Development of a Genetic Programming-based GA Methodology for the Prediction of Short-to-Medium-term Stock Markets

Manal Alghieth, Yingjie Yang, Francisco Chiclana

Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, UK

Abstract— This research presents a specialised extension to the genetic algorithms (GA) known as the genetic programming (GP) and gene expression programming (GEP) to explore and investigate the outcome of the GEP criteria on the stock market price prediction. The aim of this research is to model and predict short-to-medium term stock value fluctuations in the market via genetically tuned stock market parameters. The technology proposes a fractional adaptive mutation rate Elitism (GEP-FAMR) technique to initiate a balance between varied mutation rates and between varied-fitness chromosomes, thereby improving prediction accuracy and fitness improvement rate. The methodology is evaluated against different dataset and selection methods and showed promising results with a low error-rate in the resultant pattern matching with an overall accuracy of 95.96% for short-term 5-day and 95.35% for medium-term 56-day trading periods.

Keywords—Stock market; Time series financial forecasting; gene expressing programing.

I. INTRODUCTION

Stock price forecasting has long been a focus of intelligent soft computing techniques to improve the predictability of financial systems. Due to rapidly changing trends in current global financial markets and the ongoing commercial uncertainties, accurate forecasting of time-based financial trends has become increasingly important. Stock market forecasting provides the investors with a general overview of the changing tendency of the stock markets and supports them in making timely decisions on buying or selling stocks.

Gene Expression Programming (GEP) is regarded as an extension of GA where the chromosomes take form of programming-based solutions instead of decoded genetic phenotypes that contain fixed-length binary chromosomal strings. Similar to conventional GA, GEP has the capability to generate computer-based models based on natural genetic evolution. Conventional GA creates sequence of numbers that represent solutions, whereas GEP solutions are computer programmes that follow a tree hierarchy where each node represents a programme expressed in a functional programming language. Therefore, in GEP the algorithm optimises a family of computer programmes based upon the contextual fitness scenario. Similar to GA, GEP contains linear, fixed-length chromosomes and hierarchical structures similar to genetic programming trees. Each of the GEP solutions is represented via a tree-node or chromosome. Each generation in a GEP contains a number of these chromosomes that go through a process of recombination to induce the so-called diversity in a genetic population technically regarded as the solution search space.

A wide range of GEP approaches have been reported in the literature, including linear and graph-based approaches. The technique proposed in this paper is focused primarily on the linear approach. In the linear case of GEP, programmes are represented as sequence of programmes that are encoded or decoded as nonlinear entities. A linear interpretation system is substantially faster than the tree-based system. The algorithm uses fixed-character-solution-strings to encode solutions to the problems. These strings are eventually annotated as parse trees bearing different sizes and shapes, and are called GEP expression trees (GEP-ETs). During a GEP genetic run, the GEP results in the evolution of additional complex programmes made-up of numerous, simpler sub-programmes. As a typical GEP gene contains fixed-length character strings, these can be drawn from any element from a function-set like and a terminal set like.

This research presents a novel GEP based GA methodology that utilises a dynamic sliding-window-architecture over well-known stock datasets to predict future stocks based on two conditions: 1) medium-term stock training conditions spanning durations of 2+ months and 2) short-term training-based prediction based on a week-long sliding window. The technique proposed in this work extends on existing GEP-based time-series prediction techniques by improving on the limitations of constant mutation rates via a proposed GEP-FAMR mechanism. During consecutive runs, evolutionary techniques bear mutation rates that present equal likelihood for solutions with varied fitness rate to undergo a gene-alteration process. Often, this leads to fitter solutions getting lost. However, the same mutation operator when applied to low-to-average fitness solutions, facilitate diversity and hence paves way for better solutions to survive. To date, the majority of adaptive genetic mutation techniques address fitness-bound adaptation, which tends to increase the probability of mutation for low-fitness so high that the search effectively end-up as a random heuristic for low-fitness solutions and a pure Elitism approach for fitter solutions. The proposed

technique addresses a fractional mutation approach, which adapts the mutation probability based on chromosomal sub-groups or containers to create clustered fitness groups.

II. Literature review

Evolutionary computing is widely used in real-world time-series prediction cases including electricity load prediction, tourism demand forecasting (Hong *et al.*, 2011), bankruptcy prediction (Gordini, 2014), stock control (Sourirajan *et al.*, 2009), weather/climate prediction (Venkadesh *et al.*, 2013), and traffic congestion prediction problem (Zhang et al., 2014).

Standalone "GA" were recently integrated by Stepanek *et al*. in an existing multi-agent stock market simulation algorithm. Rough-set theory (RST) integrated with GA by Cheng *et al*. (2010) reported market trend identification (bullish or bearish). RST did improve the accuracy marginally when compared to GA, thereby proving its better capability in generating rules from data pairs. However, it also reported the shortcomings of individual RST and GA in the prediction of market trends, especially in the presence of varying and conflicting market parameters. Araújo *et al*. (2009) proposed a modified GA methodology based on a morphological-rank-linear (MRL) filter, which is capable of finding sub-time-series anomalies to filter out initial sub-optimal parameters, thereby improving the overall prediction rate. A multi-objective stock index tracking methodology was proposed by Ni and Wang (2013) based on profitability, stability and volatility of stocks. The domain has recently been extended via the generation of an ensemble of recurrent neural networks via evolutionary multi-objective GA that tune the recurrent neural architecture with the best set of models selected on the basis of a Pareto front (Smith & Jin, 2014). In a similar cash flow prediction domain, Weighted SVM and fuzzy logic were combined-mapped to "fast-messy" GA chromosomes in a bid to address vagueness via fuzzy logic and improve temporal prediction via improved input/output mapping (Cheng & Roy, 2011).

A fuzzy-genetic approach was adopted by Aladag *et al*. (2014) to remove the so-called "lagged variables" from a fuzzy time-series predicting classifier via a GA-based selection process. The work evaluated its algorithms against the Taiwanese stock exchange data and reported a marked improvement in the removal of fuzzy uncertainty from the predicted outcome. A similar study by Cai *et al*. (2013) encoded entire fuzzy inference systems along with the parameters into a GA-based run. Wei in (2013) offered a GA-weighted adaptive network based fuzzy inference system (ANFIS) methodology to tune the membership functions of a fuzzy inference system, while Xiao-qin in (2012) implemented a similar fuzzy-genetic technique to fuzzify the parameters of the GA.

Particle Swarm Optimisation (PSO) is a domain of evolutionary computing algorithms commonly used to predict data patterns based upon a set of temporally varying parameters similar to a flock of birds. Similar to other evolutionary computing approaches, this problem also suffers from getting stuck into local minima. Pulido *et al*. (2014) introduced PSO to an ensemble of ANNs to tune a Type1/Type2 fuzzy system. The technique eliminated the need of manual adjustment of fuzzy rules via a set of ANNs whose parameters were dynamically tuned via the PSO routine.

Similar work done directly in time-series analysis of stock market via GEP was reported by Jie et al. (2007) and reported prediction via Sliding Window Prediction Method (GEP-SWPM) and Differential Equation Prediction Method (GEP-DEPM). The former (GEP-SWPM) technique only addresses a past-to-future data directly. The latter (GEP-DEPM) approach proposed a differential equation prediction method to mine a differential equation from training data to forecast future stock values. However, both approaches only evaluate their outcomes on sunspot activity. Moreover, differential equations predictions are known to have a higher computational complexity compared to standard quadratic with a comparatively high sensitivity to noisy data (Chen *et al*., 2011).

In any case, due to highly complex nature of stock markets, factors tend to interact with each other in a non-deterministic way and all these interactions cannot be modelled due to the increased computational complexity of the underlying system. Models can be improved via dimensionality reduction techniques. However, parameters in any pattern recognition algorithms, such as those used in the kernel function in SVM, can be optimised.

The number of factors affecting certain stock prices is so massive that it is almost impossible even for an experienced financial specialist to pick from a combination of these factors. A lot of these contributing factors are intangible. That is, even though companies' profits can be predicted with a reasonable level of accuracy, it is not possible to predict a plethora of qualitative factors such as a company's staff skills, trading experience, competitive edges, reputation, etc. Most importantly, in real stock markets, humans are actively involved and often make completely nondeterministic and irrational decisions that directly affect the future stock prices of a firm.

III. Problem

The main stock prediction problem can be described as follows: Provided current stock value(s) and the trading volume, the task is to predict the next time-instance stock market value. The hypothesis assumed here is that the stock values during a trading day change gradually and display an organised relationship between current and future stock values. This relationship can be described as an algebraic function, whose derivation is the objective of this research (Peng *et al.*, 2014). The problem particularly focuses on the closing stock value in conjunction with the volume traded.

IV. Proposed Methodology

Based on the problem provided in the above sections, the adaptation of GP/GEP to the proposed algorithm is given below:

A. Chromosomal encoding

A set of functions *F* made of classical algebraic, Boolean and relational programming parameters including logical decision constructs that include IF-THEN-ELSE and a range of programming functions are used in the encoding stage.

B. Fitness evaluation

The most basic fitness rule in this methodology is taken to be the rate of change of stock prices over a certain historical data pattern (medium or short-term). Therefore, for an average trading period starting at $day\ 0$ and ending at $day\ t$ would simply be the difference of trade volume/closing price between the beginning and starting stock values of the sliding window instance.

C. Generation of initial population

The methodology uses an ensemble of programming trees that follow the recursive construction processing from root to leaf via the guidelines described by Potvin *et al*. (2004).

- The tree root selection is made from Boolean functions and operators
- After the root has been selected, its descendants (leaves) can be selected from Boolean constants and functions and Boolean or relational operators
- If a relational operator is selected, its descendants are selected from either real functions or terminals

D. Selection of next generation chromosoms

The selection process in this work adopts three main methods namely Elitism, rank-based selection and the roulette-wheel sampling (Whitley, 1989; Baker, 1985).
In Rank-based sampling, during any genetic run, all the programmes (chromosomes/solutions) are ranked based upon their best-to-worst raw fitness values. For each chromosome, a new fitness value is then assigned based upon the following formula:

$$\mathrm{F}_i = Max - [(Max - Min)\, {}^{i-1}/_{p-1}] \qquad (1)$$

In roulette-wheel (RW) selection, each programme is assigned a slice of the roulette wheel based upon its fitness. Therefore, the individuals with highest fitness have more chances of getting selected in the subsequent generations. Finally, the Elitism-based methodology is used to select at least one best individual with the highest fitness value in the subsequent generations. This minimises the chances of best programmes recombining with weaker candidate resulting in best solutions getting lost during the genetic runs. This however presented another challenge in the overall running where the fittest individual was often lost based on the underlying mutation probability, which was set at 0.01 (1%).

o *Composition of evolutionary operators*

The most common pitfall of conventional GA is its complete randomness during the mutation process, which is equally likely to mutate (and hence destroy) the fittest chromosome as well as low-fitness-ones. Additionally, low fitness chromosomes are less likely to generate good chromosomes due to the lack of essential building blocks to create better chromosomes (solutions). Hence, a balance must be kept where low-fitness chromosomes are to be mutated in a hope to get better solutions while making it less likely for high-fitness solutions to get mutated in order to continue with the regular fitness improvement under normal crossover probability conditions (Jiang et al., 2008, Lei et al., 2007, Limin et al., 2008, Bautu et al., 2007). The techniques primarily focus on reducing mutation rate for entire generations or individual, low fitness chromosomes. This method, though improves the survival of high fitness candidates, ultimately increases the chances of low-fitness solution survival, which lead to low diversity duration in the later stages of the genetic run. Therefore, the approaches still present a significant risk of getting the genetic process stuck into local minima based on the fact that a large number of very low fitness functions with an extremely high mutation probability are likely to transform into a completely random process.

o *Chromosomal composition as encoded real-numbered constants:*

Automatically Defined Functions (ADFs) were originally introduced by Koza (1994) which derived their logic from the manner in which human programmers as reusable components. Based on the real-numbered time-series nature of the algorithms, random numerical constants were incorporated in the Automatically Defined Functions (ADFs). Taking, for instance, the example of a 5-day stock training sequence as follows:

$$S_0 = \{231.3, 232.5, 235.2, 235.7, 234.3, 236.6\}$$

$$S_1 = \{233.4, 229.1, 230.3, 238.9, 239.2, 239.8\}$$

The underlying ADF0 and ADF1 were defined based on Koza (1994) sextic polynomial constant form concept:

$$x^6 + 2x^4 + x^2 = x^2(x-1)^2(x+1)^2 \qquad (2)$$

The proposed approach used two ADF-derived function sets deriving from the following sets:

- Simple: (+, -, *, /),
- Extended: (+, -, *, /) and common mathematical functions (sin, cos, ln, exp, sqrt)

o *Mutation via fractionally dynamic fitness-based adaptation:*

In conventional GEP, a fixed probability is assigned which provides equal opportunity for each chromosome to be selected. However, this provides equal chances for best chromosomes to be mutated and hence lose fitness as well. The methodology proposed adopts a dynamic mutation rate selection technique which adapts dynamically based on the probability of each chromosome based on the overall fitness of the running population. The rates make it likely for highly fit individuals to be selected for mutation as well which can be seen as an inherent weakness of a genetic run. Hence, dynamic fitness-based adaptation was induced in the mutation process. For a chromosome, assuming to denote

the mutation rate of , the mutation adjustment rule would be defined as follows:

$$R_{p(c)} = (f_i/f_{max}) \times p_i \qquad (3)$$

In (3) is the fitness of solution and is the maximum fitness encountered in the generation run till that point. In this paper, the proposed approach is based on fitness-based mutation rate adjustment, which only permits mutation rate adjustment for a set of top 'n%' fitness chromosomes, while keeping the mutation rate constant for the rest (Table 1).

Table 1: Proposed adaptive mutation algorithm for fractional probability adjustment based on top-n fitness members

$\mathbb{R} = \{x/0 \leq x \leq 1\}$
Sort generation $G_i \in \mathbb{G}$ as follows (where $1 \leq \mathbb{G} < Total\ generations$)
Let $C = (x_i)_{N=1}^{n}$ be a sequence of chromosomes and their fitness satisfying:
$\forall\ 1 \leq \mathbb{G} < 100 \mid \{i \in N | x_i = x_k\}| = |\{i \in N | y_i = y_k\}$
$Select\ \mathbb{R}_X = max(x\ such\ that\ \#\{s \in \mathbb{R} \mid s \geq r\} = X)$
$\mathbb{R}^X = \{r \in \mathbb{R} | r \geq \mathbb{R}_X\}$
where 'X' are the top 'n' fittest individuals
Calculate average probability $P_y = (y_i)_{N=1}^{n}$as follows:
$R_{p(c)} = (f_y/f_{max}) \times (p_{max} - p_{min})$
Where $f_y \in \mathbb{R}^X$ is an individual whose fitness is to be calculate adaptively and
$f_{max} = \frac{1}{K}\sum_{i=1}^{y_k} y_i$ is the average maximum fitness of the top X fittest chromosomes and
p_{max} and p_{min} are the minimum and maximum probabilities of the entire group G_i

- *Transposition operator:*

Transposition operator is used in evolutionary algorithm to exchange sections of genes from within a single parent (asexually) or between two parents (sexually). Regardless of the case, in GEP, the inserted sequence can appear anywhere in tree but the insertion can only be performed at the heads of genes. The operator in the proposed approach was used to asexually change single chromosomes as shown in Table 2

TABLE 2: SIMPLE TRANSPOSITION OPERATION USED TO EXCHANGE GENETIC MATERIAL IN GEP FUNCTION TREES

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Chromosome	/	-	b	/	a	b	b	a	a	*	a	-	/	a	b
Transposed Chromosome	/	-	a	*	a	-	/	a	b	b	/	A	b	b	a

- *Recombination:*

Recombination operation generally involves two parent chromosomes which are combined via single/multipoint crossover. The resulting children chromosome is deemed syntactically correct if the resultant size is the same and the resultant fragments are homologous. The proposed technique used a single-point crossover mechanism for recombination as shown in Table 3. The technique probabilistically selected two chromosomes from a generation and swapped their genetic composition over a single crossover point. The crossover point for the case shown in Table 3 is on the immediate right of the red-coloured cell.

Table 3: Sample single-point crossover operation used to diversify GEP function trees

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Chromosome 1	/	-	b	/	a	b	b	a	a	*	a	-	/	a	b
Chromosome 2	-	b	b	a	b	*	a	b	/	a	b	b	-	a	c
Chromosome 1	/	-	b	/	a	*	a	b	/	a	b	b	-	a	c
Chromosome 2	-	b	b	a	b	b	b	a	a	*	a	-	/	a	b

V. SYSTEMATIC ANALYSIS OF GEP AGAINST STANDARD DATASETS

The fractionally adaptive mutation technique was introduced in this work to improve the manner in which fittest individuals were retained in subsequent genetic runs. The underlying objective was to improve the overall Elitism-based selection mechanism on the fact that regardless of retaining the best-performing chromosome via Elitism, there was still a possibility to lose it if it was probabilistically selected by the system to be mutated. The data used for this analysis was taken from the standard BP dataset.

A. Comparison of fractionally adaptive GEP against standard optimisation methodologies

The suitability of the proposed selection mechanism, Elitism with fractional mutation rate, was assessed by comparing it against the following two existing standards of Elitism with random mutation and constant mutation

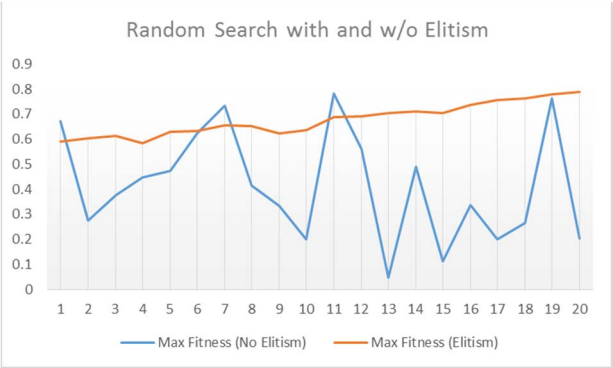
(1) Elitism with high/uncontrolled (100%) mutation rate;

(2) Elitism with constant (1%) mutation rate.

Based on the theoretical nature of GA a mutation rate of 100% effectively changes a GA run into a random search routine. In this case, the overall search lost 100% of its fittest chromosomes (subject to mutation) which were then supported by better individuals again in the subsequent generation, where sunsequently were again lost (Figure 1).

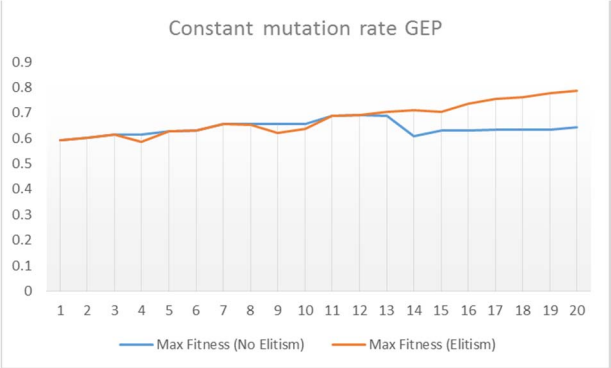
The purpose of this elaboration was to show the impact of high mutation rate on seemingly better genetic runs. However, a simple single high fitness chromosome survival leads to an organised fitness improvement of 0.7875.

FIGURE 1: REPRESENTATION OF A GENETIC RUN WITH 100% MUTATION (RANDOM SEARCH: RS) AND THE RESULTANT WORST-TO-BEST ELITISM OUTCOME



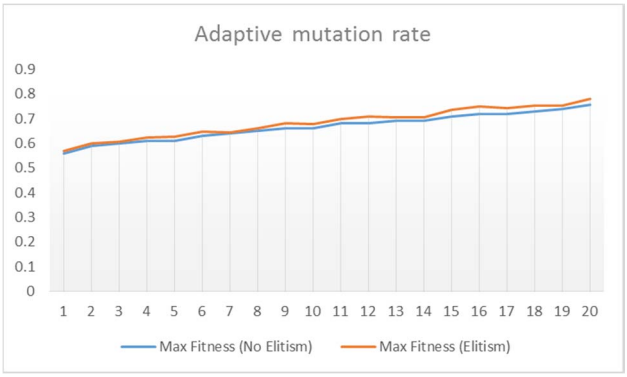
In case 2 (Figure 2), a standard 1% mutation rate was applied and a both the approaches lead to a gradually improving fitness rate. However, due to an equal (1%) mutation rate, the fittest individual was itself mutated at generation 13 and since no Elitism was applied, the resultant fitness can be seen to have taken a considerable amount of time to recover and finally achieve a fitness score of 0.6444.

FIGURE 2: REPRESENTATION OF A GENETIC RUN WITH 1% CONSTANT MUTATION RATE (CMR) AND ITS IMPACT ON THE BEST-FITNESS CHROMOSOME AT GENERATION 14



The proposed adaptive mutation showed similar outcome pertaining to not a single best chromosome getting lost to high mutation rate and finally achieving a fitness score of 0.78 (Figure 3),.

FIGURE 3: REPRESENTATION OF THE PROPOSED GEP-FAMR FRACTIONALLY ADAPTIVE MUTATION RATE MECHANISM



The rate of change of fitness for all the fitness scores can be seen in Table 4. The low rate of fitness of constant mutation rate(CMR) compared to random search(RS) must not be confused with a better performance of RS algorithm. Based on the inherent nature of the stage of randomisation, CMR is still more likely to perform better than RS in the absence of Elitism as evident from the average fitness achieved by RS under no Elitism. The rationale behind Elitism-based GEP-FAMR gaining better accuracy can still be attributed to the fact that with subsequent runs, the technique increases the total number of fitter solutions in the gene pool.

TABLE 4: RATE OF IMPROVEMENT OF FITNESS FOR VARIOUS SEARCH HEURISTICS

Algorithm used	Elitism	No elitism
RS	0.73198	0.20655
CMR	0.665501	0.971033
AMR	1.057687	0.985493

- *Functional comparison of RS, CMS and AMR techniques:*

a) Random research:

No Elitism: In each generation all chromosomes are mutated thereby changing even the highest fitness solutions. Therefore, there is no way of recovering a good solution even via Elitism.

Elitism: The technique does retain the best individual but it still gets mutation (and lost) in the subsequent generation.

b) Constant mutation rate:

No Elitism: There is a 1% chance for the fittest chromosome to get lost. However, there is still likelihood that other closer-fitness chromosomes will live or crossover with others to move to other generation and generate better fitness individuals.

Elitism: The best individual will still move to the next generation but may get selected there for crossover/mutation/recombination with any other chromosomes to generate better (or worse) candidates. This situation may lead to a lot of similar-fitness solutions leading to fitness improvement getting stuck in a local maxima.

c) GEP-Fractional Adptive Mutation Rate:

No Elitism impact on better fitness chromosomes:

The mutation probability of high fitness individuals reduce which causes them to stay in the subsequent generation, crossing with others and improving the chances of even better chromosomes

No Elitism impact on lower fitness chromosomes:

The low-fitness solutions still stay in the pool and survive to crossover with other chromosomes to induce diversity in the genetic run. This reduces the chances of the search ending-up in a local maxima.

Elitism impact on better fitness chromosomes:

The best fitness solution still replaces the lowest-fitness solution. Despite other fitter solutions moving to the next population.

Elitism impact on lower fitness chromosomes:

Despite low-fitness solution having better chance to get to the next generation, the lowest-one is still replaced by the best. This may marginally reduce the diversity of the overall population.

The overall improvement in the rate of adaptive mutation with other models proves the hypothetical

argument at the start of this research that fractional improvement is more likely to retain better individuals. Moreover, the per-generation improvement rate with Elitism presents a direct evidence that the technique will cope well with any other category of datasets.

VI. Results and Discussions

The main focus of this research is on the critical analysis of the new GEP sliding window algorithm against a diverse range of datasets and varying prediction scenarios ranging from input window containing 2+ month trailing stock data to short-term data looking back into just a week in past. The data was selected for five well-known stock companies Yahoo, British Petroleum, GSK, HSBC and the RBS. In order to understand the difference of these companies' performance, it is necessary to look into the nature of performance they showed during the past 20+ years. All the dataset described below were trained and evaluated over two window sizes of 5 days (short-term) and 56 days (medium-term). The initial results obtained are reported in table 5.

Table 5: Experimental parameters used for data presented in Figure 4,5 and 6

Time limit	5 days
Dataset	6761 days starting from 1988
Currency	US$
Function set (basic)	+, -, *, /
Population size (generations)	100
Number of genes	4
Selection operator	FAMR (proposed)
Mutation rate	Adaptive (non-constant)
Recombination rate	1%
Crossover rate	60%

Results presented in Figure 4, Figure 5 and Figure 6 represent the outcome of GEP-FAMR, green and blue plots in the figures represent actual and predicted stock rates whereas the red points show the error in prediction. The overall dataset contained a total of 6761 days of stock data where the initial 70% was used for training whereas the remaining 30% of data (from March 2006) was used for testing It must be noted that the 30% data used for testing contained a high number of outliers from the recent recession of 2006 – 2011, which made it particularly challenging for a 5-day prediction system. The high error values can particularly be seen where there are sudden and unpredicted change in stock values.

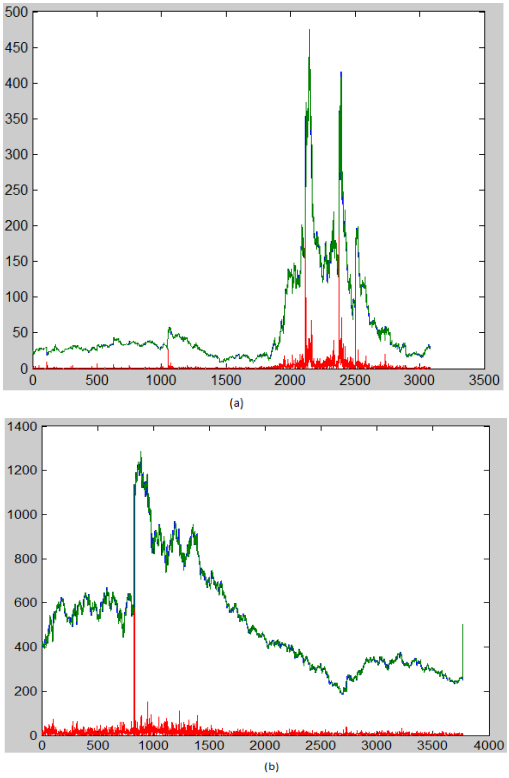


Figure 4: Actual (Green) and classified (Blue) values against error rate (Red) obtained via GEP-FAMR algorithm for (a) Yahoo and (b) British Petroleum stock data

Figure 4 is showing respective spontaneous and gradual increase in stock prices which then shows an abrupt pattern during the post-oil crisis era during the Middle East war and political turmoil for BP (2014).

Glaxo Smith Kline (GSK) was selected on its unusually spontaneous stock turmoil during early 1990s due to increasing competitions from Wellcome PLC, which launched a large number of new pharmaceutical drugs during 1980s and early 1990s. The stock currently stand at a falling pattern of 14 (0.89%) at the stock price of 1562.95.

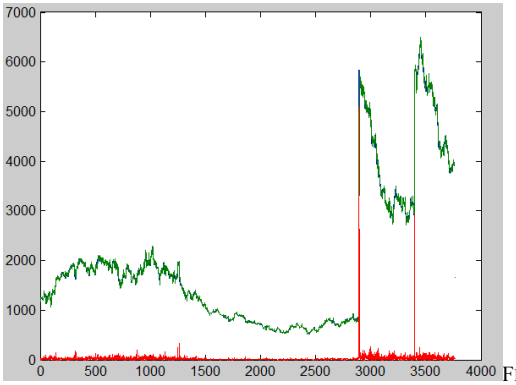


Figure 5: Actual (Green) and classified (Blue) values against error rate (Red) obtained via GEP-FAMR algorithm for GSK stock data (GSK, 2014)

The two banks involved in these tests were HSBA and RBS, whose data showed a phenomenally large drop in shares during the past 5 years of recession (Figure 6). The shares continue to fall at respective prices of 11.02 (1.96% fall) and 604.5 (0.15%) for RBS and HSBA PLCs.

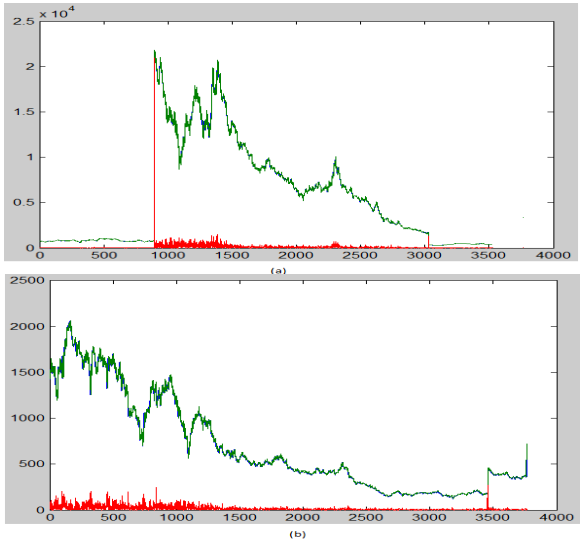


FIGURE 6: ACTUAL (GREEN) AND CLASSIFIED (BLUE) VALUES AGAINST ERROR RATE (RED) OBTAINED VIA GEP-FAMR ALGORITHM FOR HSBA/RBS STOCK DATA SHOWING A HUGE FALL DURING THE GLOBAL RECESSION (HSBA, 2014; RBS, 2014)

In the above four cases, it must be noted that both medium-term and short-term predictions bear important insight for traders to buy or sell certain shares. For instance, for long-term investors, an accurate and reliable prediction algorithm representing a correct trend (increasing or decreasing) will support in the decision making process. However, the figures (4,5and 6) report the errors rates on a short-term classifier only.

- *Initial Selection Operator Evaluation*

The algorithm's performance was initially evaluated over a fixed number of chromosomes with a total of 50 individuals (chromosomes) per generation and 1000 genetic runs with each performing/applying genetic crossover and mutation to induce diversity in the subsequent generation. The selection criteria were initially based upon Elitism, Rank-based and Roulette-wheel algorithms. As shown in Table 6, the proposed GEP-FAMR approach showed significant improvements for the cases of BP and GSK and RBS. The algorithm however showed average improvement when compared to RW selection in the case of HSBC/A, and a marginal improvement in the case of Yahoo. Prediction errors were not focused on at this stage because of possible improvement later on once the remaining parametric combination (i.e. window-size, chromosome/generation, etc.) was evaluated.

TABLE 6: COMPARISON OF LEARNING AND PREDICTION ERROR RATE FOR THREE SELECTION CRITERIA OF ELITISM, RANK-BASED AND ROULETTE-WHEEL (RW) SELECTION

Company	Selection Method	FAMR		Rank		RW	
		LE	PE	LE	PE	LE	PE
Bp		10.61	6.32	14.87	13	18.89	11
GSK		20.06	25.11	37.09	40	29.74	46.99
HSBC/A		53.07	16.5	120	0	53.45	17.99
RBS		15.31	5.76	23	8	28.89	9
Yahoo		1.62	0.89	1.63	1	1.632	1
LE: Learning Error, PE: Prediction Error RW:Roulette-Wheel							

Table 7 finally demonstrates a comparison between GEP-FAMR with both the GA variants of the conventional genetic programming (GP) and the gene expression programming (GEP) –based classification. The comparison shows GEP-FAMR to perform substantially better for both the cases of short and medium-term prediction apart from the GSK dataset where GEP with fixed mutation rate performed better. This anomaly may be attributed to the initial random conditions on which the genetic run for the particular GEP-FAMR started or the fact that a low mutation rate may still resulted in the loss of fitter individuals in the case of GEP-FAMR, which does not happen in constant-mutation-rate GEP where Elitism still retains the best performing individual anyway. Nonetheless, the overall success rate in prediction with other datasets indicates the suitability of GEP-FAMR on other methodologies.

Table7: Short and medium-term prediction accuracy based on GEP sliding-window operation (for 5-day look-ahead only)

	MT			ST		
	GEP-FAMR	GP	GEP	GEP-FAMR	GP	GEP
BP	91.14	89.39	87.01	96.32	96.23	89.76
GSK	98.45	76.34	100	99.12	76.65	100
HSBA	92.11	82.09	91.87	88.56	82.1	84.09
RBS	95.85	97.12	89.54	96.64	95	100
Yahoo	99.2	98.75	96.75	99.16	97.64	96.11
Overall accuracy	**95.35**	88.738	93.034	**95.96**	89.524	93.992

VII. CONCLUSION

The research presents a novel evolutionary methodology with modified selection operators for the prediction of stock exchange data via a specialised extension to the conventional evolutionary GA. The extension dynamically adapts the mutation rate of different fitness groups in each generation to ensure a diversification balance between high and low fitness solutions. The so-called GEP-FAMR

approach was adopted to discuss the suitability of the algorithms to predict daily and weekly stock market patterns based upon medium-to-short-term stock history. The outcomes presented an outstanding GEP accuracy compared to GP in stock prediction via simple, non-arithmetic algebraic expressions with the best performance reported at short-term forecasting of 95.96%. On the other side, the worst performance of 88.73% was reported on a GP algorithm at medium-term prediction which could be attributed to the fact the at 65-day-training-lengths, the algorithms routinely found data subjected to non-deterministic human-level changes due to global financial uncertainties such as the Middle East crisis and the global recession. However, it is understood that GEP can particularly be used to increase the overall confidence level of trading markets if it is properly evaluated for the period of a few months before actually being implemented in risky stock markets for behaviour prediction.

VIII. References

[1] Aladag C. H., Yolcu U., Egrioglu E., Bas E. (2014) Fuzzy lagged variable selection in fuzzy time series with genetic algorithms, Applied Soft Computing, Available online 29 April 2014, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2014.03.028.

[2] Alghieth, M., Yang, Y. and Chiclana, F., 2015, September. Development of 2D curve-fitting genetic/gene-expression programming technique for efficient time-series financial forecasting. In *Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on* (pp. 1-8). IEEE.

[3] Araújo R. A., Ferreira T.A.E. (2009) An intelligent hybrid morphological-rank-linear method for financial time series prediction, Neurocomputing, Volume 72, Issues 10–12, June 2009, Pages 2507-2524, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2008.11.008.

[4] Baker, J.E. (1985) Adaptive selection methods for genetic algorithms In: Proceedings of the First International Conference on Genetic Algorithms and their Applications. Hillsdale, NJ: Lawrence Erlbaum, 1985. p. 101–111

[5] Bautu, E., Bautu, A. and Luchian, H. (2007) 'AdaGEP - An Adaptive Gene Expression Programming Algorithm'. Symbolic and Numeric Algorithms for Scientific Computing, 2007. SYNASC. International Symposium on, 26-29 Sept. 2007, 403-406.

[6] BP (2014) British Petroleum Plc [Online] Available at <https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=BP&ql=1> [Accessed: 20/06/2014]

[7] Cheng C., Chen T., Wei L. (2010) A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences, Volume 180, Issue 9, 1 May 2010, Pages 1610-1629, ISSN 0020-0255, http://dx.doi.org/10.1016/j.ins.2010.01.014.

[8] Cheng M., Andreas F.V. R. (2011) Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines, International Journal of Project Management, Volume 29, Issue 1, January 2011, Pages 56-65, ISSN 0263-7863,

[9] GSK (2014) Glaxo Smith Klien Plc [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=gsk&ql=1> [Accessed: 20/06/2014]

[10] Hong W., Dong Y., Chen L., Wei S. (2011) SVR with hybrid chaotic genetic algorithms for tourism demand forecasting, Applied Soft Computing, Volume 11, Issue 2, March 2011, Pages 1881-1890, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2010.06.003.

[11] HSBA (2014) HSBC Holdings Plc [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=hsba&ql=1> [Accessed: 12/06/2014]

[12] Jiang, Y., Tang, C.-j., Zheng, H.-c., Li, C., Chen, Y., Wu, J. and Wang, D.-l. (2008) 'Adaptive Gene Expression Programming Algorithm Based on Cloud Model'. BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on, 27-30 May 2008, 226-230.

[13] Lei, D., Changjie, T., Jun, Z., Jie, Z., Yintian, L., Jiang, W. and Li, D. (2007) 'The Strategies of Initial Diversity and Dynamic Mutation Rate for Gene Expression Programming'. Natural Computation, 2007. ICNC 2007. Third International Conference on, 24-27 Aug. 2007, 265-269.

[14] Limin, H., Jinliang, Y., Lirui, G., Jie, H. and Xinqiao, F. (2008) 'Short-Term Load Forecasting Based on Improved Gene Expression Programming'. Circuits and Systems for Communications, 2008. ICCSC 2008. 4th IEEE International Conference on, 26-28 May 2008, 745-749.

[15] Ni H., Wang Y. (2013) Stock index tracking by Pareto efficient genetic algorithm, Applied Soft Computing, Volume 13, Issue 12, December 2013, Pages 4519-4535, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2013.08.012.

[16] Potvin J., Soriano P., Vallée M. (2004) Generating trading rules on the stock markets with genetic programming, Computers & Operations Research, Volume 31, Issue 7, June 2004, Pages 1033-1047, ISSN 0305-0548, http://dx.doi.org/10.1016/S0305-0548(03)00063-7. (http://www.sciencedirect.com/science/article/pii/S0305054803000637)

[17] Pulido M., Melin P., Castillo O. (2014) Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange, Information Sciences, Volume 280, 1 October 2014, Pages 188-204, ISSN 0020-0255, http://dx.doi.org/10.1016/j.ins.2014.05.006.

[18] RBS (2014) Royal Bank of Scotland [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=rbs&ql=1> [Accessed: 19/06/2014]

[19] Smith C., Jin Y. (2014) Evolutionary Multi-Objective Generation of Recurrent Neural Network Ensembles for Time Series Prediction, Neurocomputing, Available online 12 June 2014, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2014.05.062.

[20] Sourirajan K., Ozsen L., Uzsoy U. (2009), A genetic algorithm for a single product network design model with lead time and safety stock considerations, European Journal of Operational Research, Volume 197, Issue 2, 1 September 2009, Pages 599-608, ISSN 0377-2217, http://dx.doi.org/10.1016/j.ejor.2008.07.038.

[21] . Venkadesh S., Hoogenboom G. (2013) Walter Potter, Ronald McClendon, A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks, Applied Soft Computing, Volume 13, Issue 5, May 2013, Pages 2253-2260, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2013.02.003.

[22] Wei L. (2013) A GA-weighted ANFIS model based on multiple stock market volatility causality for TAIEX forecasting, Applied Soft Computing, Volume 13, Issue 2, February 2013, Pages 911-920, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2012.08.048.

[23] Xiao-qin W. (2012) Research on Prediction Model of Time Series Based on Fuzzy Theory and Genetic Algorithm, Physics Procedia, Volume 33, 2012, Pages 1241-1247, ISSN 1875-3892, http://dx.doi.org/10.1016/j.phpro.2012.05.205.

[24] Yahoo (2014) Yahoo Inc stock data YHOO [Online] Available at < https://uk.finance.yahoo.com/q/hp?s=YHOO> [Accessed: 20/06/2014]

[25] Zhang X., Onieva E., Perallos A., Osaba E., Lee V.C.S. (2014) Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction, Transportation Research Part C: Emerging Technologies, Volume 43, Part 1, June 2014, Pages 127-142, ISSN 0968-090X, http://dx.doi.org/10.1016/j.trc.2014.02.013.