ORIGINAL ARTICLE
Journal of Postdoctoral Research (JPR)

Stock Risk Assessment via Multi-Objective Genetic Programming

Amirhessam Tahmassebi[1] | Amir H. Gandomi[2]* | Anke Meyer-Baese[1]

[1]Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, USA

[2]School of Business, Stevens Institute of Technology, Hoboken, New Jersey 07030, USA

Correspondence
* School of Business, Stevens Institute of Technology, Hoboken, New Jersey 07030, USA
Email: a.h.gandomi@stevens.edu

Funding information
None

Recent exponential growth of investors in stock markets brings the idea to develop a predictive model to forecast the total risk of investment in stock markets. In this paper, an evolutionary approach was proposed to predict the total risk in stock investment based on an S&P 500 database in a time period of 1991-2010 employing a multi-objective genetic programming along with an adaptive regression by mixing algorithm. The reasonable results suggest that the proposed model can be applied to various stock databases to assess the total risk of investment. The proposed model along with stock selection decision support systems can overcome the disadvantages of weighted scoring stock selection.

KEYWORDS
Stock Market, Risk Assessment, Multi-Objective Genetic Programming, Adaptive Regression

1 | INTRODUCTION

The facts that the stock markets may not be semi-strong-form efficient or weak-form efficient in some periods bring the idea that the return of investment can be increased by employing proper statistical factors Liu and Yeh (2017). There are numerous articles in stock prediction to find consistent paths to relate value and momentum return based on common factors Asness et al. (2013)Mohanram (2005)Roko and Gilli (2008). To name a few, Holthausen and Larcker (1992) examined the profitability of a trading strategy based on a logit model to forecast the sign of subsequent twelve-month excess returns from accounting ratios over the 1978–1988 period. Moreover, Sorensen et al. (2000)Velikova and Daniels (2004) employed classification trees in various contexts including stocks portfolio and housing price. There are nonlinearities among most of the fundamental statistical factors Duda et al. (1973). Therefore, most of the

time linear models should be discarded and nonlinear robust models are needed. In addition to this, there is no clear interpretation of the relations in statistical favor models since they work like a black box Roko and Gilli (2008). Various papers have addressed the same issues by considering several linear models to predict the relations among factors Fama (1970)Albanis and Batchelor (2000)Arentze and Timmermans (2003). To address the black box concept and nonlinear relations among the model factors, Liu and Yeh (2017) have designed a stock selection decision support systems using neural networks. In this paper, to complete the proposed model by Liu and Yeh (2017), a multi-objective genetic programming was proposed to be added to the available decision system to predict the total risk in stock market. To test out the performance of the proposed predictive model, an S&P 500 database in a time period of 1991-2010 was employed.

Algorithm 1: ARM

Input: Input Variables X_i, Target Variables Y_i, i $\in$ (1,N), Function $\hat{f}$

Output: Best Model

Random permutation the order of the observations M;

for $m \in \{1, \ldots, M-1\}$ **do**

  Randomly permute the order of the observations.;

  Split the data into two parts ;

  $Z^{(1)} = (X_i, Y_i)_{i=1}^{\frac{N}{2}}$;

  $Z^{(2)} = (X_i, Y_i)_{i=\frac{N}{2}+1}^{N}$;

  for $j \in \{1, \ldots, J\}$ **do**

    Estimate $\hat{f}_{j,\frac{N}{2}}(x; Z^{(1)})$ of f ;

    Estimate the variance function $\sigma^2(x)$ by $\hat{\sigma}^2_{j,\frac{N}{2}}(x)$;

    for $i \in \{\frac{N}{2}+1, \ldots, N\}$ **do**

      Predict Y_i by $\hat{f}_{j,\frac{N}{2}}(X_i)$;

    end

    $E_j = \frac{\prod_{i=\frac{N}{2}+1}^{N} h((Y_i - \hat{f}_{j,\frac{N}{2}}(X_i))/\hat{\sigma}_{j,\frac{N}{2}}(X_i))}{\prod_{i=\frac{N}{2}+1}^{N} \hat{\sigma}_{j,\frac{N}{2}}(X_i)}$;

    Compute the current weight $\hat{W}_j = \frac{E_j}{\sum_{l=1}^{J} E_l}$;

  end

  The final estimate is $\hat{f}_N(x) = \sum_{j=1}^{J} \hat{W}_j \hat{f}_{j,N}(x)$;

end

2 | GENETIC PROGRAMMING

Genetic Programming (GP) Koza (1992) is a symbolic optimization technique that searches the feature space to find the best fitted mathematical model for both accuracy and simplicity. Based on functional programming language, GP applies the selection framework on objectives of the problem such as fitness and complexity measures. The whole procedure follows the principle of Darwinian natural selection to use computer programs for solving a problem through evolutions. In fact, GP is a population-based method through generations instead of choosing only one candidate. This

stage is built by randomly mixing mathematical building blocks such as mathematical operators, analytic functions, constants, and state variables. Genetic operators make new generations guided by objective functions to ensure the quality of each individual. GP function regressor Veeramachaneni et al. (2015)Veeramachaneni et al. (2013) was implemented as a Multi-Objective Genetic Programming (MOGP) approach based on Non-Dominated Sorting Genetic Algorithm II (NSGA-II) introduced by Deb et al. (2002). The algorithm employs two different objective functions including model errors and the subtree complexity measure. Then, based on Adaptive Regression by Mixing (ARM) algorithm as shown in Algorithm 1, a fused model was proposed. The proposed fused model based on the ARM algorithm has the ability of adapting itself over different procedures to perform well under various conditions. Essentially, the goal of employing the ARM algorithm was to produce a model by giving different weights to some of the models in the Pareto front via proper assessment of performance of the estimators Yang (2001). GP has shown great performance in predicting complex patterns using its evolutionary nature Gandomi et al. (2015)Tahmassebi et al. (2017c)Tahmassebi et al. (2017a)Tahmassebi and Gandomi (2018) and flexibility to be combined with parallel algorithms to run multiple jobs using high performance computing (HPC) Tahmassebi et al. (2017b).

TABLE 1 Parameters setting for the GP function regressor.

Parameter	Setting
Population Size	1000
Number of Generations	500
Tournament Size	20
Number of Inputs	5
Crossover Rate	0.1
Mutation Rate	0.9
Number of Examples in Training Set	189
Number of Examples in Testing Set	63
Number of CPU Threads	4
1^{st} Objective	MSE
2^{nd} Objective	Subtree Complexity
Population Initialization	Ramped-Half-and-Half
Function Set	$+, -, \times, /, \sqrt{\ }, (\ )^2, (\ )^3, (\ )^4$
	log, exp, sin, cos

3 | RESULTS & DISCUSSION

To test out the performance of the proposed model, an S&P 500 database presented by Liu and Yeh (2017) was employed. Liu and Yeh (2017) built a stock selection decision support model using mixture design and neural networks. In this

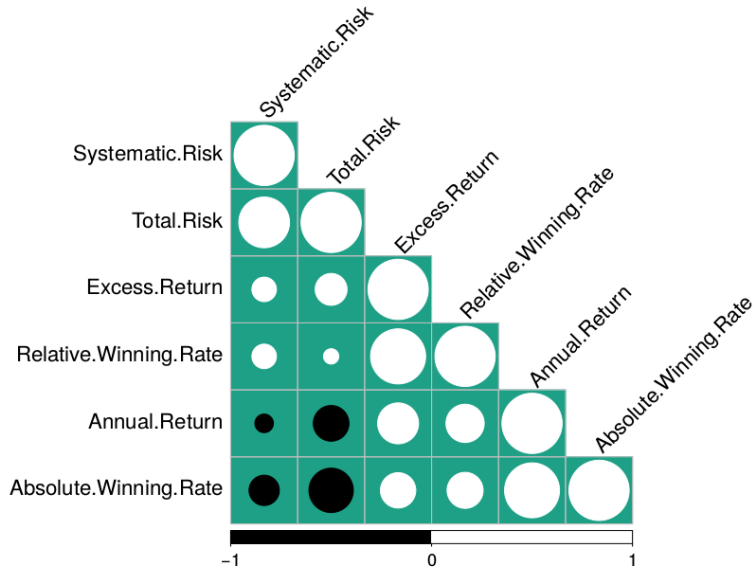


FIGURE 1 Correlation matrix illustration of the input variables along with the output variable using hierarchical clustering.

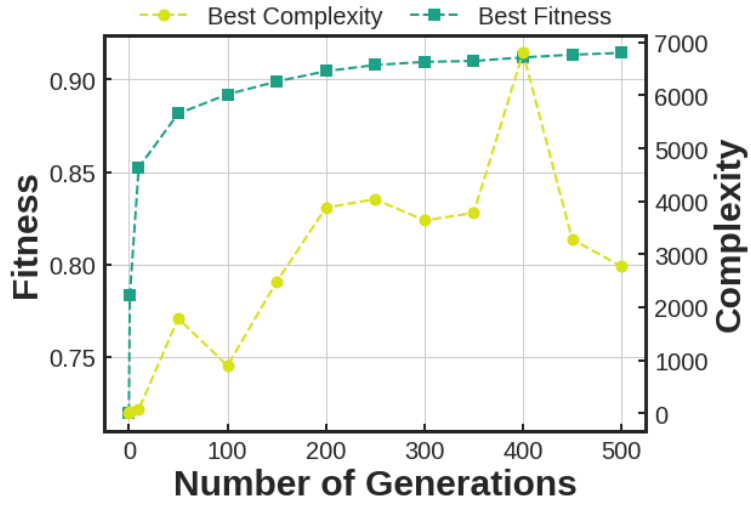
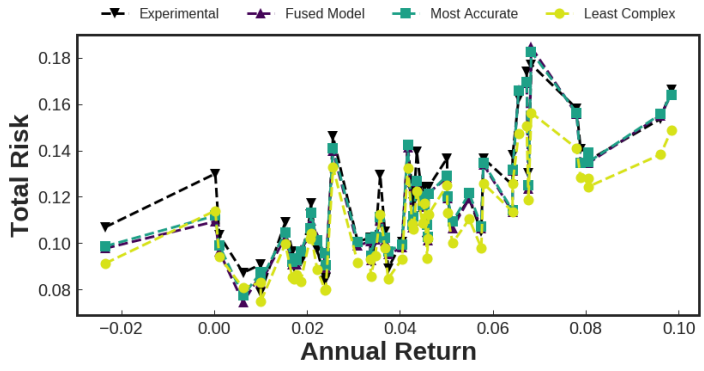
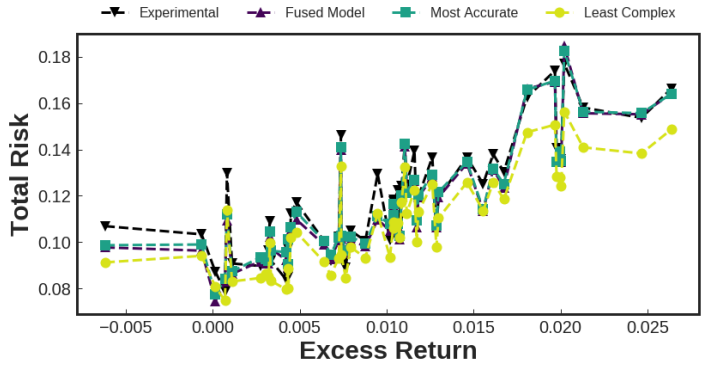


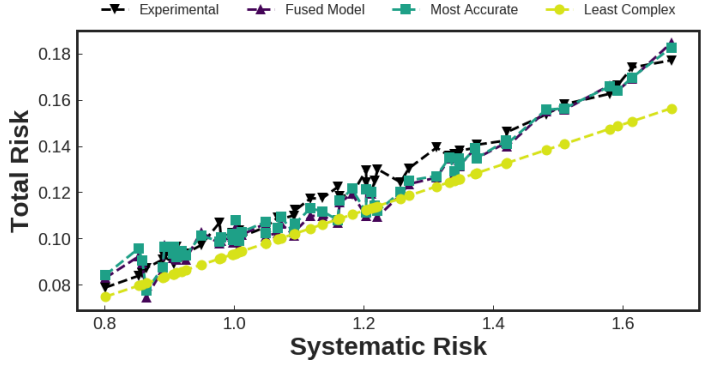
FIGURE 2 The evolution of the employed objective functions, fitness and complexity measures for the developed GP model through different numbers of generations.



(a) Annual Return



(b) Excess Return



(c) Systematic Risk

FIGURE 3 An exhaustive comparison of the predicted total risk for (a) the annual return, (b) the excess return, and (c) the systematic risk using the most accurate model, the least complex model, and the fused model versus the experimental data.

regard, they computed the optimal weighting combinations of stock-picking concepts in four different periods of time from 1991 until 2010 Liu and Yeh (2017)Yeh and Cheng (2010)Yeh and Hsu (2011). Employing the calculated weight throughout a densely connected neural network model, they have chosen six different output targets including (1) annual return, (2) excess return, (3) systematic return, (4) absolute winning rate, (5) relative winning rate, and (6) total risk. Figure 1 presents the correlation matrix illustration of the input variables along with the output variable using hierarchical clustering. Positive correlations are displayed in white and negative correlations in black. It is clear that the diagonal has the probability of one (full white circle). The size of the circles are proportional to the correlation coefficients. Thus, as the circle gets progressively larger this indicates the features are more correlated which in turn can be both positive or negative (black or white). For example, the total risk and the absolute winning rate are inversely correlated and the absolute winning rate and the annual return are linearly correlated.

In this paper, robust models using a combination of the MOGP and the ARM algorithm were proposed to find the relation between the total risk with the rest of the targets. All the proposed models were trained using the first three time periods and were tested on the fourth time period to stay away from overfitting. Table 1 presents the parameter settings for the GP function regressor. Figure 2 presents the evolution of the employed objective functions, fitness and complexity measures for the developed GP model through 500 generations over the training data set. As shown, the fitness measure reached a value of 93% after 500 generations. In addition to this, the subtree complexity measure increased through numbers of generations at first, but after 500 generations, it decreased and reached a stable value of 2870. In addition to this, a fused model based on ARM algorithm as shown in Algorithm 1 was presented as well. The proposed fused model based on the ARM algorithm has the ability of adapt itself over different procedures to perform well under various conditions. In other words, the goal of employing the ARM algorithm was to produce a model by giving different weights to some of the models in the Pareto front via proper assessment of performance of the estimators Yang (2001).

Figure 3 presents an exhaustive comparison of the predicted total risk for the annual return (Figure 3a), the excess return (Figure 3b), and the systematic risk (Figure 3c) using the most accurate model, the least complex model, and the fused model versus the experimental data. As shown, the most accurate model with a value of 0.9297 for R^2 and a value of 4.7×10^{-5} for MSE showed the best performance in predicting the total risk. As seen, the predicted values are truly close to the experimental data. Table 2 presents the summary statistics including correlation coefficient (R^2), mean-square error (MSE), and mean absolute error (MAE) of the results of the GP function regressors including the most accurate model, the least complex model, and the fused model. Higher R^2 values and lower MSE values result in a more precise model. Although the proposed fused model could not outperform the most accurate model on the employed database, it has shown great potential in complex databases Ilario da Silva et al. (2017)Veeramachaneni et al. (2013)Veeramachaneni et al. (2015).

TABLE 2 Regression score metrics of the selected models in the Pareto front.

Model	R^2	MSE	MAE
Least Complex	0.7812	1.2×10^{-4}	0.0105
Fused	0.9134	5.13×10^{-5}	0.0055
Most Accurate	0.9297	4.17×10^{-5}	0.0049

4 | CONCLUSIONS

This paper aims at developing an evolutionary symbolic implementation for stock prediction. In this regard, a multi-objective genetic programming strategy based on non-dominated sorting genetic algorithm II with considering the optimization of mean-square error as the fitness measure and the subtree complexity as the complexity measure simultaneously was employed. The GP model ran for 500 generations with 1000 populations considering training/testing sets to overcome any possible over-fitting. As shown in Table 2, higher R^2 values and lower MSE values result in a more precise model. The most accurate model with an R^2 of 0.9297 showed the best performance on the employed S&P 500 database.

REFERENCES

Albanis, G. T. and Batchelor, R. A. (2000) Five classification algorithms to predict high performance stocks. In *Advances in Quantitative Asset Management*, 295–317. Springer. URL: https://doi.org/10.1007/978-1-4615-4389-3_13.

Arentze, T. and Timmermans, H. (2003) Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: methods and empirical illustration. *Journal of Geographical Systems*, **5**, 185–206. URL: https://link.springer.com/article/10.1007%2Fs10109-003-0097-9.

Asness, C. S., Moskowitz, T. J. and Pedersen, L. H. (2013) Value and momentum everywhere. *The Journal of Finance*, **68**, 929–985. URL: https://doi.org/10.1111/jofi.12021.

Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, **6**, 182–197. URL: http://ieeexplore.ieee.org/document/996017/.

Duda, R. O., Hart, P. E. and Stork, D. G. (1973) *Pattern classification*. Wiley, New York.

Fama, E. F. (1970) Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, **25**, 383–417. URL: https://doi.org/10.2307/2325486, https://doi.org/10.1111/j.1540-6261.1970.tb00518.x.

Gandomi, A. H., Alavi, A. H. and Ryan, C. (2015) *Handbook of genetic programming applications*. Springer. URL: https://doi.org/10.1007/978-3-319-20883-1.

Holthausen, R. W. and Larcker, D. F. (1992) The prediction of stock returns using financial statement information. *Journal of Accounting and Economics*, **15**, 373–411. URL: https://doi.org/10.1016/0165-4101(92)90025-W.

Koza, J. R. (1992) *Genetic programming: on the programming of computers by means of natural selection*, vol. 1. MIT press.

Liu, Y.-C. and Yeh, I.-C. (2017) Using mixture design and neural networks to build stock selection decision support systems. *Neural Computing and Applications*, **28**, 521–535. URL: https://doi.org/10.1007/s00521-015-2090-x.

Mohanram, P. S. (2005) Separating winners from losers among lowbook-to-market stocks using financial statement analysis. *Review of accounting studies*, **10**, 133–170. URL: https://doi.org/10.1007/s11142-005-1526-4.

Roko, I. and Gilli, M. (2008) Using economic and financial information for stock selection. *Computational Management Science*, **5**, 317–335. URL: https://doi.org/10.1007/s10287-007-0056-x.

Ilario da Silva, C. R., Orra, T. H. and Alonso, J. J. (2017) Multi-objective aircraft design optimization for low external noise and fuel burn. In *58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 1755.

Sorensen, E. H., Miller, K. L. and Ooi, C. K. (2000) The decision tree approach to stock selection. *The Journal of Portfolio Management*, **27**, 42–52. URL: https://doi.org/10.3905/jpm.2000.319781.

Tahmassebi, A. and Gandomi, A. H. (2018) Building energy consumption forecast using multi-objective genetic programming. *Measurement*, **118**, 164 – 171. URL: https://doi.org/10.1016/j.measurement.2018.01.032.

Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Schmaal, L., Goudriaan, A. E. and Meyer-Baese, A. (2017a) An evolutionary approach for fmri big data classification. *IEEE Congress on Evolutionary Computation*. URL: https://doi.org/10.1109/CEC.2017.7969421.

Tahmassebi, A., Gandomi, A. H. and Meyer-Bäse, A. (2017b) High performance gp-based approach for fmri big data classification. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 57. ACM. URL: https://dl.acm.org/citation.cfm?doid=3093338.3104145.

Tahmassebi, A., Gandomi, A. H., Schulte, M. H., Schmaal, L., Goudriaan, A. E. and Meyer-Baese, A. (2017c) fmri smoking cessation classification using genetic programming. *Workshop on Data Science meets Optimisation*.

Veeramachaneni, K., Arnaldo, I., Derby, O. and O'Reilly, U.-M. (2015) Flexgp. *Journal of Grid Computing*, **13**, 391–407. URL: https://doi.org/10.1007/s10723-014-9320-9.

Veeramachaneni, K., Derby, O., Sherry, D. and O'Reilly, U.-M. (2013) Learning regression ensembles with genetic programming at scale. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, 1117–1124. ACM. URL: https://doi.org/10.1145/2463372.2463506.

Velikova, M. and Daniels, H. (2004) Decision trees for monotone price models. *Computational Management Science*, **1**, 231–244. URL: https://doi.org/10.1007/s10287-004-0014-9.

Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association*, **96**, 574–588. URL: https://doi.org/10.1198/016214501753168262.

Yeh, I.-C. and Cheng, W.-L. (2010) First and second order sensitivity analysis of mlp. *Neurocomputing*, **73**, 2225–2233. URL: https://doi.org/10.1016/j.neucom.2010.01.011.

Yeh, I.-C. and Hsu, T.-K. (2011) Growth value two-factor model. *Journal of Asset Management*, **11**, 435–451. URL: https://doi.org/10.1057/jam.2010.24.

AMIRHESSAM TAHMASSEBI received his B.Sc., and M.Sc. degrees in Physics from The University of Tehran, Iran, and The University of Akron, Ohio, in 2010, and 2015 respectively. He is currently Ph.D. candidate in the Department of Scientific Computing at Florida State University. His research interests include data mining, machine learning, medical imaging, scientific computing, genetic programming, and deep learning.

Amir H. Gandomi is an assistant professor of Analytics & Information Systems at School of Business, Stevens Institute of Technology. Prior to joining Stevens, Dr. Gandomi was a distinguished research fellow in headquarter of BEACON NSF center located at Michigan State University. He received his PhD in engineering and used to be a lecturer in several universities. Dr. Gandomi has published over 130 journal papers and 4 books. Some of those publications are now among the hottest papers in the field and collectively have been cited about 8,500 times (h-index = 46). Recently, he has been named as 2017 Clarivate Analytics Highly Cited Researcher (The Top 1%) and ranked 20th in GP bibliography among more than 10,000 researchers. He has also served as associate editor, editor, and guest editor in several prestigious journals and has delivered several keynote/invited talks. Dr. Gandomi is part of a NASA technology cluster on Big Data, Artificial Intelligence, and Machine Learning. His research interests are global optimization and (big) data mining using machine learning and evolutionary computations in particular.

Anke Meyer-Baese received her Ph.D. in Electrical and Computer Engineering from Darmstadt University of Technology, Germany in 1995. She is now Full Professor at the Department of Scientific Computing at Florida State University. Her research interests are data mining, pattern recognition techniques, and neural networks applied to various fields including (1) medical imaging such as breast MRI, computer-aided diagnosis, fMRI data analysis, (2) computational biology including dynamical analysis of gene regulatory networks, graph theoretical concepts applied in therapeutics of glioblastoma, stem cells, phosphoproteomics, and (3) computational neuroscience such as brain-based classification techniques, nonlinear stability analysis of cortical systems, graph theory applied to cortical networks. Dr. Meyer-Baese has published over one hundred journal papers and books. Some of those publications are now among the hottest papers in the field, and collectively have been cited more than 2,500 times (h-index = 25).