Combining Deep Learning and GARCH Models for Financial Volatility and Risk Forecasting

Jakub Michańków,
Krakow University of Economics, Krakow, Poland,
University of Warsaw, Warsaw, Poland *j.michankow@uw.edu.pl*

Łukasz Kwiatkowski,
Krakow University of Economics,
Krakow, Poland *kwiatkol@uek.krakow.pl*

Janusz Morajda,
Krakow University of Economics,
Krakow, Poland *morajdaj@uek.krakow.pl*

Abstract

We develop a hybrid approach to forecasting the volatility and risk of financial instruments by combining econometric GARCH models with deep learning networks. For the latter, we employ Gated Recurrent Unit (GRU) networks, whereas four different specifications are used as the GARCH component: standard GARCH, EGARCH, GJR-GARCH and APARCH. Models are tested using daily returns on the S&P 500 index and Bitcoin prices. As the main volatility estimator, and the target function of our hybrid models, we use the modified Garman-Klass estimator. Volatility forecasts resulting from the hybrid models are employed to evaluate the assets' risk using the Value-at-Risk (VaR) and Expected Shortfall (ES). Gains from combining the GARCH and GRU approaches are discussed in the contexts of both the volatility and risk forecasts. It can be concluded that the hybrid solutions produce more accurate point volatility forecasts, although it does not necessarily translate into superior risk forecasts.

Keywords: neural networks, GRU networks, financial time series, Value-at-Risk, Expected Shortfall.

1. Introduction and literature review

Measuring and predicting volatility and investment risk of financial assets are perennial problems of great importance both for scientists and practitioners, with the relevant literature abounding in model specifications and quantitative methods designed to address the tasks. Currently the most common approach has been developed within the area of financial econometrics, where the prices of financial instruments are typically assumed to form some conditionally heteroscedastic stochastic processes, the exact specification of which, along with their estimation and statistical inference, constitute a key part of the researchers' endeavours (see, e.g., [32]). A basic group of this type of tools for modelling and forecasting volatility (and, consequently, risk) are the GARCH models developed by [1] & [30], generalising the ARCH specification proposed by [8]. The input information in the GARCH models, driving current volatility, comprises primarily the past return rates and their conditional variances. Voluminous subsequent research aimed at modifications and extensions of the original GARCH structure, also by admitting various types of the conditional distribution. This resulted in a considerable diversity of the GARCH class, with EGARCH ([28]), APARCH ([6]), GJR-GARCH ([15]), and TGARCH ([35]) being among the most widely recognised.

A parallel trend in financial time series modelling and forecasting follows the development of machine learning tools, particularly artificial neural networks (ANNs). These models, often treated as "black boxes", are regarded as nonlinear and nonparametric techniques in which no *a priori* assumption concerning the mathematical form (equation) of the model is formulated. The function mapping input data into output signals (forecasts) is formed at the stage of training

the model, implemented on the basis of a learning set including historical quotations. Over recent years, both researchers and practitioners have increasingly been using dynamic ANNs equipped with the ability to remember and process information from some recent period of time. These tools include mainly deep-learning-based recurrent neural networks (RNNs; introduced by [19], and further developed by [33]), in particular Long Short-Term Memory networks (LSTM; [18]), and also (utilised in the presented research) Gated Recurrent Unit (GRU) neural networks ([2]), which constitute simplified modifications of LSTM.

Quite recently, a new promising research trend has emerged (including also our present paper), in which attempts are made to integrate formal tools based on the GARCH methodology with currently developed neural models based on deep learning with memorising the dynamics of the analysed phenomenon. Research on this type of hybrid models has been undertaken in many works. In particular, to cite only the most pertaining to the current paper, Kristjanpoller & Minutolo have applied hybrid models (based on feed-forward back-propagation neural network and GARCH) to predict the volatility of gold ([22]) and oil prices ([23]). [20] developed a hybrid deep learning method combining GARCH with LSTM neural networks and applied it to forecasting the volatility of copper price. Finally, in ([25]), a GARCH model was incorporated into an LSTM network for improving the prediction of stock volatility.

To the best of our knowledge, no attempts have been made so far to merge GARCH structures with GRU neural networks (particularly, for the purpose of financial modelling). To fill this gap, it is the main objective of this paper to introduce a new tool for financial assets volatility and risk prediction by combining the two approaches, with the resulting hybrid specification referred to as GARCH-GRU, henceforth. Next, using data concerning S&P500 and Bitcoin, we analyse the predictive effectiveness of the GARCH-GRU models in comparison to 'pure' GARCH models, mainly to examine the synergistic benefits of the former. The focus is not only on the point volatility forecasts, typically analysed in the literature, but also on forecasting financial risk, measured by Value at Risk (VaR) and Expected Shortfall (ES).

The remainder of the paper is organised as follows. Section 2 presents the methodological framework, with Subsections 2.1 and 2.2 describing a brief theoretical background of existing methods concerning GARCH models and GRU networks, respectively. Subsection 2.3 introduces our proposition of the GARCH-GRU model, along with its parameter settings, whereas Subsection 2.4 outlines the ex post evaluation framework of the volatility and VaR forecasts. Section 3 is devoted to the empirical analysis, starting with statistical description and pre-processing of the data sets under study (Subsection 3.1). Subsections 3.2 and 3.3 cover the empirical results and their analyses for S&P500 and Bitcoin, respectively. Finally, Section 4 concludes.

2. Methodology

Below we briefly present the methodological framework underlying our study, and combining popular and widely recognised GARCH models (a selection of which is briefly outlined in Subsection 2.1) and GRU neural networks (see [2]; Subsection 2.2). The 'merge' results in a novel GARCH-GRU hybrid specification, presented in Subsection 2.3. We close this section with a brief description of the ex post volatility and VaR forecast accuracy measures employed in our work (Subsection 2.4).

2.1 GARCH models

Let $r_t = 100\ln(P_t/P_{t-1})$ denotes the logarithmic rate of return on some asset at time *t*, with P_t and P_{t-1} standing for the instrument's prices at time t and $t-1$ respectively. Let us then consider a model of the form: $r_t = E(r_t|\psi_{t-1}) + \varepsilon_t$, combining the conditional mean of the returns (given the past information, ψ_{t-1}) and an error term ε_t defined as:

$$\varepsilon_t = z_t\sigma_t, \tag{1}$$

where random variables $z_t \sim iiD(0,1)$ form a sequence of independent and identically distributed standardised errors (with zero mean and unit variance), and $\sigma_t = \sqrt{Var(r_t|\psi_{t-1})}$ is the return's conditional standard deviation, usually referred to as the volatility.

In our research, for modelling the conditional variance σ_t^2, we employ four most commonly entertained in the extant literature GARCH specifications: a 'standard' GARCH ([1]), the Glosten-Jagannathan-Runkle GARCH (GJR-GARCH; [15]), the Exponential GARCH (EGARCH; [28]), and the Asymmetric Power ARCH (APARCH; [6]). Below, we briefly present their specific volatility equations (a detailed and comprehensive review of univariate GARCH model specifications can be found, e.g., in [11] and [31]).

GARCH

The volatility equation takes the form:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{2}$$

where σ_t^2 is the conditional variance at time *t*, and the parameters are subject to restrictions ensuring positive σ_t: $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, \dots, q$, and $\beta_j \geq 0$ for $j = 1, \dots, p$.

GJR-GARCH

The volatility equation takes the form:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} (\alpha_i + \omega_i I_{t-i}) \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{3}$$

where $I_{t-i} = 1$ when $\varepsilon_{t-i} \geq 0$, and $I_{t-i} = 0$ otherwise. Additionally, $\alpha_0 > 0$, $\alpha_i \geq 0$ and $\omega_i \geq 0$ for $i = 1, \dots, q$, and $\beta_j \geq 0$ for $j = 1, \dots, p$.

EGARCH

The volatility equation takes the form:

$$\ln \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \{\theta z_{t-i} + \gamma [|z_{t-i}| - E(|z_{t-i}|)]\} + \sum_{j=1}^{p} \beta_j \ln \sigma_{t-j}^2, \tag{4}$$

where $\alpha_1 \equiv 1$ for the identification of the model.

APARCH

The volatility equation takes the form:

$$\sigma_t^{\delta} = \alpha_0 + \sum_{i=1}^{q} \alpha_i [|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i}]^{\delta} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\delta}, \tag{5}$$

where $\delta > 0$, $\alpha_0 > 0$, $\alpha_i \geq 0$ and $-1 < \gamma_i < 1$ for $i = 1, \dots, q$, and $\beta_j \geq 0$ for $j = 1, \dots, p$

Three types of conditional distributions, most commonly entertained in the literature, are used in this study for the standardised error term, z_t: the normal distribution, the symmetric Student's *t*-distribution and the skewed Student's *t*-distribution (see [13]).

Estimation (through the maximum likelihood approach) and forecasting in the GARCH models have been implemented in numerous libraries available in the R programming environment, among which the `rugarch` package (see [13], [14]) appears one of the most popular and comprehensive, and is also employed in this work.

2.2 GRU neural networks

The GRU neural networks, introduced by Chung *et al*. in [2], constitute simplified versions of more popular LSTM networks. The GRU networks use a single unit to forget the information or update the network state, which allows them to achieve similar results to LSTMs, while significantly reducing the training time.

The functions within a GRU network cell can be described by the following equations:

$$d_t = \varsigma_g(W_d x_t + U_d o_{t-1} + b_d), \tag{6}$$

$$s_t = \varsigma_g(W_s x_t + U_s o_{t-1} + b_s), \tag{7}$$

$$\widehat{o_t} = \phi_o(W_o x_t + U_o(s_t \odot o_{t-1}) + b_o), \tag{8}$$

$$o_t = (1 - d_t) \odot o_{t-1} + d_t \odot \widehat{o_t}\,, \tag{9}$$

where x_t is the input vector, o_t is the output vector, while d_t and s_t are the update gate and the reset gate vectors, respectively. The matrices W and U (subscripted according to pertinent equations) as well as the vectors b comprise the net's parameters, whereas ς_g and ϕ_0 are sigmoid and hyperbolic tangent activation functions, respectively. Finally, $\odot$ denotes the Hadamard product. For a detailed description of the GRU networks and a comparison with other types of recurrent networks we refer the reader to [16].

2.3 A proposition of a GARCH-GRU model and its parameters settings

In this paper we propose to combine the above-presented GARCH models (Subsection 2.1) and the GRU neural networks (Subsection 2.2). We name the hybrid structure created from this conjunction as the GARCH-GRU models.

The main idea of the approach is the incorporation of the volatility forecasts derived from a given GARCH model as an input variable to a GRU network. Additionally, to potentially further improve the hybrid model's performance, the GRU component is fed with two more inputs: absolute log returns and additional (out-of-GARCH) volatility estimates obtained by means of the Garman & Klass ([12]) estimator modified by Yang & Zhang ([34]) to account for the gap between the previous day's closing and the current day's opening prices. Specifically, the variance estimate (denoted as $\sigma^{2,GKYZ}$) is given by the formula:

$$\sigma^{2,GKYZ} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\ln\frac{O_i}{C_{i-1}}\right)^2 + \frac{1}{2}\left(\ln\frac{H_i}{L_i}\right)^2 - (2\ln 2 - 1)\left(\ln\frac{C_i}{O_i}\right)^2\right], \tag{10}$$

where O_i, H_i, L_i and C_i, respectively, denote the opening, the highest, the lowest and the closing price at time i, and n denotes the number of daily log returns used to calculate the estimate (we set $n = 10$). Additionally, we scale the estimator to match the magnitude of the volatility estimates with the ones retrieved from GARCH models. To that end, the following formula proposed by Fiszeder ([9], [10]) is employed:

$$Scaled\ \sigma^{2,GKYZ} = \frac{a}{b}\sigma^{2,GKYZ}, \tag{11}$$

$$a = \frac{1}{T}\sum_{t=1}^{T} {r_t}^2,\ b = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^{2,GKYZ} \tag{12}$$

where *T* denotes the initial sample size used for the estimation of a GARCH model.

Specific architecture of the GRU component used in this research consists of three GRU-type layers with 512/256/128 neurons and one single neuron dense layer on the output. Each of these GRU layers uses ReLU (Rectified Linear Unit) activation function, a dropout regulariser set to 0.3, and *l2* kernel regulariser set to 0.00001, which allows to select the best MSE value based on the validation set loss from all the epochs.

For the network optimisation, we use the Adam optimiser ([21]), with the learning rate set to 0.0009. The network component is trained the GKYZ volatility estimates at time *t* + 1, with the loss function defined as the mean square error between the volatility estimates and the network output (volatility predictions). Datasets feeded into the network are divided into mini-batches, with the size of 500 data points, while each batch is divided into sequences of 6 days based on which a single day prediction is produced. The tuning process is performed by means of the KersTuner with Hyperband algorithm ([29]). Finally, the model is trained for 150 epochs, with a model checkpoint callback function using the lowest value of the loss that occurred during the training.

2.4 Ex post volatility and VaR forecasts evaluation

The ex post assessment of the GARCH and GARCH-GRU models' predictive performance is carried out here with respect to the two: point volatility forecasts and risk forecasts.

The point volatility predictions' accuracy is measured by three standard forecast error metrics: mean squared error (MSE), mean absolute error (MAE), and heteroscedasticity-adjusted (HMSE). Differences between the MSEs for some two competing models are tested for their statistical significance *via* the Diebold & Mariano ([5]) test, with a modification proposed by Harvey, Leybourne & Newbold ([17]). In our setting, we focus on comparing the

MSE obtained for a given GARCH model with the one produced by a corresponding GARCH-GRU specification. The null hypothesis states that both of the MSE values are equal, while the alternative – that the hybrid model is more accurate than the 'pure' GARCH structure. Finally, the correlation between the conditional variance forecasts ($\sigma_{t+1}^{2,f}$) and corresponding GKYZ volatilities ($\sigma_{t+1}^{2,GKYZ}$) is assessed by the coefficient of determination from the Mincer & Zarnowitz ([27]) regression:

$$\sigma_{t+1}^{2,GKYZ} = \beta_0 + \beta_1 \sigma_{t+1}^{2,f} + \xi_{t+1}, \tag{13}$$

with ξ_{t+1} denoting an error term.

The second aspect of the models' predictive evaluation in this paper is risk forecasts accuracy. To that end, volatility predictions resulting from GARCH and GARCH-GRU models are used to produce long position Value at Risk (VaR) and Expected Shortfall (ES) forecasts:

$$VaR_{t+1}(\alpha) = -r_{t+1}^{f} - \sigma_{t+1}^{f} q_{\alpha}^{z}, \tag{14}$$

$$ES_{t+1}(\alpha) = E(r_{t+1} | r_{t+1} < VaR_{t+1}(\alpha)) = r_{t+1}^{f} + \sigma_{t+1}^{f} E(z_t | z_t < q_{\alpha}^{z}), \tag{15}$$

where α denotes the tolerance probability level, r_{t+1}^{f} and σ_{t+1}^{f} are, respectively, the forecasted return and volatility at time t+1, and finally, q_{α}^{z} denotes the α^{th}-quantile of the distribution assumed for z_t (see [7], [26]). Notice that the return predictions, r_{t+1}^{f}, are generated in our paper only through the underlying GARCH model, and thereby are not further processed through the hybrid GARCH-GRU structure. This limitation is intended here to ensure that any differences between the VaR and ES predictions stemming from GARCH and GARCH-GRU models remain attributable solely to the accuracy of the volatility forecasts produced by the two approaches. Analysis of potential further gains from using return predictions from the hybrid model instead, remains beyond the scope of the current research.

Backtesting of VaR and ES forecasts is performed by means of standard tools. For the former, we use two procedures. First, the Kupiec ([24]) test is employed to examine the unconditional coverage property (stated by the null hypothesis), that is the consistency between the empirical VaR hit ratio and the assumed tolerance level. Both significantly higher and lower number of VaR exceedances (or, violations) cause the null to be rejected. Second, through the conditional coverage test by Christoffersen ([3], [4]) test we check whether the VaR hits are independent (thus do not occur in clusters) and the empirical VaR hit ratio coincides with the assumed tolerance probability. The two statements jointly form the null hypothesis.

To backtest the Expected Shortfall predictions, we resort to the McNeil and Fray ([26]) test, with the null assuming that the mean of the ES exceedances equals zero. The test results are reported in two variants: one under the exact distribution of the test statistics, and the other using a bootstrapped distribution. The latter accounts for a possible misspecification of the underlying distribution of the standardised residuals.

3. Empirical analysis

3.1 Data

Two data sets of daily logarithmic rates of return are analysed in our research, each representing quite a distinct type of financial assets: the S&P 500 index (5-day week, quotations over 6 April 2009 to 31 December 2020) and Bitcoin (BTC/USD; 7-day week, quotations over 5 August 2013 to 31 December 2020). The time ranges of the data sets ensure an equal number of 2707 observations in each case.

For the forecasting evaluation of the models (presented in the following subsection), all the data sets are divided in such a manner as to ensure the same amounts of observations for each asset at corresponding stages of analysis. Specifically, a rolling window scheme is employed for both GARCH and hybrid GARCH-GRU models. The size of the rolling window for the GARCH models is set to 504 days (the models are re-estimated upon arrival of each new observation). The GARCH component order is fixed to p=q=1 (see Eqs. 2-5) for all the assets (as typically done in the empirical literature), while the ARMA component is reduced to AR(1) for S&P 500, and only a constant for Bitcoin, with the choices supported by a preliminary

analysis employing the Bayesian information criterion (the results left unreported for the sake of brevity).

For the neural network stage, the data is divided into a series of rolling training sets, each of 1008 observations, and a series of rolling test sets, each comprising 504 observations. Each time, 33% of the training set (336 observations) is used for a validation set.

The total number of the ex post evaluated predictions obtained from the GARCH and hybrid models is 1194, and is the same for each asset (although the corresponding time ranges vary: 7 April 2016 to 31 December 2020 for S&P 500 and 29 September 2017 to 31 December 2020 for Bitcoin). Sample sizes were based on initial hyperparameter tuning.

Empirical results obtained for each of the two assets are discussed below (Subsection 3.2 and 3.3) in the following fashion. First, we compare the GARCH and GARCH-GRU models in terms of MSE, along with testing its values through the Diebold-Mariano test, with low p-values favouring the hybrid model (Tables 1 and 4). Then, results for VaR exceedances are presented (Tables 2 and 5). Finally, based on the previous, a selection of the best performing models is analysed in more detail, both with respect to the overall volatility forecast accuracy and risk prediction (Tables 3 and 6), the latter including 1% and 5% Value at Risks as well as 5% Expected Shortfall.

3.2 Results for S&P 500

Table 1 indicates that the best performing (in terms of MSE) is the EGARCH-GRU model with a skewed Student's *t*-distribution, although only by a rather narrow margin as compared with some other specifications, like GJR-GARCH-GRU with either a symmetric or skewed *t*-distribution, and even 'standard' GARCH-GRU with a skewed *t*-distribution. Overall, the results presented in Table 1 imply unanimously that combining GARCH models with GRU networks significantly enhances the forecast accuracy, with all of the Diebold-Mariano test p-values remaining below 0.05.

Table 1. Comparison of volatility forecasts in terms of MSE across all models and distributions for S&P 500.

Metrics / Model	G(N)	G(N)-GRU	G(STD)	G(STD)-GRU	G(SSTD)	G(SSTD)-GRU
MSE	0.0844	0.0168	0.0646	0.0183	0.0575	0.0159
DM p-value	0.0379		0.0494		0.0189	
Metrics / Model	E(N)	E(N)-GRU	E(STD)	E(STD)-GRU	E(SSTD)	E(SSTD)-GRU
MSE	0.1539	0.0167	0.1362	0.0160	0.1350	**0.0152**
DM p-value	0.0018		0.0025		0.0046	
Metrics / Model	GJR(N)	GJR(N)-GRU	GJR(STD)	GJR(STD)-GRU	GJR(SSTD)	GJR(SSTD)-GRU
MSE	0.1388	0.0181	0.1244	0.0153	0.1053	0.0156
DM p-value	0.0347		0.0395		0.0290	
Metrics / Model	AP(N)	AP(N)-GRU	AP(STD)	AP(STD)-GRU	AP(SSTD)	AP(SSTD)-GRU
MSE	0.1103	0.0333	0.1015	0.0253	0.0890	0.0221
DM p-value	0.0237		0.0160		0.0009	

Note: G stands for GARCH, E for EGARCH, GJR for GJR-GARCH, AP for APARCH. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. DM denotes the Diebold-Mariano test. The best result according to MSE is given in bold.

Next, we compare the models in terms of VaR unconditional coverage, with Table 2 presenting the actual number of VaR exceedances and hit ratios (in percentage terms) for both VaR tolerance levels under consideration, i.e. 5% and 1%. The results indicate that the most accurate VaR hit coverage is attained by the GJR-GARCH-GRU models with a normal and a skewed Student's *t*-distribution for the 5% tolerance level, and the APARCH model with a skewed Student's *t*-distribution for the 1% tolerance (paths of the 5% and 1% VaR forecasts along with their violations are displayed in Figure 1). Overall, and contrary to Table 1, the results here provide only mixed conclusions as to gains from the hybrid models, since

combining GARCH with the GRU networks does not necessarily bring the VaR hit ratios closer to the expected 5% and 1% tolerance levels.

Table 2. Number of VaR exceedances (and VaR hit ratios) across all models for S&P 500.

VaR / Model	G(N)	G(N)-GRU	G(STD)	G(STD)-GRU	G(SSTD)	G(SSTD)-GRU
VaR 5%	69 (5.77%)	62 (5.19%)	79 (6.61%)	72 (6.03%)	74 (6.19%)	63 (5.27%)
VaR 1%	31 (2.59%)	31 (2.59%)	21 (1.75%)	22 (1.84%)	19 (1.59%)	21 (1.75%)
VaR / Model	E(N)	E(N)-GRU	E(STD)	E(STD)-GRU	E(SSTD)	E(SSTD)-GRU
VaR 5%	78 6.53%)	58 (4.85%)	90 (7.53%)	63 (5.27%)	78 (6.53%)	56 (4.69%)
VaR 1%	35 (2.93%)	32 (2.68%)	24 (2.01%)	25 (2.09%)	17 (1.42%)	21 (1.75%)
VaR / Model	GJR(N)	GJR(N)-GRU	GJR(STD)	GJR(STD)-GRU	GJR(SSTD)	GJR(SSTD)-GRU
VaR 5%	65 (5.44%)	**59 (4.94%)**	75 (6.28%)	66 (5.52%)	68 (5.69%)	**59 (4.94%)**
VaR 1%	28 (2.34%)	34 (2.84)	20 (1.67%)	26 (2.17%)	18 (1.50%)	22 (1.84%)
VaR / Model	AP(N)	AP(N)-GRU	AP(STD)	AP(STD)-GRU	AP(SSTD)	AP(SSTD)-GRU
VaR 5%	76 (6.36%)	58 (6.70%)	82 (6.86%)	65 (5.44%)	74 (6.19%)	53 (4.43%)
VaR 1%	34 (2.84%)	32 (2.68%)	22 (1.84%)	24 (2.01%)	**16** (**1.34**%)	22 (1.84%)

Note: G stands for GARCH, E for EGARCH, GJR for GJR-GARCH, AP for APARCH. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. Expected number of VaR exceedances, corresponding to 5% and 1% tolerance levels, are equal to 59 and 12, respectively. The best outcomes are indicated in bold.

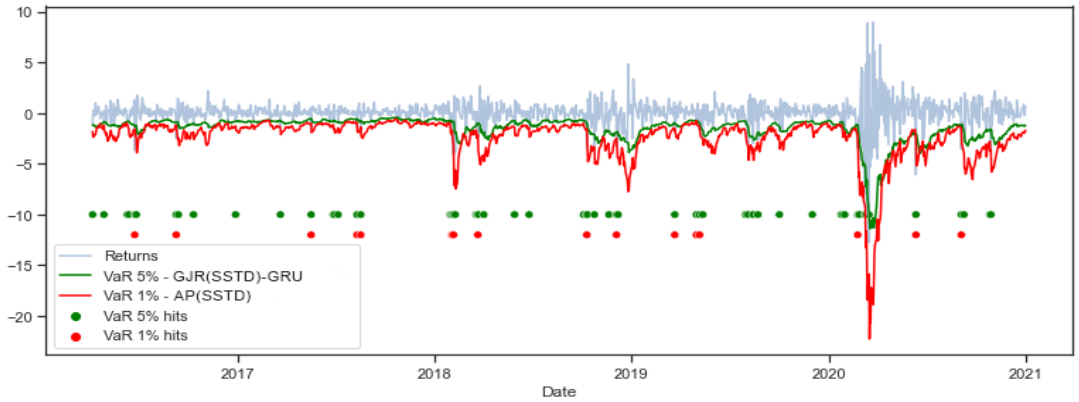

Fig. 1. VaR forecasts for S&P 500.

Further, in Table 3, we analyse in more detail a selection of four models that proved superior according to MSE and/or VaR exceedances: EGARCH-GRU with a skewed *t*-distribution (minimising all the three criteria for the volatility point forecasts: MSE, MAE and HMSE), GJR-GARCH-GRU with a normal and a skewed *t*-distribution (both ensuring the ideal VaR hit ratio at 5% tolerance; the latter model additionally yielding the highest R^2 in the Mincer-Zarnowitz regression), and APARCH with a skewed *t*-distribution (featuring the best, although not ideal, unconditional VaR coverage at the 1% tolerance level). The results indicate that the two GJR-GARCH-GRU models pass the unconditional coverage test for the 5% tolerance level, but rather fail the conditional coverage test, thus implying some clustering of the VaR violations (as might have already been expected from Figure 1, to some extent). In addition, and to one's dismay, these models also fail the ES backtest.

On the other hand, the APARCH model with a skewed Student's *t*-distribution, preferred in terms of the 1% VaR prediction, performs well in all three tests. Nevertheless, the model's performance for the 5% VaR tolerance level is clearly surpassed by the other specifications.

Table 3. Detailed comparison of the best performing models for S&P 500.

Metrics / Model	E(SSTD)-GRU	GJR(N)-GRU	GJR(SSTD)-GRU	AP(SSTD)
MSE	**0.0152**	0.0181	0.0156	0.0890
MAE	**0.0787**	0.0842	0.0813	0.1919
HMSE	**0.0128**	0.0142	0.0134	0.0550
R^2	0.9736	0.9563	**0.9806**	0.8465
VaR exceedances: 5%/1%	56/21	**59**/34	**59**/22	74/**16**
VaR hit ratio: 5%/1%	4.69%/1.75%	4.94%/2.84%	4.94%/1.84%	6.19%/1.34%
Kupiec p-value: 5%/1%	0.6197(F)/0.0173(R)	0.9258(F)/1.6e-07(R)	0.9258(F)/0.0088(R)	0.0667(F)/0.2616(F)
Christof. p-value: 5%/1%	0.0130(R)/0.0010(R)	0.0017(R)/9.9e-09(R)	0.0279(R)/0.0007(R)	0.0555(F)/0.2408(F)
ES p-value bootstr./sample	0.0551(F)/0.0155(R)	4.9e-05(R)/1.0e-06(R)	0.0463(R)/0.0126(R)	0.4351(F)/0.3791(F)

Note: VaR 5% and VaR 1% stands for 5% and 1% tolerance level. Expected number of VaR exceedances, corresponding to 5% and 1% tolerance levels, are equal to 59 and 12, respectively. F means the test failed to reject the H0 at 5% significance level, R means the H0 was rejected. G stands for GARCH model, E for EGARCH model, GJR for GJR-GARCH model, AP for APARCH model. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. R^2 is the coefficient of determination. The best (across the models) outcome according to a given metric is given in bold.

3.3 Results for Bitcoin

As indicated by Table 4, the best performing model for Bitcoin (in terms of point volatility forecasts) is the conditionally normal APARCH-GRU structure, with four other hybrid specifications being close seconds: GARCH-GRU and GJR-GARCH-GRU, with both symmetric and skewed *t*-distributions. Incidentally, the result may imply that simpler GARCH specifications, constituting some special cases of APARCH, may require more sophisticated, heavy-tailed conditional distributions to offset their simpler volatility structure.

Overall, and similar to the case of S&P 500, combining GARCH and GRU models largely improves the point volatility forecasts. Enhancing a GARCH model with a GRU network results in a one- or two-order-of-magnitude drop in MSE, even though in some cases the difference appears either statistically insignificant (conditionally normal EGARCH vs. EGARCH-GRU) or at least not as statistically significant as one could expect (conditionally *t*-distributed EGARCH vs. EGARCH-GRU, and APARCH vs. APARCH-GRU with a skewed *t*-distribution). However surprising or aberrant these results may appear, they remain largely attributable to a single erratically high (compared to the target GKYZ estimates) volatility forecast obtained from the above-mentioned 'sheer' GARCH models, linked to the COVID-19 pandemic outbreak. Conceivably, this volatility over-prediction is due to additional reverse transformations required to calculate the forecast of conditional standard deviation from the volatility equation defined inherently either for the logarithm of the variance (as in EGARCH; see Eq. 4) or some power transformation thereof (as in APARCH; see Eq. 5). Ultimately, these discrepancies between the 'sheer' and hybrid EGARCH (and APARCH) volatility forecasts lead to an overly high long-run variance estimate underlying the Diebold-Mariano (DM) test, which dwindles the test statistics value, thus increasing the p-value.

Table 5 presents the number and hit ratios of VaR exceedances. To one's dismay, and similar to the S&P 500 case, we notice that despite the earlier results indicating a considerable gain from combining GARCH with GRU models for the sake of volatility forecasting, the effect does not necessarily translate into superior VaR prediction performance of the hybrid structures. For both of the tolerance levels under consideration, it is a 'sheer' APARCH model that produces VaR estimates with a hit ratio nearest to the expected one: the conditionally normal APARCH for the 5% tolerance level, and APARCH with a *t*-distribution for the 1% tolerance (see Figure 2). Overall, as inferred from Table 5, combining GARCH with GRU models may lead to either more conservative or more liberal VaR predictions, as compared with the ones from the underlying 'pure' GARCH structures.

Table 4. Comparison of volatility forecasts in terms of MSE across all models and distributions for Bitcoin.

Metrics / Model	G(N)	G(N)-GRU	G(STD)	G(STD)-GRU	G(SSTD)	G(SSTD)-GRU
MSE	5.6795	0.4042	5.0206	0.3938	5.0327	0.3937
DM p-value	2.537e-05		4.399e-16		4.233e-16	
Metrics / Model	E(N)	E(N)-GRU	E(STD)	E(STD)-GRU	E(SSTD)	E(SSTD)-GRU
MSE	7.4872	0.4406	11.3832	0.4698	62.5108	0.5980
DM p-value	0.1355		0.006316		1.429e-13	
Metrics / Model	GJR(N)	GJR(N)-GRU	GJR(STD)	GJR(STD)-GRU	GJR(SSTD)	GJR(SSTD)-GRU
MSE	5.8335	0.4202	5.0074	0.3946	5.0188	0.3982
DM p-value	5.838e-05		1.844e-15		1.8e-15	
Metrics / Model	AP(N)	AP(N)-GRU	AP(STD)	AP(STD)-GRU	AP(SSTD)	AP(SSTD)-GRU
MSE	5.8356	**0.3818**	3.7345	0.4308	12.9127	0.4184
DM p-value	0.000837		3.167e-08		0.01944	

Note: G stands for GARCH, E for EGARCH, GJR for GJR-GARCH, AP for APARCH. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. DM denotes the Diebold-Mariano test. The best result according to MSE is given in bold.

Table 5. Number of VaR exceedances (and VaR hit ratios) across all models for Bitcoin.

VaR / Model	G(N)	G(N)-GRU	G(STD)	G(STD)-GRU	G(SSTD)	G(SSTD)-GRU
VaR 5%	52 (4.35%)	33 (2.76%)	70 (5.86%)	50 (4.18%)	69 (5.77%)	46 (3.85%)
VaR 1%	23 (1.92%)	14 (1.17%)	16 (1.34%)	13 (1.08%)	16 (1.34%)	14 (1.17%)
VaR / Model	E(N)	E(N)-GRU	E(STD)	E(STD)-GRU	E(SSTD)	E(SSTD)-GRU
VaR 5%	51 (4.27%)	32 (2.68%)	63 (5.27%)	81 (6.78%)	64 (5.36%)	94 (7.87%)
VaR 1%	23 (1.92%)	13 (1.08%)	13 (1.08%)	19 (1.59%)	11 (0.92%)	33 (2.76%)
VaR / Model	GJR(N)	GJR(N)-GRU	GJR(STD)	GJR(STD)-GRU	GJR(SSTD)	GJR(SSTD)-GRU
VaR 5%	54 (4.52%)	33 (2.76%)	71 (5.94%)	47 (3.93%)	67 (5.61%)	43 (3.60%)
VaR 1%	24 (2.01%)	14 (1.17%)	16 (1.34%)	13 (1.08%)	16 (1.34%)	13 (1.08%)
VaR / Model	AP(N)	AP(N)-GRU	AP(STD)	AP(STD)-GRU	AP(SSTD)	AP(SSTD)-GRU
VaR 5%	**56 (4.69%)**	32 (2.68%)	63 (5.27%)	67 (5.61%)	55 (4.60%)	65 (5.44%)
VaR 1%	23 (1.92%)	14 (1.17%)	**12 (1.00%)**	14 (1.17%)	11 (0.92%)	18 (1.50%)

Note: G stands for GARCH, E for EGARCH, GJR for GJR-GARCH, AP for APARCH. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. Expected number of VaR exceedances, corresponding to 5% and 1% tolerance levels, are equal to 59 and 12, respectively. The best outcomes are indicated in bold.

Fig. 2. VaR forecasts for Bitcoin.

Table 6 presents detailed results for a selection of the best performing models: conditionally normal APARCH-GRU (winning in terms of MSE, MAE, HMSE and R^2 in the Mincer-Zarnowitz regression) and two 'pure' APARCH models: with either a normal or a symmetric *t*-distribution.

In general, the discrepancy between the models' performance in terms of either volatility forecasting or VaR and ES prediction is striking. The conditionally normal APARCH model combined with a GRU network produces by far superior forecasts of the returns' volatility, yet fails to yield unanimously satisfactory outcomes for the risk prediction. Conversely, it is 'sheer' APARCH models with a normal and Student's *t*-distribution that yield the best VaR and ES forecasts for 5% and 1% tolerances, respectively. Nonetheless, both APARCH models leave much to be desired as it comes to forecasting the volatility itself. Such a divergence of the models' performance may lead to a conclusion that the GKYZ volatility estimates, employed in training the GRU component of the hybrid model structures, are somewhat deficient for the follow-up task of risk assessment. Conceivably, this may be the case not only due to a generally high volatility of the Bitcoin returns, but also on account of a strikingly high 'volatility of volatility', numerous and pronounced spikes in the modelled series, interspersing otherwise relatively 'regular' returns (see Figure 2).

Table 6. Detailed comparison of the best performing models for Bitcoin.

Metrics / Model	APARCH(N)-GRU	APARCH(N)	APARCH(STD)
MSE	**0.3818**	5.835	3.7345
MAE	**0.3906**	1.5928	1.3160
HMSE	**0.0145**	0.1024	0.1290
R^2	**0.9586**	0.3186	0.5928
VaR exceedances: 5%/1%	32/14	**56**/23	63/**12**
VaR hit ratio: 5%/1%	2.68%/1.17%	**4.69%**/1.92%	5.36%/**1%**
Kupiec p-value: 5%/1%	5.81e-05(R) / 0.5597(F)	0.6197(F) / 0.0043(R)	0.6639(F) / 0.9860(F)
Christof. p-value: 5%/1%	0.0003(R) / 0.7143(F)	0.6265(F) / 0.0108(R)	0.1548(F) / 0.8850(F)
ES p-value bootstr./sample	0.0368(R) / 0.0360(R)	0.0118(R) / 0.0097(R)	0.7616(F) / 0.8199(F)

Note: VaR 5% and VaR 1% stands for 5% and 1% tolerance level. Expected number of VaR exceedances, corresponding to 5% and 1% tolerance levels, are equal to 59 and 12, respectively. F means the test failed to reject the H0 at 5% significance level, R means the H0 was rejected. G stands for GARCH model, E for EGARCH model, GJR for GJR-GARCH model, AP for APARCH model. N stands for Normal distribution, STD for Student's *t*-distribution, SSTD for skewed Student's *t*-distribution. R^2 is a coefficient of determination. The best (across the models) outcome according to a given metric is given in bold.

4. Conclusions

The main aim of the paper was to propose, develop and evaluate the effectiveness of hybrid GARCH-GRU models in forecasting financial volatility and risk, thereby bridging the most common, 'classic' econometric tools for volatility dynamics (GARCH models) with deep machine learning methods. The approach was tested on two financial assets displaying distinct volatility dynamics: S&P 500 and Bitcoin. The empirical analysis generally confirmed that the introduced hybrid models may prove effective in improving GARCH predictions, particularly the volatility forecasts.

Although no the same single model specification proved the best for each of the analysed two assets, it was the hybrid GARCH-GRU models that emerged unanimously superior for the point volatility forecasting, winning over 'standard' GARCH structures (in terms of MSE). In particular, the best volatility forecasts for S&P500 are produced by two EGARCH-GRU models (under a skewed and a symmetric *t*-distribution), while for Bitcoin – by the APARCH-GRU model (with a normal distribution). Nonetheless, this general outcome is hardly a surprise, given that the main task of the GRU network in the hybrid models was to minimise the MSE loss.

On the other hand, and somewhat to one's dismay, the apparent gains from the volatility prediction obtained from the hybrid GARCH-GRU structures do not translate unanimously into

superior Value at Risk and Expected Shortfall forecasts. From Tables 3 and 6 it can be inferred that the choice of a winning specification largely hinges on both the asset at hand as well as tolerance level. Using, for brevity, the models' acronyms used in the tables, the following models proved the most valid with respect to the risk assessment at the tolerance levels of 5% and 1%: S&P 500: GJR(SSTD)-GRU at the 5% tolerance, and AP(SSTD) at 1%; Bitcoin: AP(N) at 5%, and AP(STD) at 1%.

The above list indicates clearly that hybridising GARCH with GRU models does not necessarily yield superior risk forecasts (although, as stated earlier, improves the volatility predictions). Moreover, all of the listed specifications differ from the ones that proved the most accurate (in terms of MSE) for the volatility forecasting, mentioned in the previous paragraph. This, in turn, may actually put into question the very choice of the target function underlying the GRU components in the hybrid models advanced in this paper (hinged on the point volatility forecast accuracy). Thus it may be necessary to redefine the GRU target function specifically for the task of VaR and/or ES prediction. We leave this line of research for future work.

On the whole, the research findings corroborate the potential and purposefulness of combining 'classic' econometric models for volatility dynamics with deep machine learning approaches for the purpose of improving results produced by the former. The results presented in the current paper preclude unanimous conclusions as to the empirical advantages of such 'hybridisation', leaving it largely to a particular financial asset and task at hand. Nevertheless, the GARCH-GRU approach developed in this paper appears to systematically and considerably enhance volatility predictions, while still leaving some room for improvement with respect to risk forecasting.

Acknowledgements

Jakub Michańków acknowledges financial support from a project funded by the IDUB program: BOB-661-1057/2024 at the University of Warsaw. Janusz Morajda acknowledges financial support from project No. 019/ZII/2024/POT financed from the subsidy granted to the Krakow University of Economics. Łukasz Kwiatkowski acknowledges financial support from a subsidy for the maintenance of research potential, granted to the Krakow University of Economics.

References

1. Bollerslev, T.: Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics 31 (3): 307–3 (1986)
2. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE], https://arxiv.org/abs/1412.3555, Accessed 22 August 2023, (2014)
3. Christoffersen, P.F.: Evaluating Interval Forecasts. International Economic Review 39(4): 841–862 (1998)
4. Christoffersen, P., Hahn, J., Inoue, A.: Testing and Comparing Value-at-Risk Measures. Journal of Empirical Finance 8(3), 325-342 (2001)
5. Diebold, F.X., Mariano, R.S.: Comparing Predictive Accuracy. Journal of Business & Economic Statistics 13(3), 253–263 (1995)
6. Ding Z., Engle R.F., Granger C.W.J.: A long memory property of stock market return and a new model, Journal of Empirical Finance 1(1), 83-106 (1993)
7. Doman M., Doman R.: Modelowanie zmienności i ryzyka. Metody ekonometrii finansowej, Wolters Kluwer, Kraków (2009)
8. Engle R.F.: Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation, Econometrica 50 (4), 987–1008. (1982)
9. Fiszeder P.: Forecasting the volatility of the Polish stock index – WIG20. W: Forecasting Financial Markets. Theory and Applications. Łódź. (2005)
10. Fiszeder P.: Prognozowanie VaR – zastosowanie wielorównaniowych modeli GARCH, Modelowanie i prognozowanie gospodarki narodowej, Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego, 365-376 (2007)
11. Francq, Ch., Zakoian, J.-M.: GARCH models: structure, statistical inference and financial applications, 2nd edition, John Wiley & Sons (2019)

12. Garman, M.B., Klass, M.J.: On the Estimation of Security Price Volatilities from Historical Data. The Journal of Business 53(1), 67–78 (1980)
13. Ghalanos, A.: Introduction to the rugarch package (Version 1.4-3), https://cran.r-project.org/web/packages/rugarch/vignettes/Introduction_to_the_rugarch_package.pdf, Accessed 22 August 2023 (2022)
14. Ghalanos, A.: rugarch: Univariate GARCH models, https://cran.r-project.org/web/packages/rugarch/rugarch.pdf, Accessed 22 August 2023 (2022)
15. Glosten, L.R., Jagannathan, R., Runkle, D.E.: Relationship between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance 48(5), 1779-1801 (1993)
16. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016)
17. Harvey, D., Leybourne, S., Newbold, P.: Testing the equality of prediction mean squared errors. International Journal of forecasting 13(2), 281-291 (1997)
18. Hochreiter S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (1997)
19. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences 79(8), 2554–2558 (1982)
20. Hu, Y., Ni, J., Wen, L.: A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction, Physica A: Statistical Mechanics and its Applications, Vol. 557, Article 124907 (2020)
21. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [Cs], http://arxiv.org/abs/1412.6980, Accessed 22 August 2023 (2017)
22. Kristjanpoller, W., Minutolo, M.C.: Gold price volatility: A forecasting approach using the Artificial Neural Network–GARCH model. Expert Systems with Applications 42, 7245-7251 (2015)
23. Kristjanpoller, W., Minutolo, M.C.: Forecasting volatility of oil price using an artificial neural network-GARCH model. Expert Systems with Applications 65, 233-241 (2016)
24. Kupiec, P.: Techniques for Verifying the Accuracy of Risk Measurement Models (SSRN Scholarly Paper ID 6697). Social Science Research Network, https://papers.ssrn.com/abstract=6697, Accessed 22 August 2023 (1995)
25. Liu, W.K., So, M.K.P.: A GARCH model with artificial neural networks, Information 11(10), 489 (2020)
26. McNeil, A.J., Frey, R.: Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance 7(3), 271–300 (2000)
27. Mincer, J., Zarnowitz, V.: The Evaluation of Economic Forecasts. In: Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance, National Bureau of Economic Research, Inc, https://EconPapers.repec.org/RePEc:nbr:nberch:1214 (1969)
28. Nelson D.B.: Conditional heteroscedasticity in asset returns: a new approach, Econometrica 59(2) 347-370 (1991)
29. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L. and others: KerasTuner. https://github.com/keras-team/keras-tuner, Accessed 22 August 2023 (2019)
30. Taylor, S.: Modelling Financial Time Series. Wiley (1986)
31. Teräsvirta, T.: An Introduction to Univariate GARCH Models. In: Mikosch, T., Kreiß, JP., Davis, R., Andersen, T. (eds) Handbook of Financial Time Series. Springer, Berlin, Heidelberg (2009)
32. Tsay, R.S.: Analysis of Financial Time Series, John Wiley & Sons, Chicago (2010)
33. Williams, R.J., Hinton, G.E., Rumelhart, D.E.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (1986)
34. Yang, D., Zhang, Q.: Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. The Journal of Business 73(3), 477–492 (2000)
35. Zakoian J.M.: Threshold heteroscedasticity models. Journal of Economic Dynamics and Control 18(5), 931-955 (1994)