Shadow Gene Guidance: A Novel Approach for Elevating Genetic Programming Classifications and Boosting Predictive Confidence

Hassan Gharoun
Hassan.Gharoun@Student.UTS.edu.au
PhD Student, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Mohammad Sadegh Khorshidi
msadegh.Khorshidi.ak@gmail.com
PhD Student, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Navid Yazdanjue
navid.yazdanjue@gmail.com
PhD Student, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Fang Chen
Fang.Chen@uts.edu.au
Distinguished Professor, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU

Amir H. Gandomi
a.h.gandomi@gmail.com
Professor, Faculty of Engineering & IT, University of Technology Sydney
Sydney, NSW, AU
Distinguished Professor, University Research and Innovation Center (EKIK), Óbuda University
Budapest, HU

ABSTRACT

This paper introduces a novel classification method that utilizes genetic programming (GP). The primary purpose of the proposed method is to enhance future generations of GP, through continuously refining the genetic makeup of the population for improved classification results. Accordingly, this paper developed the novel method by modifying Boruta feature selection method in such a way that allows to evaluate the significance of individuals' genes. This method creates modified versions of the genes called "shadow genes", evaluates their impact on model performance in competing with shadow genes, and identifies key genes. These key genes are then used to enhance future generations. The obtained results demonstrated that the proposed method not only enhances the fitness of the individuals but also steers the population toward optimal solutions. Furthermore, empirical validation on multiple datasets reveals that the proposed method significantly outperforms classic GP models in both accuracy and reduced prediction entropy, showcasing its superior ability to generate confident and reliable predictions.

CCS CONCEPTS

• **Computing methodologies** → **Genetic programming**; *Supervised learning by classification.*

KEYWORDS

Genetic Programming, Cross Over, Uncertainty-aware Classification.

ACM Reference Format:
Hassan Gharoun, Mohammad Sadegh Khorshidi, Navid Yazdanjue, Fang Chen, and Amir H. Gandomi. 2024. Shadow Gene Guidance: A Novel Approach for Elevating Genetic Programming Classifications and Boosting Predictive Confidence. In *Proceedings of The Genetic and Evolutionary Computation Conference 2024 (GECCO '24).* Melbourne, Australia, 4 pages. https://doi.org/10.1145/3638530.3664175

GECCO '24 Companion, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0494-9/24/07…$15.00
https://doi.org/10.1145/3638530.3664175

1 BACKGROUND AND BACKGROUND

The rapid expansion of machine learning (ML) applications across various domains underscores its transformative potential, particularly in achieving high predictive accuracy. However, ML models often encounter challenges related to their reliability. The majority of ML models draw conclusions without evaluating the caliber of these conclusions. This situation risks guiding either a human operator or an automated controller toward erroneous decisions [3].

Concerning such lapses in ML outcomes, it's not enough for a model to just be accurate; it's also crucial that it can measure the level of confidence in its predictions. Despite GP's success in various domains, its application in uncertainty awareness remains underexplored.

Recent developments in GP have addressed classification tasks focusing on enhancing accuracy. However, the significance of uncertainty aware decisions in classification tasks is often overlooked in GP literature and efforts to integrate uncertainty quantification (UQ) in GP mainly focus on areas beyond classification tasks. Contrastingly, existing UQ methods in the literature primarily used in NNs assess solely uncertainty without inherently boosting model confidence. On the other hand, GP with its capability to utilize

sophisticated fitness functions, holds significant potential in developing uncertainty-aware models for classification tasks, especially when combined with advancements in NN methodologies. This area represents a promising direction for further research and development.

Accordingly, this study introduces a new variant of GP that integrates a Boruta feature selection to enhance classification tasks. The cornerstone of this approach is the innovative use of shadow genes—not directly in decision-making, but as a strategic tool to identify and select significant genes, thereby generating potentially robust individuals within the GP framework. It presents a framework that fills a notable gap in GP research—particularly the lack of focus on uncertainty awareness.

2 METHODOLOGY

2.1 Preliminary: Boruta Feature Selection

The Boruta algorithm is a straightforward yet effective method for identifying the most relevant features in a dataset. It accomplishes this through a comparative analysis with artificially generated shadow features. Here's a simplified breakdown of the process:

- Generation of Shadow Features: The algorithm starts by duplicating each feature in the dataset. The values of these duplicates are then shuffled to generate shadow features, which mimic the structure of the original features but not their exact content.
- Evaluating with Random Forest: The importance of both original and shadow features is assessed repeatedly.
- After each iteration, the importance scores from the Random Forest classifier are used to compare each original feature against the highest importance score among the shadow features. An original feature is deemed significant if its importance consistently exceeds that of the best-performing shadow feature across multiple iterations.
- Iterative Refinement: This cycle of comparison and evaluation continues across several iterations, focusing on pinpointing features with genuine predictive value.
- Selecting Key Features: Ultimately, features that reliably outperform shadow features are deemed important and retained for model development.

The proposed method, detailed in subsequent sections, extends the foundational principle of shadow features in the Boruta algorithm, to the concept of shadow genes. This principle forms the basis for the creation of superior individuals, key in attaining the highest level of fitness across each generation.

2.2 Proposed Method

The proposed method employs Multi-gene Genetic Programming (MGGP) that integrates artificial neural network (ANN) classification to evolve a population of individuals for effective performance on classification tasks. Figure S1 from the Supplementary Document illustrates the sequential overview of the proposed method. The process is outlined in simple steps as follows:

- Initialization of Population: The algorithm commences by creating an initial group of individuals, each comprised of several genes structured as trees. These genes encode various operations and terminals that are relevant to the task at hand.
- Differential Fitness Evaluation: MGGP evaluates the collective output of all genes in an individual. This is achieved through a shallow neural network, which uses the outputs of the genes as inputs. The classification probability is thus determined, and the individual's fitness is calculated as a negative log loss comparing the actual and predicted class labels.
- New Population Creation: After fitness Evaluation, three subsets of the population form via:
  - The first set is derived from genetic operations:
    - Crossover: The crossover in MGGP involves swapping subtrees between genes of parent individuals, introducing diversity in the offspring's genetic composition.
    - Mutation: Mutation randomly alters genes, encouraging genetic diversity and aiding in the exploration of new solutions within the genetic landscape.
    - Selection and Reproduction: The selection process in MGGP prioritizes the fitness of the entire multi-gene structure. Individuals with higher fitness are selected for reproduction through the tournament, ensuring the propagation of beneficial genetic characteristics through generations.
  - The second set consists of elite individuals chosen based on their fitness.
  - The third set is generated from combination of elite individuals through the Boruta enforced operation elaborated in section 2.3.
- Hybrid Classification Method with Neural Network Classification: The classification in the proposed framework is enhanced by employing an ensemble approach, wherein multiple genes of an individual collectively contribute to the final decision-making process as follows:
  - *Gene evaluation*: The symbolic structure of genes within an individual is evaluated based on training and evaluation datasets.
  - *Gene (Ensemble) Integration*: The output from multiple genes is integrated using shallow ANN with fixed structure. This integration involved training the ANN on the output of genes from the training data and predict the probability of validation data class labels.
  - *Fitness Evaluation*: The fitness of synthesized genes within each individual (*"individual fitness"*) is computed based on the prediction probability of validation data and ground truth using a log loss function, which is particularly effective for classification tasks.
- Termination Criteria: The MGGP algorithm iterates through this process over successive generations. With each cycle, the quality of the solutions is refined. The process concludes once predetermined termination criteria are met.

2.3 Boruta Enforced Individual (BEI): Shadow Gene Guidance on Creating Elite Individual

Here a fraction of individuals with the highest fitness value among the population in each generation are selected to perform Boruta

operation to produce a single individual for the next generation. All genes of selected individuals for this operation are used to create their shadow counterparts. This method is conceptualized as follows:

- *Creation of Shadow Genes*: Initially, shadow genes are generated. These entities are noise-augmented versions of the original genes within the dataset. For each gene, Gaussian noise, scaled to the gene's standard deviation, is added, followed by a shuffling of values. This results in a set of shadow genes, which are essentially noisy variations of the original genes.
- *Gene Importance Assessment*: The evaluation process involves several steps:
  - A shallow ANN is trained using both original and shadow genes,
  - The model's performance, quantified by the F1-score, serves as a baseline for assessing each gene's importance.
  - Subsequently, each gene undergoes perturbation through noise addition and reshuffling. The model is retrained with this altered dataset, and a new F1-score is computed.
  - The performance decrement, attributed to the perturbation of each gene, is documented.
- *Comparative Importance Assessment and Hit Counting*: The algorithm undertakes a comparative evaluation of the original genes against their shadow counterparts. This assessment hinges on contrasting the performance decrement of each original gene with the maximum decrement observed for shadow genes. A more substantial performance drop for an original gene, relative to the maximum observed for shadow genes, signifies its importance. For each gene, a 'hit' is registered if its performance decrement exceeds the highest decrement observed among the shadow genes. The 'hit' count for each gene is accumulated over multiple iterations, providing a robust measure of the gene's relative importance.
- *Selection of Important Genes*: Genes demonstrating consistent superiority in importance across iterations, as indicated by their hit history (H), are flagged as significant. The algorithm selects the top-ranking genes, guided by the *"Maximum allowable gene"* parameter, which denotes the maximum number of genes that is allowable within each individual.
- *Producing Boruta Enforced Individual (BEI)*: The chosen genes in the previous step form a new single individual for the next generation. In this step, the most significant genes, as determined by their hit history and importance, are combined to form a new, singular individual for the next generation.

3 EXPERIMENTAL SETUP

3.1 Data Sets

Study uses four datasets to assess method performance: (1) the dataset of Activity Recognition Using Wearable Physiological Measurements (ARWPM)[1], capturing physiological signals with 4,480 instances and 533 features, (2) the Gene Expression Cancer RNA-Seq Data Set (GECR) dataset[2], offering gene expression profiles for cancer classification with 801 instances and 20,531 features, (3) the Gas Sensor Array Drift Dataset at Different Concentrations (GSAD) dataset[5], recording gas sensor responses with 13,910 instances and 129 features, and (4) the Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set (HAPT) dataset[4], detailing human activities and postural transitions with 10,929 instances and 561 features.

3.2 Experiment Configurations

The proposed GP model undergoes 20 runs, each with up to 150 generations, using a population of 25 individuals limited to five genes each to balance complexity and computational efficiency. Trees are initialized using the 'Half-and-Half' method and restricted to a depth of 10 to prevent overfitting. A tournament selection with an elite fraction of 0.05 ensures the retention of superior genes. The model incorporates a 0.05 fraction for Boruta recombination, enhancing gene selection with a high crossover probability of 0.8 and a mutation probability of 0.1. An ANN using Cross-Entropy loss for evaluation and specific configurations used for classification and the Boruta selection mechanism.

4 DISCUSSION AND RESULT

The effectiveness of the proposed method was compared with a standard GP model through 20 independent evaluations. Each run divided the dataset into 60% for training, 10% for validation, and 30% for testing. The fitness of the BEI was monitored throughout each run, alongside the peak fitness within the population for each generation. As depicted in Figure 1, the evolution of fitness values across generations for a randomly selected run is presented. Initially, the BEI's fitness may lag behind that of the best randomly generated individuals. However, with progression through generations, the BEI's fitness surpasses that of the population's peak fitness. Despite occasional surpassing by some individuals within the population, the BEI, through the recombination mechanism intrinsic to the proposed method, not only enhances the fitness of those individuals but also steers the population towards optimal solutions. This leadership role of the BEI becomes particularly apparent in the latter generations, consistently emerging as the fittest individual. This phenomenon was consistently observed across all 80 runs for the four datasets, with the BEI concluding as the fittest individual. For more comprehensive view of all runs, please refer to the Supplementary Document, Figures S2, S3, S4, and S5.

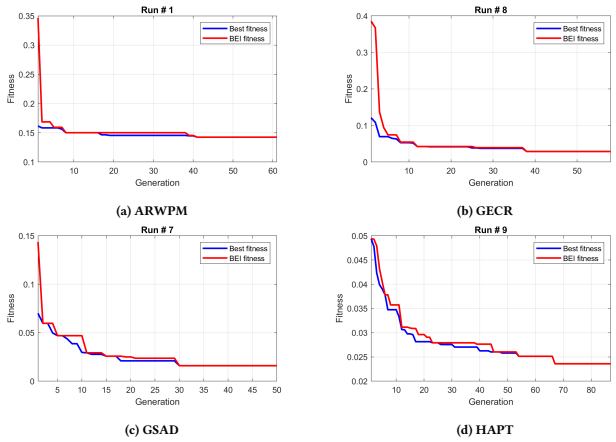
Cross-entropy distribution comparisons, depicted in Figure 2, revealed proposed method superior performance over the baseline GP, particularly in training and testing phases. The limited size of the GECR dataset and its allocation for validation introduced variability, yet proposed method showed consistent outperformance in training and testing for this dataset too.

Statistical comparison using the Wilcoxon rank-sum test, with results summarized in Table 1, confirmed proposed method superiority, with its results outperforming the baseline across all datasets except for the test stage of the GECR dataset.

Analysis of prediction entropy, shown in Figure 3, demonstrated proposed method enhanced prediction confidence, with a noticeable pattern of lower entropy values for correct predictions and reduced high-entropy incorrect predictions. This suggests proposed method refined capability for uncertainty-aware decision-making, making it superior in providing more reliable outcomes compared to the baseline GP model.

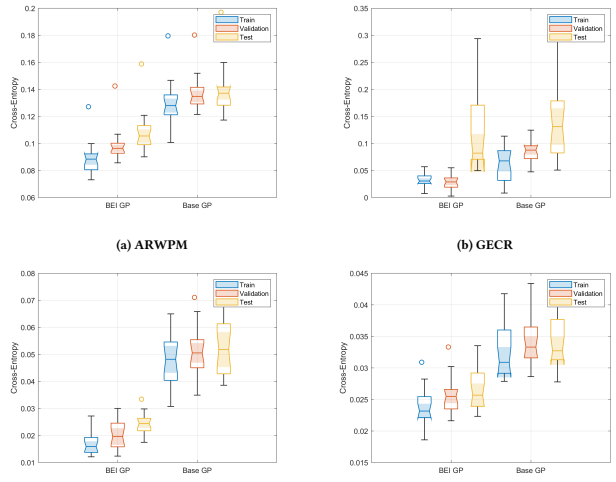
Table 1: Comparison of cross-entropy loss.; Highlighted shows BEI outperforms base models

		ARWPM	GECR	GSAD	HAPT
BEI	Train	**0.0883 ± 0.0117**	**0.0327 ± 0.0118**	**0.0173 ± 0.0044**	**0.0239 ± 0.0029**
	Val	**0.0982 ± 0.0115**	**0.0272 ± 0.0139**	**0.0201 ± 0.0051**	**0.0255 ± 0.0027**
	Test	**0.1087 ± 0.0146**	0.1218 ± 0.0716	**0.0243 ± 0.0036**	**0.0268 ± 0.0034**
Base	Train	0.1300 ± 0.0158	0.0627 ± 0.0320	0.0475 ± 0.0099	0.0325 ± 0.0040
	Val	0.1371 ± 0.0127	0.0859 ± 0.0196	0.0509 ± 0.0092	0.0343 ± 0.0036
	Test	0.1392 ± 0.0173	0.1412 ± 0.0693	0.0520 ± 0.0100	0.0344 ± 0.0042
P-Value	Train	2.56×10^{-7}	0.004320	6.8×10^{-8}	3.94×10^{-7}
	Val	6.92×10^{-7}	1.06×10^{-7}	6.8×10^{-8}	3.42×10^{-7}
	Test	1.38×10^{-6}	0.239323	6.8×10^{-8}	5.87×10^{-6}



(a) ARWPM (b) GECR (c) GSAD (d) HAPT

Figure 1: Fitness evolution of BEI and the top individual for a randomly selected run and all generations.

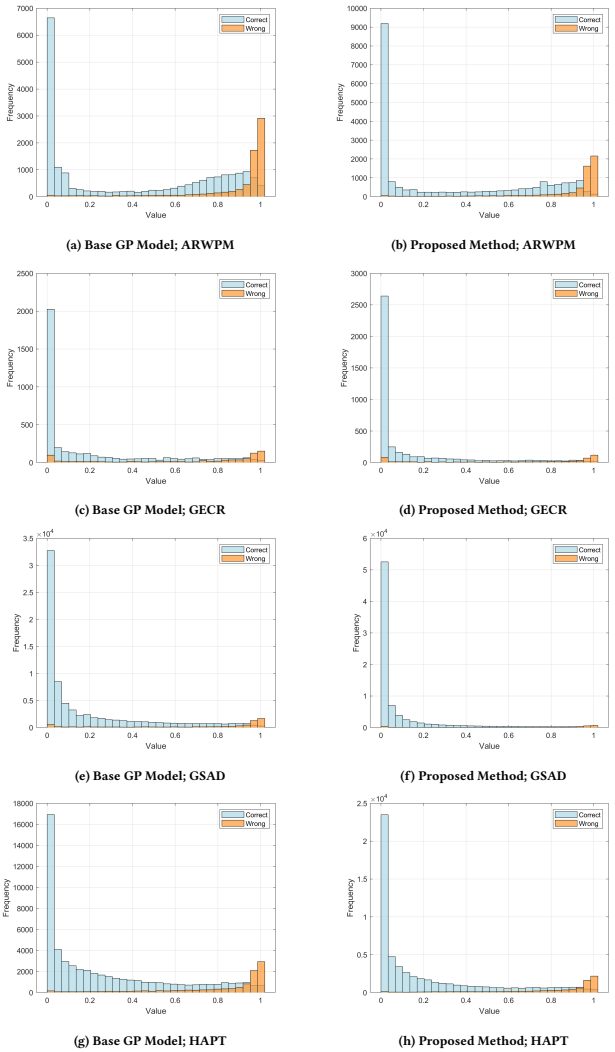


(a) ARWPM (b) GECR (c) GSAD (d) HAPT

Figure 2: Boxplots show cross-entropy in proposed method vs. baseline GP over 20 runs.

5 CONCLUSION

This paper presents a novel approach using a modified Boruta feature selection method that significantly outperforms traditional GP models by enhancing populations. Future work will explore combining UQ criteria to further refine populations over generations.



(a) Base GP Model; ARWPM (b) Proposed Method; ARWPM (c) Base GP Model; GECR (d) Proposed Method; GECR (e) Base GP Model; GSAD (f) Proposed Method; GSAD (g) Base GP Model; HAPT (h) Proposed Method; HAPT

Figure 3: Histogram of the predictive entropy results.

ACKNOWLEDGEMENTS

This work was supported by the Australian Government through the Australian Research Council under Project DE210101808.

REFERENCES

[1] 2019. Activity recognition using wearable physiological measurements. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5RK6V.
[2] Samuele Fiorini. 2016. gene expression cancer RNA-Seq. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5R88H.
[3] Lance Kaplan, Federico Cerutti, Murat Sensoy, Alun Preece, and Paul Sullivan. 2018. Uncertainty aware AI ML: why and how. *arXiv preprint arXiv:1809.07882* (2018).
[4] Anguita Davide Oneto Luca Reyes-Ortiz, Jorge and Xavier Parra. 2015. Smartphone-Based Recognition of Human Activities and Postural Transitions. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54G7M.
[5] Alexander Vergara. 2013. Gas Sensor Array Drift Dataset at Different Concentrations. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MK6M.