An agent-based approach to predicting lymph node metastasis status in breast cancer

Sean Grimes
Department of Computer Science
Drexel University
Philadelphia, PA, USA
spg63@drexel.edu

Mark D. Zarella, Ph.D.
Department of Pathology and Laboratory Medicine
Drexel University College of Medicine
Philadelphia, PA, USA
mark.zarella@jhu.edu

Fernando U. Garcia, M.D.
Department of Pathology
Tower Health
Reading, PA, USA
fernando.garcia@towerhealth.org

David E. Breen, Ph.D.
Department of Computer Science
Drexel University
Philadelphia, PA, USA
david@cs.drexel.edu

***Abstract*—We present a flexible, multi-agent approach to predictive classification problems which uses simple, modular agents that interact and share information socially in an arena with a variable number of participants. Opinion aggregation is accomplished using a honey-bee-derived optimization algorithm that improves accuracy and reduces variance compared with existing weighted and unweighted voter mechanisms. Confidence metrics may be derived from the agent interactions. We apply our system to a data set of 483 de-identified breast cancer patients to predict node-positive or node-negative disease with over 78.5% accuracy in general. When eliminating low-confidence predictions, which leaves 79.5% of patients, classification accuracy improves to 84.5%.**

***Index Terms*—classification, prediction, multi-agent, collective-intelligence, swarm, nomogram, wisdom-of-crowds, breast cancer**

I. Introduction

This work presents an alternative to artificial neural networks (ANNs). The basis for this approach, as previously reported in [1], can be found in both prediction markets (PMs) and wisdom-of-crowds (WoC) theories. Prediction markets aim to determine the probability of a future event by collecting truthful input from agents (human or computer), aggregating that information and forming a collective knowledge [2]. However, PMs expect participants to be well-informed agents, something that is both 'hard' and computationally infeasible for any sufficiently large system [3]. WoC presents an alternative to PMs, where participants are not required or expected to be well-informed. Indeed, WoC relies on a more diverse crowd to arrive at a small predictive error [4]. In order to generate a correct prediction, a WoC system requires an advanced aggregation mechanism to elicit an overall prediction from the crowd of participants [5]. Previous work has investigated Unweighted Mean Model (UWM), Weighted Voter Models (WVM), and other simple aggregation mechanisms with varying levels of success [1], [6], [7]. We now employ a different approach that uses a honey-bee-derived swarm optimization algorithm as an aggregation mechanism to provide improved results.

We apply this work by retrospectively examining tumor characteristics acquired in the routine care of patients at Drexel University College of Medicine to make a binary classification, node-positive or node-negative, that predicts lymph node metastasis status, a determination that could obviate surgical dissection of the lymph nodes. Multiple methods exist for predicting lymph node metastasis in breast cancer patients, the Memorial Sloan-Kettering Cancer Center (MSKCC) nomogram is an early attempt at solving this problem, that uses standard patient characteristics, such as patient age, primary tumor size, presence of lymphatic invasion, histologic grade, among others [8]. The MSKCC has, however, shown inconsistent results across data sets and may not be viable as a tool for decision making in some patient populations [9], [10]. Similarly, we use available patient characteristics to predict whether lymph node metastasis has occurred in breast cancer positive patients.

Further, a swarm aggregation mechanism is used to place predictions in confidence intervals, which leads to an improvement of overall system accuracy for a subset of high-confidence subjects.

II. Methods & Design

A. Patient Features

Our agent-based method was used to predict lymph node metastasis in 483 de-identified breast cancer patients. This research was approved by the Drexel University IRB office, with the data being distributed to the investigators using an honest broker. Our work considered all features which were available for at least 50% of patients, unlike previous work which considered only highly correlated features and required feature completeness [11], [12]. Table I shows the clinical features available for classification. Not all features were available for each patient.

TABLE I
CHARACTERISTICS OF PATIENT POPULATIONS

Feature	$n = 483$	%
Age		
≤ 45	103	21.3
> 45	380	78.7
Primary Tumor Max size (mm)		
≥ 200	5	1.0
100-199	15	3.1
50-99	51	10.6
25-49	125	25.9
0-24	271	56.1
unknown	16	3.3
Angio Lymphatic Invasion		
Absent	127	26.3
Present	200	41.4
Indeterminate	25	5.2
Unknown	131	27.1
pT Stage		
Unknown	36	7.5
pT1	210	43.5
pT2	173	35.9
pT3/pT4	64	13.3
Histologic Grade		
Unknown	33	6.8
1	53	11.0
2	164	34.0
3	233	48.2
Tubule Formation		
Unknown	30	6.2
1 ($> 75\%$)	13	2.7
2 ($10 - 75\%$)	98	20.3
3 ($< 10\%$)	342	70.8
Nuclear Grade		
Unknown	29	6.0
1	20	4.1
2	151	31.3
3	283	58.6
Lobular Extension		
Unknown	202	41.8
Absent	147	30.4
Present	134	27.7
Pagetoid Spread		
Unknown	213	44.1
Absent	177	36.6
Present	93	19.3
Perineureal Invasion		
Unknown	267	55.3
Absent	186	38.5
Present	30	6.2
Calcifications		
Unknown	115	23.8
Absent	126	26.1
Present	176	36.4
Present w/ DCIS	66	13.7
ER Status		
Unknown	51	10.6
Negative	155	32.1
Positive ($> 10\%$)	277	57.3
PR Status		
Unknown	54	11.2
Negative	201	41.6
Positive ($> 10\%$)	228	47.2
P53 Status		
Unknown	81	16.8
Negative	255	52.8
Positive ($> 5\%$)	147	30.4
Ki67 Status		
Unknown	56	11.6
Negative	114	23.6
Positive ($> 14\%$)	313	64.8
Her2 Score		
Unknown	83	17.2
0	119	24.6
1	169	35.0
2	54	11.2
3	58	12.0

In addition to the basic score values for ER, PR, P53, Ki67, and Her2, the agents were given access to the raw count values (when available) for 1+, 2+, 3+, 0 cells, total cells, membrane intensity (where applicable), and positive and negative intensities (where applicable). These values were collected using quantitative image analysis applied to whole-slide images of representatives slides acquired at the time of diagnosis. Slide scanning was done on either Aperio XT or Hamamatsu S210 slide scanner with quantification performed using Aperio Imagescope software and conformed to College of American Pathologists (CAP) recommendations. In testing, none of these values were shown to be highly correlated with the outcome individually, but showed an increase in overall swarm performance when the data were included.

B. Design Overview

Our approach uses simple agents (WoC-Bots) without expert knowledge that interact with each other in a "social interaction arena" to transfer information and formulate opinions [1]. The agents are trained with different, small, subsets of features that describe the classification task at hand. This initially gives us a group of agents with a diverse and independent set of knowledge. Previous work in this area used simple agents built around a small multi-layer perceptron classifier (MLP). During social interactions in the arena, agents learn not just the current prediction of other agents, but also their past performance and trust values. The social interactions allow agents to generate and update trust values associated with each agent they interact with based on prior prediction performance, certainty in the current prediction, and other performance-based metrics. Following an interaction period an overall prediction is generated with an aggregation mechanism that uses trust as one of the weights when determining the amount of votes each agent received. The work presented in this paper uses a similar system and agent design as described above, however each agent uses a 5-layer (3 hidden) MLP classifer instead of the 3 layer classifier described in [1].

Eight to ten randomly selected features are distributed to each agent, with each agent also receiving two highly correlated additional features, "primary tumor size" and "histologic grade". Correlation for "primary tumor size" and "histologic grade" was determined using standard principal component analysis (PCA) externally to the system presented here. The MLP classifier for each agent is first trained, agents are then initialized in an MxN grid-based interaction arena where the final arena size is set such that there are two times as many spaces as there are participating agents. Arena sizing was determined through performance testing, with $2x$ showing a good balance between freedom of movement and agent interaction opportunities.

C. Interaction Period

This work follows the interaction period described in [1], initializing agents randomly in the arena, and ensures that no agents share the same space in the arena at initialization. Agents move throughout the arena in a "Manhattan-like"

fashion, moving one step north, south, east, or west within the bounds of the arena. The direction is randomly selected, agents are not allowed to interact with the same agent two times in a row or more than twice within five separate movements. Agents who cannot move while avoiding these restrictions are "teleported" to a randomly selected empty space within the arena. Further, agents are "teleported" to a random, empty, space every 10-15 iterations to facilitate additional information dispersal. The interaction operates in discrete iterations where each agent is moved during each iteration, and if two agents meet, the interaction between them must finish before the next iteration (and movement) can continue.

The system handles missing data for any given patient by simply disallowing participation of agents that have missing data for the current patient. This method allows us to consider all available information for each patient without requiring data completeness for all patients.

During the interaction period the participating agents interact with each other when two (but not more than two) agents share a single grid space within the arena. The goal of the interaction period is information sharing. Agents share their input features, initial MLP classification results, current prediction, and variables that represent past performance, allowing each agent participating in the interaction to update their current prediction based on shared knowledge. When two agents interact, agent a and agent b, the internal state of the agents will change, updating the agent's current prediction.

Previous work in [1] provides additional details about the interaction period and the equations that govern information sharing between agents and how the interactions change the internal state of each agent.

D. Swarm Aggregation

Prediction aggregation in multi-agent / social systems is an open problem with various different proposed solutions [13]. We tested previously described methods, the Unweighted Mean Model [6] and Weighted Voter Model [7], with both producing inconsistent results; simulation accuracy varied based on the randomized agent initialization within the interaction arena and the order of agent interactions. A trust-based aggregation model described in [1] performed more consistently than either the UWM or WVM, but had, on average, worse accuracy than the WVM for this data set.

Due to the successful aggregation of human opinions found in [14] and [15] using "swarm intelligence", we implemented a swarm-based aggregation mechanism, specifically modeled on honey-bee foraging behavior. Bee colonies are able to forage across a large area in multiple directions, often finding optimal sources for nectar and pollen – a similar task to many optimization problems in computer science. A subset of bees (the scouts) move throughout the colony's foraging area, searching for high yield food sources [16]. Upon returning to the colony the scouts can perform a "waggle dance", advertising the location of the food source to other bees within the colony, which can then forage at the advertised location. The length of time a bee is dancing is generally proportional to the quality of the food source, with additional bees recruited to higher quality food sources due to longer "waggle dances" [17]. This allows for high yielding food sources to be advertised and foraged as long as they remain productive, with a drop off in interest as better sources are found or as the source becomes depleted [18].

Our algorithm is an approximation of the bee behavior described above. The swarming period has a single goal, for all agents to support a single opinion, either 1 or 0, rather than all agents supporting a single 'presenting' agent. 20% of the agents are randomly selected to present their opinion from the overall population of agents; these agents are the 'presenters' and are analogous to the 'scout' bees. The remaining 80% of non-presenting agents are 'watchers'. All agents that participated in the interaction step – those with complete data for the current patient – must participate in the swarming step.

The presenting agents do not perform a dance, simulated or otherwise, instead the watching agents are assigned to support the presenting agents based on roulette wheel selection, also called "fitness proportionate selection" [19]. Each presenting agent is assigned some probability, a_{prob}, between 0 and 1 such that the sum of all probabilities of all presenting agents is 1 after normalization. a_{prob} is calculated by Eq. 1, an equally weighted combination of the presenting agent's `prior_performance`, $a_{priorPerf}$, `confidence`, $a_{confidence}$, and `trust_score`, a_{trust}.

$$a_{prob} = (a_{priorPerf} * a_{confidence} * a_{trust})/3 \qquad (1)$$

Once probabilities have been computed for all presenting agents the roulette wheel selection algorithm assigns watchers to support each presenting agent using a_{prob} such that presenting agents with a high a_{prob} value will be assigned watching agents more frequently than presenters with a low a_{prob} value. In this context, a watching agent "supports" a presenting agent by being assigned to the presenting agent. The watching agent cannot be assigned to multiple presenting agents simultaneously and can be re-assigned to a different presenting agent up to two times if its prediction is different from the presenting agent's prediction **and** if the watching agent has a higher $a_{priorPerf}$ score than the presenting agent. This allows agents with a history of strong performance an opportunity to move to a presenting agent it thinks best represents its predictive belief. After assignment, and re-assignment if applicable, is complete each presenting agent will represent the watchers it is assigned. For example, if a presenter is assigned 10 watchers the presenting agent's opinion will be worth 11 'votes' during the final aggregation step – the 10 watchers and its own prediction.

The swarm goes through a series of iterative steps to arrive at a prediction. That is, the swarm will generate a final binary prediction with an assigned confidence value through repeating the above process of assigning watchers to presenters and taking a vote of all presenters, iteratively lowering the decision threshold after a set number of iterations if the threshold is not met. Initially the threshold is 100% agreement; if the

initial selection of presenting agents all agree on a prediction this is considered a "Very High Confidence" prediction, the prediction is made, and the swarming period is ended. If there is no immediate agreement the agents will perform an interaction period (the same interaction that happens in the interaction arena) with all other agents assigned to the presenting agent they are also assigned to, as well as the presenting agent, to facilitate additional information sharing. A new selection of presenting agents will be randomly selected and the swarm will go through the same steps previously outlined. This process is allowed to run for 100 iterations and the threshold for agreement is lowered to 90%, if at any point 90% or more agents support the same prediction the swarming period is ended and the prediction is considered "High Confidence".

If neither of the previous cases are met the support threshold is lowered to 75% and the swarm is allowed an additional 50 iterations of information sharing and selection. If the 75% support threshold is met during this period the prediction is considered "Medium Confidence". If the swarm has still not arrived at a prediction by this point, a simple unweighted vote is taken from all agents (presenters and watchers) and the prediction is considered "Low Confidence".

The above process allows us to assign a pseudo-confidence value to each prediction based on the number of iterations it takes the swarm to arrive at a conclusion and the percent of agents supporting the conclusion. The confidence values are summarized as:

- Very High Confidence: All agents agree on a prediction class immediately following roulette-wheel selection,
- High Confidence: 90% of agents agree on a prediction class immediately following roulette-wheel selection or after 100 additional iterations,
- Medium Confidence: 75% of agents agree on a prediction class after 100-150 additional iterations,
- Low Confidence: Unweighted vote of presenting and watching agents.

III. Results

A. Swarm Aggregation

Fig. 1 shows 5-fold validation results for predicting lymph node metastasis status from the clinical features listed in Table I, with the output from the WVM on the left and our bee-inspired swarm aggregation method on the right. The folds used to test the WVM and the swarm aggregation method were the same, i.e. they were not re-randomized between tests. The averaged 5-fold results for accuracy, sensitivity, and specificity can be seen in Fig. 1. The swarming method improved average accuracy by 6.2%, the sensitivity by 10.9%, and the specificity by 2.0% compared with the WVM aggregation method. The worst case values for accuracy, sensitivity, and specificity were improved using the swarm method, with a worst-case accuracy in the WVM of 67% in fold 1 improved to 79.4% using the swarm. The bee-inspired swarm method was strictly an improvement for both the accuracy and sensitivity statistics. However, the WVM out-performed the swarm method for specificity in 2 of the 5 folds (folds 2 and 4), despite the 2% average improvement.

TABLE II
Swarm Aggregation vs. Weighted Voter Model Variance (σ^2)

Method	Accuracy (σ^2)	Sensitivity (σ^2)	Specificity (σ^2)
Swarm	2.30	20.99	0.35
WVM	24.52	74.0	20.68

TABLE III
Confidence Interval Distribution and Accuracy

Interval	$n = 483$	% of n	Accuracy (%)
Very High Confidence	3	0.62	100
High Confidence	156	32.3	90.1
Medium Confidence	225	46.6	80.9
Low Confidence	99	20.5	59.7
Very High + High + Medium	384	79.5	84.79

The WVM showed significant variance across each fold, which we aimed to reduce using the swarm aggregation method. Table II shows the variance across the same five folds shown in Fig. 1 for both the WVM and swarm aggregation methods, with the swarm method showing significantly reduced variance for accuracy, sensitivity, and specificity measurements.

B. Confidence Intervals

We applied the four confidence intervals described in Section II-D to the patients in our data set. The results shown in Table III represent the accuracy values for each confidence interval. The "Very High Confidence" category captured only 0.62% of patients and was not considered useful for this data set. The "High Confidence" category captured 32.3% of patients, with an accuracy of 90.1% and the "Medium Confidence" category captured 46.6% of patients with an average accuracy of 80.9%. 20.5% of patients fell into the "Low Confidence" category with an accuracy of 59.7%. If we eliminate all "Low Confidence" predictions (20.5% or 99 patients) from the system we are able to achieve an accuracy of 84.8%, accounting for 384 of the original 483 patients (79.5%). This demonstrates our approach's ability to stratify the subjects into confidence categories that improve the overall accuracy of the prediction.

IV. Conclusion & Future Work

We have demonstrated a flexible, multi-agent approach to binary classification problems, applying our system based on wisdom-of-crowds and swarm intelligence to a data set of breast-cancer positive patients, predicting node-positive or node-negative disease with a combined accuracy of 84.8% for the "Very High", "High", and "Medium" confidence intervals, capturing 79.5% of patients. The honey-bee-inspired aggregation mechanism has improved the accuracy, specificity, and sensitivity of our system over existing aggregation methods, as well as reducing the prediction variance.

As with any classification or learning system, we expect better performance with more training data. We intend to work

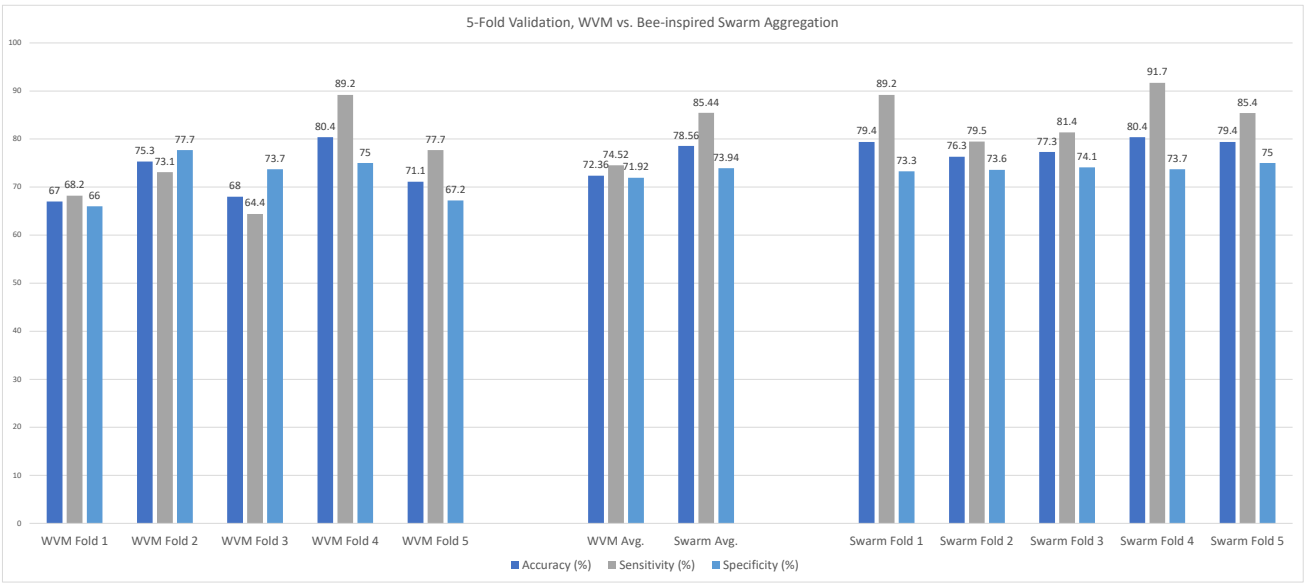

Fig. 1. 5-fold validation results for predicting lymph node metastasis status from the clinical features listed in Table I. Weighted Voter Model (left) vs. Swarm Aggregation (right)

with other breast-cancer positive data sets to explore both performance increases with more data, as well as to validate the system against other data sets. Additional data will also allow us to further explore the confidence interval sub-system with the goal of making the "Very High Confidence" category more inclusive without sacrificing much, if any, accuracy. Further, we would like to apply this method to a data set of pre-surgical features, relying on information available at time of biopsy in order to help guide pre-surgical decision making. Previous work using an agent-based system has shown this approach is resistant to feature dropout. We aim to do further testing using the breast cancer data set presented here to confirm the prior results as we remove protein expression data such as P53 or Her2, since they may not typically be collected.

REFERENCES

[1] S. Grimes and D. E. Breen, "Woc-bots: An agent-based approach to decision-making." *Applied Sciences (2076-3417)*, vol. 9, no. 21, 2019.

[2] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry, "The wisdom of the crowd in combinatorial problems," *Cognitive science*, vol. 36, no. 3, pp. 452–470, 2012.

[3] A. Othman, "Zero-intelligence agents in prediction markets," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*. Citeseer, 2008, pp. 879–886.

[4] E. Ostrom, "The difference: How the power of diversity creates better groups, firms, schools, and societies. by scott e. page. princeton: Princeton university press, 2007. 448p. 19.95 paper." *Perspectives on Politics*, vol. 6, no. 4, pp. 828–829, 2008.

[5] Ş. Ertekin, C. Rudin, and H. Hirsh, "Approximating the crowd," *Data Mining and Knowledge Discovery*, vol. 28, no. 5, pp. 1189–1221, 2014.

[6] R. Hastie and T. Kameda, "The robust beauty of majority rules in group decisions." *Psychological Review*, vol. 112, no. 2, p. 494, 2005.

[7] G. Valentini, H. Hamann, and M. Dorigo, "Self-organized collective decision making: The weighted voter model," in *Proc. International Conference on Autonomous Agents and Multi-agent Systems*, 2014, pp. 45–52.

[8] M. L. Smidt, D. M. Kuster, G. J. van der Wilt, F. B. Thunnissen, K. J. Van Zee, and L. J. Strobbe, "Can the memorial sloan-kettering cancer center nomogram predict the likelihood of nonsentinel lymph node metastases in breast cancer patients in the netherlands?" *Annals of surgical oncology*, vol. 12, no. 12, pp. 1066–1072, 2005.

[9] M. D. Zarella, D. E. Breen, A. Reza, A. Milutinovic, and F. U. Garcia, "Lymph node metastasis status in breast carcinoma can be predicted via image analysis of tumor histology." *Analytical and quantitative cytopathology and histopathology*, vol. 37, no. 5, pp. 273–285, 2015.

[10] I. Van den Hoven, G. P. Kuijt, A. Voogd, M. van Beek, and R. Roumen, "Value of memorial sloan-kettering cancer center nomogram in clinical decision making for sentinel lymph node-positive breast cancer," *Journal of British Surgery*, vol. 97, no. 11, pp. 1653–1658, 2010.

[11] K. J. Van Zee, D.-M. E. Manasseh, J. L. Bevilacqua, S. K. Boolbol, J. V. Fey, L. K. Tan, P. I. Borgen, H. S. Cody, and M. W. Kattan, "A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy," *Annals of surgical oncology*, vol. 10, no. 10, pp. 1140–1151, 2003.

[12] H. E. Kohrt, R. A. Olshen, H. R. Bermas, W. H. Goodson, D. J. Wood, S. Henry, R. V. Rouse, L. Bailey, V. J. Philben, F. M. Dirbas *et al.*, "New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients," *BMC cancer*, vol. 8, no. 1, pp. 1–15, 2008.

[13] Q. Du, H. Hong, G. A. Wang, P. Wang, and W. Fan, "Crowdiq: A new opinion aggregation model," in *Proc. 50th Hawaii International Conference on System Sciences*, 2017.

[14] L. Rosenberg, N. Pescetelli, and G. Willcox, "Artificial swarm intelligence amplifies accuracy when predicting financial markets," in *Proc. IEEE 8th Annual Conference on Ubiquitous Computing, Electronics and Mobile Communication*, 2017, pp. 58–62.

[15] L. Rosenberg, "Artificial swarm intelligence, a human-in-the-loop approach to AI," in *Proc. 13th AAAI Conference on Artificial Intelligence*, 2016.

[16] V. Tereshko and A. Loengarov, "Collective decision making in honey-bee foraging dynamics," *Computing and Information Systems*, vol. 9, no. 3, p. 1, 2005.

[17] K. Von Frisch, *The dance language and orientation of bees*. Harvard University Press, 2013.

[18] M. Beekman and F. Ratnieks, "Long-range foraging by the honey-bee, apis mellifera l." *Functional Ecology*, vol. 14, no. 4, pp. 490–496, 2000.

[19] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2193–2196, 2012.