kostas Tsatsaronis. "Stress-testing macro stress testing: does it live up to expectations?." *Journal of Financial Stability* 12 (2014): 3-15. - 3
2. Dees, Stéphane, and Jérôme Henry. "Stress-Test Analytics for Macroprudential Purposes: Introducing STAMP€." *SATELLITE MODELS* (2017): 13. 5
3. Anca Maria, Podpiera, Otker-Robe Inci, and Ã. Inci. "The social impact of financial crises: evidence from the global financial crisis." *Policy Research Working Paper Series* (2013). -38
4. EBA stress testing methodology 2018, <http://www.eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing/2018> 48
5. Greek Banking System Diagnostic Exercise 2013, <http://www.bankofgreece.gr/BogEkdoseis/2013%20Stress%20test%20of%20the%20Greek%20banking%20sector.pdf> 49
6. Henry, J., Kok, C., Amzallag, A., Baudino, P., Cabral, I., Grodzicki, M., ... & Pancaro, C. (2013). A macro stress testing framework for assessing systemic risks in the banking sector. ECB WP No. 152. 50
7. d'Angleterre, Banque. "A Framework for Stress Testing the UK Banking System." *Bank of England Discussion Paper (London)* (2013). 51
8. The Fed - Comprehensive Capital Analysis and Review - <https://www.federalreserve.gov/supervisionreg/ccar.htm> 52
9. Juselius, M and M Kim (2011): "Sustainable financial obligations and crisis cycles", *Helsinki Economic Centre of Research Discussion Papers* 313. 55
10. Alfaro, R and M Drehmann (2009): "Macro stress tests and crises: What can we learn?," *BIS Quarterly Review*, December, pp 29-41. 56
11. Drehmann, M, C Borio and K Tsatsaronis (2011b): "Characterising the financial cycle: don't lose sight of the medium term!", paper presented at the Reserve Bank of Chicago-ECB 14th Annual International Banking Conference, The role of central banks in financial stability: How has it changed?, Chicago, 10-11 November. 57
12. Oura, Hiroko, and Liliana B. Schumacher. "Macrofinancial stress testing-principles and practices." (2012). 65
13. Hirtle, Beverly, and Andreas Lehnert. "Supervisory stress tests." *Annual Review of Financial Economics* 7 (2015): 339-355. 67
14. Gounopoulos, Dimitrios, Johannes Höbelt, and Nikolaos I. Papanikolaou. "Bank Stress Tests: An Active Treatment or a Placebo?." (2016). 68

15. Bookstaber, Rick, et al. "Stress tests to promote financial stability: Assessing progress and looking to the future." *Journal of Risk Management in Financial Institutions* 7.1 (2014): 16-25. 79
16. Acharya, Viral, Robert Engle, and Diane Pierret. "Testing macroprudential stress tests: The risk of regulatory risk weights." *Journal of Monetary Economics* 65 (2014): 36-53. 75
17. Chatzis, Sotirios P., et al. "Forecasting stock market crisis events using deep and statistical machine learning techniques." *Expert Systems with Applications* 112 (2018): 353-371. -77
18. Fischer, Thomas, and Christopher Krauss. "Deep learning with long short-term memory networks for financial market predictions." *European Journal of Operational Research* 270.2 (2018): 654-669. -78
19. Kraus, Mathias, and Stefan Feuerriegel. "Decision support from financial disclosures with deep neural networks and transfer learning." *Decision Support Systems* 104 (2017): 38-48. -79
20. Castro Carvalho, A., et al. "Proportionality in banking regulation: a cross-country comparison." *FSI Insights on policy implementation* 1 (2017). -80
21. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014), "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," *Proc. EMNLP*. 81
22. Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. Modeling coverage for neural machine translation. *Proc. ACL* (2016). 82
23. Quadrana, M., Hidasi, B., Karatzoglou, A. and Cremonesi, P. (2017), Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks, *Proc. ASM RecSys*. 83
24. Srivastava, N, Hinton, J., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014) 1929-1958. 84
25. Ley, E. and M. Steel (2008): On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regressions. Working paper 85
26. Fernandez, C. E. Ley and M. Steel (2001): Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100(2), 381–427 86
27. Feldkircher, Martin, et al. "ARNIE in Action: the 2013 FsAp stress tests for the Austrian banking system." *Financial stability report* 26 (2013): 100-118.
28. Anand, Kartik, Guillaume Bédard-Pagé, and Virginie Traclet. "Stress testing the Canadian banking System: a System-wide approach." *Financial System Review* 61 (2014).
29. Hasan, Maher Mohamad, Christian Schmieder, and Claus Pühr. "Next Generation Balance Sheet Stress Testing." (2011).
30. Bookstaber, Rick, et al. "Stress tests to promote financial stability: Assessing progress and looking to the future." *Journal of Risk Management in Financial Institutions* 7.1 (2014): 16-25.

31. Drehmann, M, A Patton and S Sorensen (2007): "Non-linearities and stress testing", in Risk measurement and systemic risk, Proceedings of the fourth joint central bank research conference, ECB.
32. BIS stats. <http://stats.bis.org/bis-stats-tool/org.bis.stats.ui.StatsApplication/StatsApplication.html>
33. Ferri, Giovanni, and Valerio Pesic. "Bank regulatory arbitrage via risk weighted assets dispersion." *Journal of Financial Stability*(2016).
34. Hirtle, Beverly, and Andreas Lehnert. "Supervisory stress tests." *Annual Review of Financial Economics* 7 (2015): 339-355.
35. Gersl, Adam, and Jakub Seidler. "How to improve the quality of stress tests through backtesting." *Finance a Uver* 62.4 (2012): 325.

Appendix: Prior selection in BMA models

It is a popular choice to set a uniform prior probability for each model to represent the lack of prior knowledge. It is often the case in BMA to assume no prior knowledge for each model and assign a uniform prior probability i.e. $p(M_\gamma) \propto 1$. Regarding the marginal likelihoods $p(M_\gamma|Y, X)$ and the posterior distributions $p(\beta|M_\gamma, Y, X)$ the literature standard is to use a specific prior structure called Zellner's g prior in order to estimate posterior distributions in an efficient mathematical way. In this setup the prior knowledge for the coefficients is assumed to be a normal distribution with pre-specified mean and variance. Specifically the parametric formulation is given by (8).

$$\beta_\gamma | g \sim N\left(0, \sigma^2 \left(\frac{1}{g} X_\gamma' X_\gamma\right)^{-1}\right) \quad (8)$$

According to (8) coefficients are assumed to have zero mean and a variance-covariance structure which is broadly in line with that of the data X_γ . The hyper-parameter g denotes the prior level of confidence that the coefficients are zero. The posterior distribution of the coefficients follows a t-distribution with expected value $\frac{g}{1+g} \widehat{\beta}_\gamma$ where $\widehat{\beta}_\gamma$ is the standard OLS estimator for model γ . Thus as $g \rightarrow \infty$ the coefficient estimator approaches the OLS estimator. Similarly, the posterior variance of β_γ is affected by the value of g (9).

$$\text{Cov}(b_\gamma | Y, X, g, M_\gamma) = \frac{(Y - \bar{Y})(Y - \bar{Y})'}{N-3} \frac{g}{1+g} \left(1 - \frac{g}{1+g} R_\gamma^2\right) (X_\gamma' X_\gamma)^{-1} \quad (9)$$

The posterior covariance is similar to that of the OLS estimator, times a factor that includes g and R_γ^2 (OLS R squared for model γ). For BMA, this prior framework results in α marginal likelihood which includes a size penalty factor adjusting for model size k_γ given by

$$p(Y|M_\gamma, X, g) \propto (Y - \bar{Y})'(Y - \bar{Y})^{-\frac{N-1}{2}} (1+g)^{-\frac{k_\gamma}{2}} \left(1 - \frac{g}{1+g}\right)^{-\frac{N-1}{2}} \quad (10)$$

The "default" approach for hyper-parameter g is the "unit information prior" (UIP), which sets $g = N$ for all models.