JOURNAL OF
EMERGING INVESTIGATORS

Article

Simple solving heuristics improve the accuracy of sudoku difficulty classifiers

Madeleine Higgins[1], **Clayton Greenberg**[2], **Angeline Huang**[1]

[1] Brearley School, New York, New York

[2] Saarland University, Saarbrücken, Germany

SUMMARY

Sudokus are logical puzzles that vary in difficulty. They are typically solved with some commonly accepted strategies such as "obvious singles," "hidden singles," and "naked pairs." Since strategies are an important part of how solvers approach sudoku puzzles, solving strategy analysis may be useful in building sudoku difficulty classifiers. Our study aimed to improve the accuracy of sudoku difficulty classification by analyzing the predictive power of 17 variables, including metrics based on sudoku solving strategies. Other classification attempts have been made using a convolutional neural network, a model trained on real-time human solving patterns, and a rating system based on the number of solving rounds needed to solve a puzzle. We collected 6,000 sudoku puzzles from puzzle-sudoku.com for our study. We based our classifier on the website's difficulty ratings, which were basic, easy, intermediate, advanced, extreme, and evil. We paired the levels together so that our classifier only distinguished between three levels of difficulty. We trained two models; the Simple Model was trained on the number of clues, the average possibilities per empty cell, clue variation, and clue placement. The Simple Model had a 44% testing accuracy. The Solving Strategies Model was trained on all variables from the Simple Model and eight additional solving strategies features. These features measured the accuracy of the "obvious singles," "hidden singles," and "naked pairs" strategies on the puzzles. We hypothesized that including these solving strategies variables would improve accuracy in classifying sudoku difficulty because they reflect human solving behavior. The Solving Strategies Model classified between the three difficulty levels with 78% testing accuracy, significantly higher than the Simple Model's accuracy. This result indicates that the sudoku strategy metrics improved the model's ability to classify sudoku difficulty.

INTRODUCTION

Sudoku puzzles have been studied by both mathematicians and computer scientists for their combinatorial complexity and the challenges they pose in solving and classification. They are often modeled as variants of Latin squares, which are n x n grids for which each row and column contains numbers 1 to n (1). Sudokus, however, have the additional constraint of their 3x3 subgrids. One sudoku-related query is the number of possible sudoku grids, which has been approached with graph theory and brute-force calculations (1-2). Researchers have long been studying the minimum number of clues needed for a solvable puzzle, proven to be 17, and the maximum number of clues for an unsolvable puzzle, proven to be 79 (3-4).

Our study focused on classifying sudoku difficulty. Identifying factors that predict sudoku difficulty is a complex task because it involves computationally quantifying a human's perception of a puzzle's challenge. Humans and computers approach sudokus differently; computers often rely on methods like backtracking, which involves systematic trial and error, or SAT solvers, computer programs that can solve any Boolean satisfiability problem (5). To be solvable with SAT methods, sudoku puzzles must be transformed into mathematical formulas that intake Boolean inputs and return "True" or "False" (5). Simulated annealing has also been tested as an efficient way for computers to solve sudoku puzzles (6). Simulated annealing involves determining a space of possible solutions and shrinking it using a cost function and a positive temperature parameter.

Unlike computer solvers, humans commonly rely on shorthand strategies to solve sudoku puzzles. For instance, the "obvious singles" strategy solves a cell by eliminating all numbers that are in its row, column, or 3x3 box (**Figure 1**)**.** The "hidden singles" strategy solves a cell when it has a possible value not present in any other cell in its row, column, or 3x3 box (**Figure 1**). "Naked pairs" is a more complex strategy which relies on locating groups of 2 or 3 cells in the same row, column, or 3x3 box for which the number of cells in the group is equal to the total number possibilities of the group (**Figure 2**).

For difficulty classifiers to be used in a practical setting, they should classify based on a human solver's ability rather than that of a computer and they should also maximize computational efficiency. Some past approaches have fallen short in one of these areas. For instance, a convolutional neural network (CNN) has been used to classify sudoku difficulty into three levels (easy, medium, and difficult) with 80% accuracy (7). CNNs are neural network algorithms that learn by performing unsupervised feature engineering using kernel optimization (7). This model, however, was trained on how many steps were taken by a backtracking sudoku solver. While the CNN is efficient, it has not been proven capable of classifying based on human solving difficulty. Additionally, it is difficult to interpret how CNNs make decisions due to what some call their "black-box" nature, since their inner workings are often not intuitive. Thus, using CNNs may limit our understanding of the predictors of sudoku difficulty.

Prior work has also shown that a computational model trained on real-time human solving patterns can classify

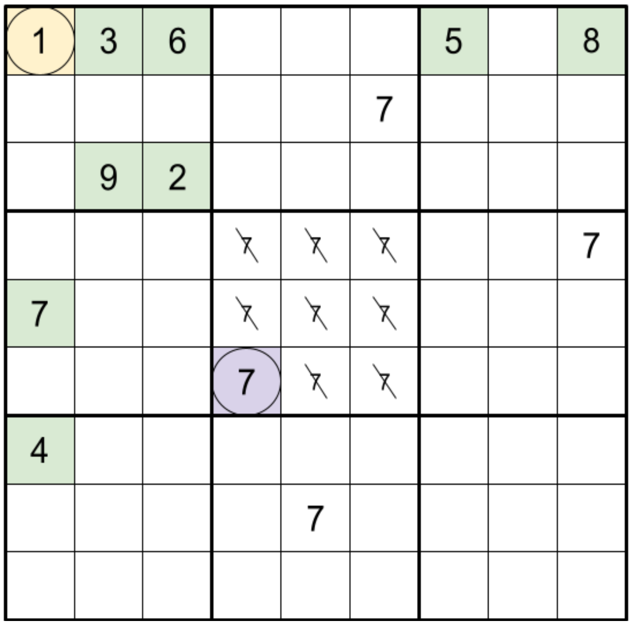


Figure 1: Sudoku board displaying "obvious single" and "hidden single" solving strategies. The yellow highlighted cell is filled in "1" using obvious single strategy (the green highlighted cells eliminate all other possibilities). The purple highlighted cell is filled in "7" using hidden single strategy (no other cell in its 3x3 box has possibility "7").

sudoku difficulty well when compared to human solving times (8). Another paper suggested classifying sudoku difficulty by counting the number of solving rounds (a "round" being the number of simple moves or one high-level move) necessary to solve a puzzle (9). While this method prioritizes human solving ability, it is less efficient than other methods because the classification process requires the complete solving of each sudoku. This study also found that the number of clues somewhat correlates to the level of solving strategy needed to solve the sudoku (9).

In this study, we investigated how different variables, especially variables relating to the efficacy of solving strategies on sudoku puzzles, predict sudoku difficulty. Our initial dataset was 6,000 sudoku puzzles from puzzle-sudoku.com and their given difficulty ratings (10). Here, we used two models: the Simple Model, which was trained only on the number of clues and other easily quantified difficulty metrics, and the Solving Strategies Model, which was trained on additional features pertaining to the efficacy of several solving strategies on the puzzles. These variables were determined primarily by our computer program performing one or two rounds of a strategy on a puzzle and determining the number of cells solved.

Examining differences in accuracy between the two models enabled us to understand the effects of using solving strategies variables. Both models were random forest classifiers, which are simple, fast, and effective on datasets for which there is little knowledge; they also tend to perform better on smaller datasets like ours than CNNs or other neural networks. To maximize efficiency, we used only simple solving strategies, such as "hidden singles", "obvious singles", and "naked pairs." Our computer program also did not completely solve the puzzles, saving time.

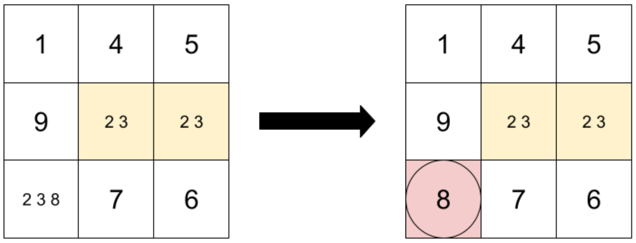


Figure 2: "Naked pair" method in 3x3 box. Two of the cells in the 3x3 box only have possibilities "2" and "3", and therefore the third empty cell cannot be filled with either. This red cell can only be an "8".

Our hypothesis was that variables pertaining to the efficacy of various solving strategies on sudoku puzzles would improve predictions of sudoku difficulty. When classifying between three levels of difficulty, the Solving Strategies model performed with significantly higher testing accuracy than the Simple Model, affirming our hypothesis. By training more models with different combinations of variables, we discovered the particular effectiveness of using the "naked pairs" strategy in classifying sudoku difficulty. We speculate that the Solving Strategies Model was effective in its classification because it considers human solving practices.

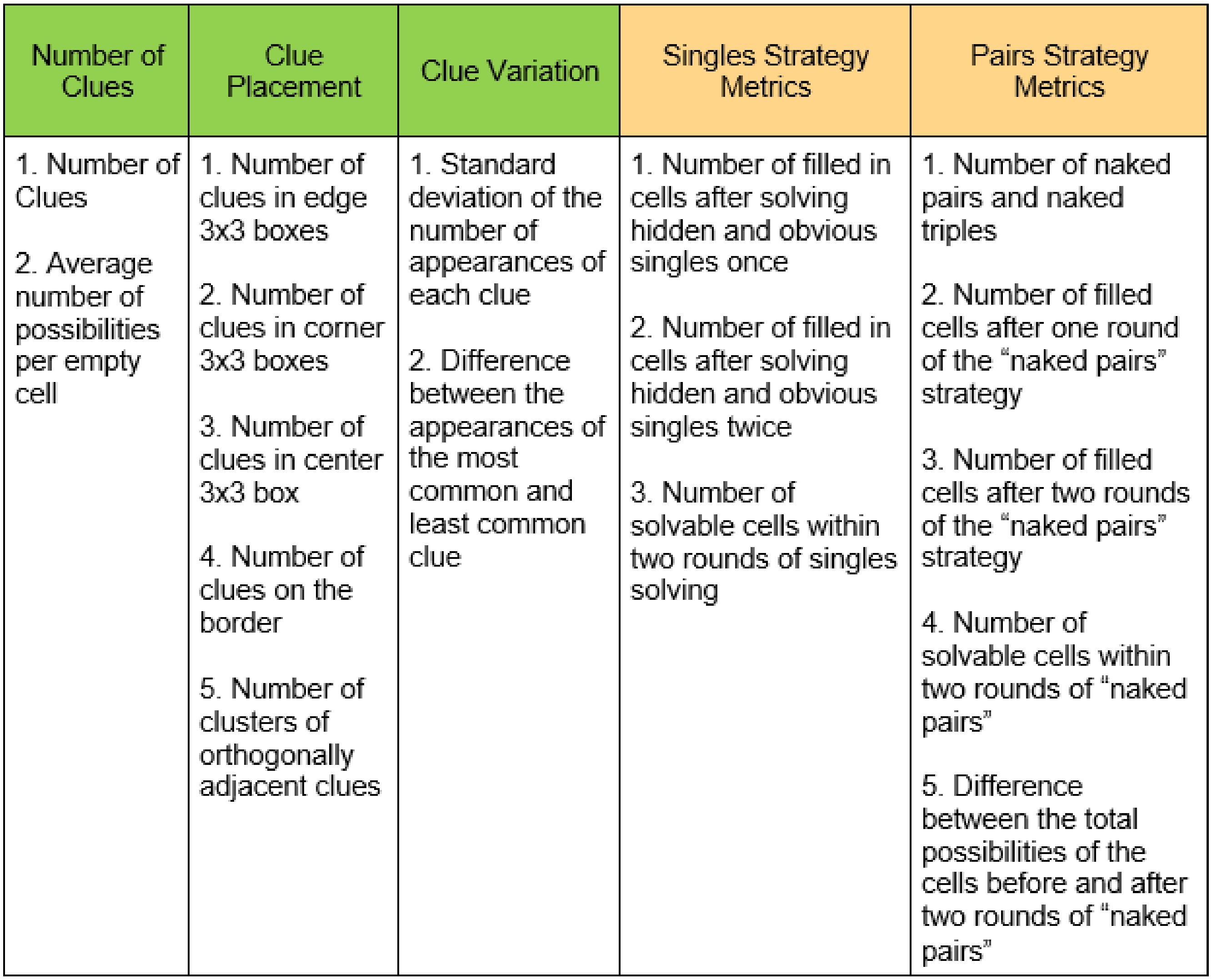
RESULTS

To examine the predictive power of different types of features for determining sudoku difficulty, we computed 17 predictive variables for sudoku difficulty from 6,000 puzzles from puzzle-sudoku.com **(Table 1**). The first nine variables, the simple metrics, considered the number, variation, and placement of clues. The eight solving strategy variables related to the efficacy of the "obvious singles", "hidden singles", and "naked pairs" strategies on the puzzles (**Figures 1,2**). They were primarily calculated by performing one or two "rounds" of a certain strategy on the puzzles and counting how many cells were filled in the process.

The classification task had three classes (0, 1, and 2) based on the six difficulty ratings given on puzzle-sudoku.com (basic, easy, intermediate, advanced, extreme, and evil). Each class grouped together two difficulty levels from the website. For example, basic and easy difficulties were grouped together into class 0. Our first step was training two random forest classifiers, one that excluded solving strategies variables (the Simple Model) and one that included all variables (the Solving Strategies Model). For each classifier, we split the data into 80% training and 20% testing, with both training and testing data evenly divided between the three difficulty classes. The Simple Model performed with 44% accuracy on the testing data, and the Solving Strategies Model performed with 78% accuracy on the testing data; both models performed with 100% accuracy on the training data (**Table 1**). The difference between the accuracies of the two models was statistically significant (p<0.05).

We then compared the difficulty predictions of the Solving Strategies Model on the testing data to the true difficulty labels in a confusion matrix; these results showed that the Solving Strategies Model had the highest accuracy identifying Class 0 sudokus and the lowest accuracy identifying Class 1. The most common mistakes made by the model were in differentiating between Classes 1 and 2 (**Figure 4**).

https://doi.org/10.59720/24-144



Number of Clues	Clue Placement	Clue Variation	Singles Strategy Metrics	Pairs Strategy Metrics
1. Number of Clues 2. Average number of possibilities per empty cell	1. Number of clues in edge 3x3 boxes 2. Number of clues in corner 3x3 boxes 3. Number of clues in center 3x3 box 4. Number of clues on the border 5. Number of clusters of orthogonally adjacent clues	1. Standard deviation of the number of appearances of each clue 2. Difference between the appearances of the most common and least common clue	1. Number of filled in cells after solving hidden and obvious singles once 2. Number of filled in cells after solving hidden and obvious singles twice 3. Number of solvable cells within two rounds of singles solving	1. Number of naked pairs and naked triples 2. Number of filled cells after one round of the "naked pairs" strategy 3. Number of filled cells after two rounds of the "naked pairs" strategy 4. Number of solvable cells within two rounds of "naked pairs" 5. Difference between the total possibilities of the cells before and after two rounds of "naked pairs"

Table 1: List of features examined for determining sudoku puzzle difficulty. Green indicates simple metrics, and yellow indicates solving strategy metrics.

To determine which features in the Solving Strategies Model were the most powerful predictors of sudoku difficulty, we calculated the mean decrease in impurity (MDI) for each feature. MDI measures the importance of a feature in a model. The feature with the highest MDI, 0.314, was the number of solvable cells within two rounds of the strategy "naked pairs." We also trained 17 new models, each with only one feature from the Solving Strategies Model. Of these new models, the feature that classified with the highest accuracy of 57% was the same "naked pairs" feature which had the highest MDI (**Figure 3**). The models with features pertaining to pairs solving, singles solving, and simple metrics classified with accuracies 46–57%, 44–45%, and 33–42%, respectively (**Figure 3**).

Finally, to determine which of the eight solving strategy variables were significant in improving the accuracy from the Simple Model to the Solving Strategies Model, we trained eight models that included one solving strategy variable alongside all variables from the Simple Model. Only one strategy feature significantly improved the accuracy of the Simple Model: the total number of solvable cells within two rounds of "naked pairs" ($p<0.05$) (**Table 2**). The feature which was the number of naked pairs and triples also trended towards improving the model, although the improvement was not statistically significant ($p=0.056$) (**Table 2**).

DISCUSSION

We aimed to classify sudoku difficulty by testing different variables for their ability to predict sudoku difficulty. We especially tested variables that reflect the efficacy of various simple solving strategies on sudoku puzzles. We trained the Simple Model on variables relating to the number, placement, and variation of clues and the Solving Strategies Model on variables relating to the efficacy of solving strategies. The Solving Strategies Model and the Simple Model had testing accuracies of 78% and 44%, respectively, a difference proven to be statistically significant. We then trained more models on different combinations of variables to determine which specific variables were the most powerful in predicting sudoku difficulty.

https://doi.org/10.59720/24-144

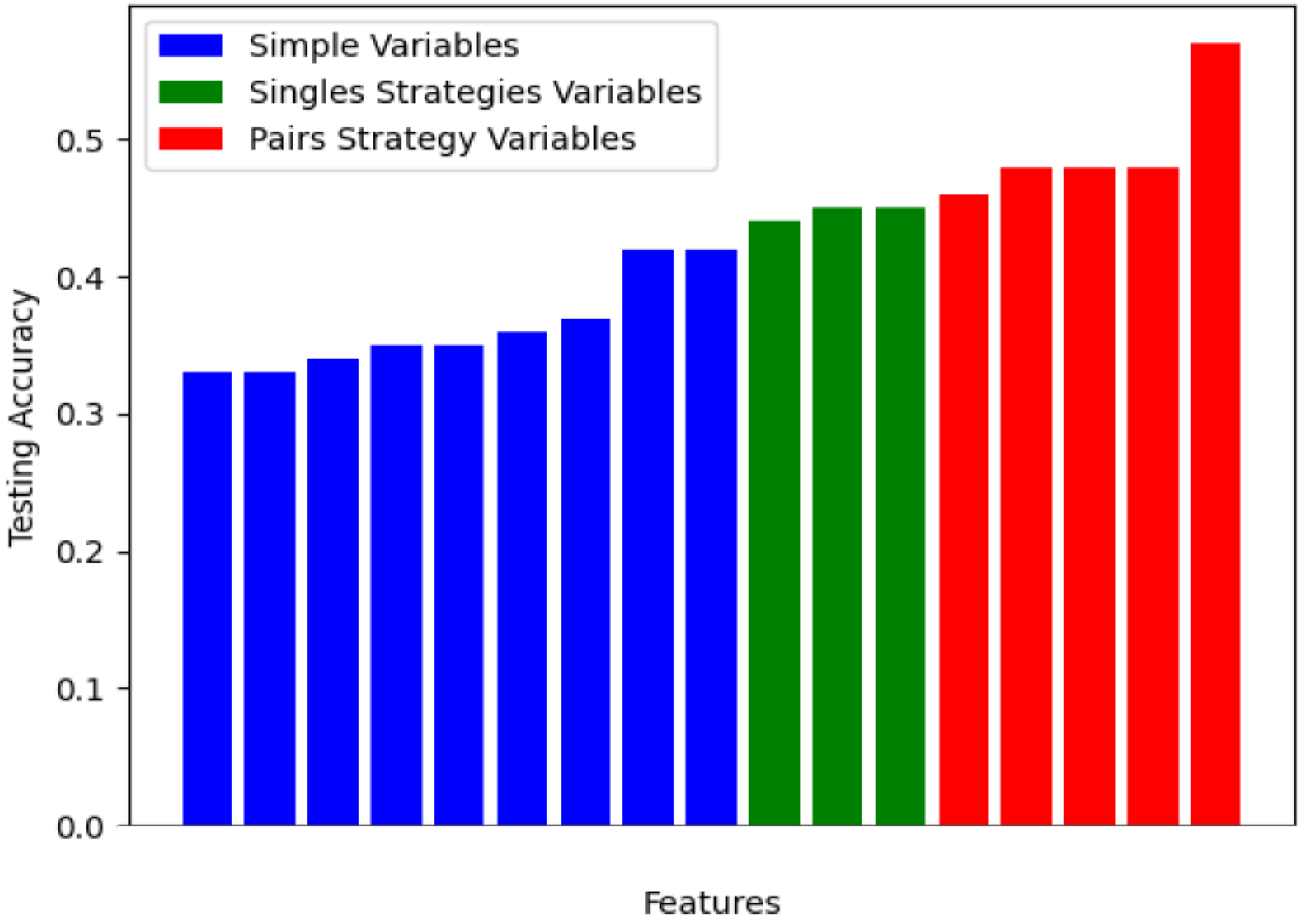


Figure 3: Testing accuracies of random forest models trained on individual variables from the Solving Strategies Model. The model included three categories of variables and the variables based on the "naked pairs" strategy performed with the highest accuracy.

The Simple Model had a testing accuracy of 44% when classifying between the three difficulty levels. Therefore, while the number of clues, average number of possibilities per empty cell, clue placement, and clue variation may be somewhat correlated to sudoku difficulty, these variables alone predicted sudoku difficulty poorly. The Solving Strategies Model's significantly higher testing accuracy of 78% indicates that including solving strategy variables improved the predictions of sudoku difficulty. The 100% training accuracies of both models suggest that there was likely some overfitting, but the final 78% testing accuracy indicates that the Solving Strategies Model went beyond memorizing training examples. The effectiveness of the solving strategy variables may be caused by their link to human solving behavior.

Only one solving strategy variable, the number of solvable cells within two rounds of the "naked pairs" strategy, significantly improved the Simple Model by itself. The number of naked pairs and triples trended towards improving the model, but the difference was not significant. None of the variables pertaining to the singles solving strategies, "obvious singles" and "hidden singles," significantly improved the accuracy of the Simple Model, indicating that the "naked pairs" strategy variables had greater predictive power. The superiority of the "naked pairs" variables was also supported by their higher MDI values and individual predictive accuracies. The greater complexity of the "naked pairs" strategy might explain its better performance within the Solving Strategies model, indicating that more difficult strategies may better predict puzzle difficulty.

Our experimentation had several limitations and sources of bias. For example, we measured only five values to represent clue placement and only two values to represent clue variation. Additionally, we accounted for only a few of the many sudoku solving strategies in this study. While the accuracy of the final model was 78%, the model accuracy decreased when distinguishing between difficulty Classes 1 and 2 (intermediate/advanced and extreme/evil). This is likely because there is more similarity between those difficulty levels, at least by our metrics. Perhaps including more complex strategies in the study, such as Swordfish

Model	Training Accuracy	Testing Accuracy
Simple Model	100%	44%
Solving Strategies Model	100%	78%

Table 2: Testing accuracies for the Simple Model and the Solving Strategies Model. The Simple Model was trained on number of clues, clue placement, clue variation, and the average number of possibilities per empty cell. The Solving Strategies Model contained these features and features related to solving strategies. The Solving Strategies Model had significantly higher testing accuracy according to the Stuart-Maxwell test (p<0.05).

https://doi.org/10.59720/24-144

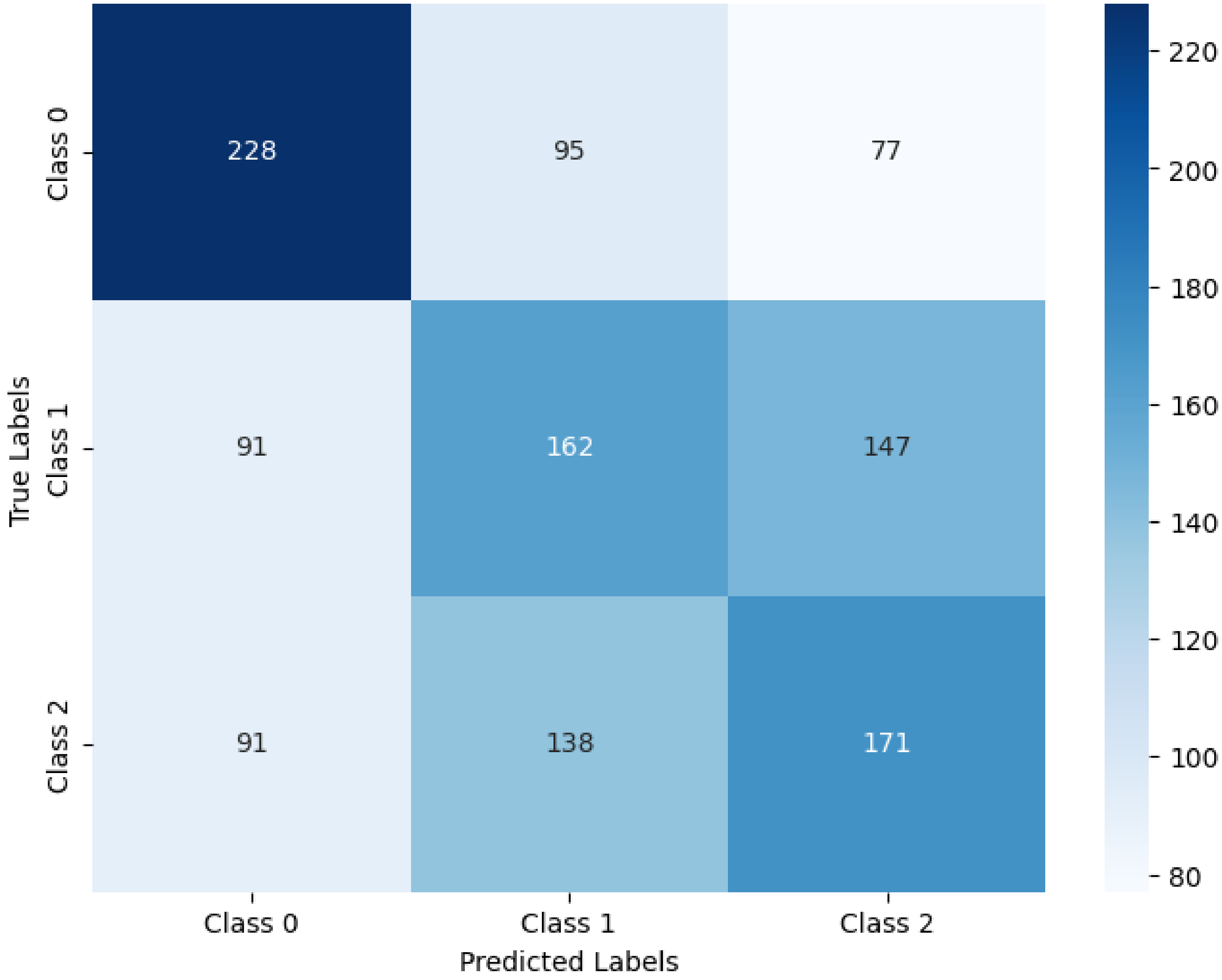


Figure 4: Confusion matrix showing predictions of Solving Strategies Model and the true difficulty classes for the testing data. Class 0 is the easiest level, Class 1 is medium, and Class 2 is the most difficult. The model had greater success identifying Class 0 than the other classes and appears to have difficulty distinguishing between Classes 1 and 2.

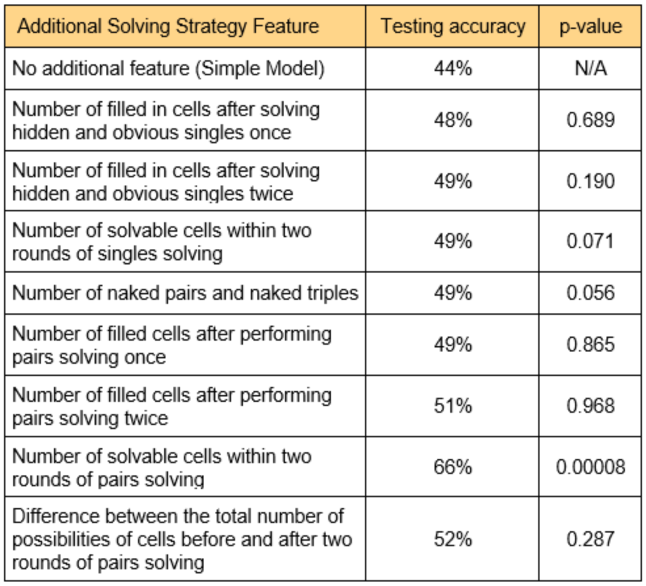
or X-wings, could increase the model's ability to distinguish more difficult puzzles where those strategies might be more applicable. Finally, we only used sudokus and ratings from one site, puzzle-sudoku.com, which likely biased our results towards this website's system for separating levels. Other sources of puzzles may not use the exact same criteria to separate sudokus by difficulty level.

Our results suggested that considering human solving strategies improves accuracy in classifying sudoku difficulty. We also saw that those solving strategies can be analyzed with just a few rounds of solving. It is difficult to compare these results to those of other studies because sudoku difficulty classifiers often differ in their source of base ratings. We relied on ratings from puzzle-sudoku.com to label our training data, but others have used the number of steps taken by a backtracking solver or solving times by real human solvers (7-8). Future experiments could consider more solving strategies of greater complexity, collect more variables relating to clue placement and clue variation, or gather a larger dataset of sudokus and sudoku difficulty ratings from diverse sources.

Our method for quantifying solving strategy usefulness has proven to be successful in determining sudoku difficulty, a somewhat subjective quality. Understanding the causes of differences in sudoku difficulty may help illuminate how humans approach logic puzzles in general and other cognitive tasks which involve strategic methods and logical thinking.

MATERIALS AND METHODS

Using Selenium IDE, we scraped each clue and empty space from 6,000 puzzle-sudoku.com puzzles evenly distributed among each of the six levels on the website (basic, easy, intermediate, advanced, extreme, and evil). In this study, we considered 17 variables. Two related to clue variation: the first was the difference between the number of appearances of the most common and least common clue, and the second was the standard deviation of the number of appearances of each number. To account for clue placement, we computed five variables: the number of clues in corner boxes, the number of clues in edge boxes, the number of clues in the center box, the number of clues on the border,



Additional Solving Strategy Feature	Testing accuracy	p-value
No additional feature (Simple Model)	44%	N/A
Number of filled in cells after solving hidden and obvious singles once	48%	0.689
Number of filled in cells after solving hidden and obvious singles twice	49%	0.190
Number of solvable cells within two rounds of singles solving	49%	0.071
Number of naked pairs and naked triples	49%	0.056
Number of filled cells after performing pairs solving once	49%	0.865
Number of filled cells after performing pairs solving twice	51%	0.968
Number of solvable cells within two rounds of pairs solving	66%	0.00008
Difference between the total number of possibilities of cells before and after two rounds of pairs solving	52%	0.287

Table 3: Comparing models trained on all simple metrics and one solving strategy feature. Each random forest model was trained on all of the features in the Simple Model and one feature relating to solving strategies. The p-values, calculated with the Stuart-Maxwell test, show the significance of their difference from the Simple Model.

and the number of blocks of orthogonally adjacent clues.

To account for the efficacy of various solving strategies (obvious singles, hidden singles, naked pairs/triples), we computed a total of eight values. Three of those values measured the efficacy of the singles strategy on a sudoku. First, our program counted the number of filled in cells after solving for all "obvious singles" (cells for which there is only one possibility by process of elimination) and "hidden singles" (cells with a possible value that is not present in any other cell in its 3x3 box, row, or column) (**Figure 1**). The program also counted the number of filled in cells after solving for all obvious and hidden singles two times on a single sudoku, given the fact that more empty cells become obvious or hidden singles after the singles strategy is used once.

The other five solving strategies variables measured the applicability of the "naked pairs" strategy on a sudoku. One value measured was the number of naked pairs and naked triples, which are groups of empty cells in the same row, column, or 3x3 box for which the total number of possibilities among the cells is equal to the number of cells in the group (**Figure 2**). The naked pairs/triples strategy involves eliminating possibilities from cells based on a group of *n* cells in the same box, row, or column with a combined total of *n* possibilities (**Figure 2**). The number of filled in cells after one pair solve, the number of filled cells after two pair solves, and the number of solvable cells within two pair solves were calculated and appended to the dataset, as was the difference between the total number of possibilities of the cells before and after two pair solves.

We used random forest classifiers in all testing. Random forest classifiers divide a dataset into sub-samples and fit a random decision tree classifier to each one. The trees are combined into an ensemble (forest) using weights based on the individual accuracies of the trees. Each random forest classifier used had 160 estimators, a value that was chosen through testing to optimize the testing accuracies of the final model. No maximum depth was added because our dataset was not large enough to warrant one; if this classifier should be used in a practical setting, there may be a need to add a maximum depth to increase the speed of the model. The testing accuracies were calculated as the proportion of sudokus in the testing dataset which were classified into the correct difficulty level. The training accuracies were the proportion of sudokus in the training dataset which were classified into the correct difficulty level. All comparisons between models were tested for statistically significant differences using Stuart-Maxwell statistical tests.

ACKNOWLEDGMENTS

We would like to thank Anna Higgins of Northeastern University who helped us with the cluster comprehension section of the code and the figure formatting.

Received: May 19, 2024
Accepted: September 5, 2024
Published: May 7, 2025

REFERENCES

1. Felgenhauer, B. and F. Jarvis. "Mathematics of Sudoku I." *Mathematical Spectrum*, vol. 39, no. 1, 2006, pp. 15-22.
2. Lala, Chiraag. "Graph theory of Sudoku." *Indian Institute of Science Education and Research Bhopal*, April 2013.
3. McGuire, Gary, *et al.* "There Is No 16-Clue Sudoku: Solving the Sudoku Minimum Number of Clues Problem via Hitting Set Enumeration." *Experimental Mathematics*, vol. 23, no. 2, Jun 2014, pp. 190-217. https://doi.org/10.1080/10586458.2013.870056
4. Delahaye, Jean-Paul. "The Science behind SUDOKU." *Scientific American*, vol. 294, no. 6, Jun 2006, pp. 80-87.
5. Lynce, I. and J. Ouaknine. "Sudoku as a SAT Problem." *AI&M*, 2006.
6. Chi, Eric C. and K. Lange. "Techniques for Solving Sudoku Puzzles." *arXiv preprint arXiv: 1203.2295*, May 2013. https://doi.org/10.48550/arXiv.1203.2295
7. Wei, Xuan. "Difficulty level classification of sudoku puzzles based on convolutional neural network." *Academic Journal of Computing & Information Science*, vol. 6, no. 11, Nov 2023, pp. 35-39. https://doi.org/10.25236/AJCIS.2023.061105
8. Pelánek, Radek. "Difficulty rating of sudoku puzzles: An overview and evaluation." *arXiv preprint arXiv:1403.7373*, 2014. https://doi.org/10.48550/arXiv.1403.7373
9. Hunt, Martin, *et al.* "Difficulty-Driven Sudoku Puzzle Generation." *UMAP Journal*, vol. 29, no. 3, 2008, pp. 343-362.
10. "Sudoku - online puzzle game." *Puzzle-Sudoku.com*, https://www.puzzle-sudoku.com/.

Appendix

github.com/maddy-higgins/Sudoku-Classification