



PATTERN COMPUTER[®]

PATTERN DISCOVERY
BEYOND PUBLISHED
RESULTS:

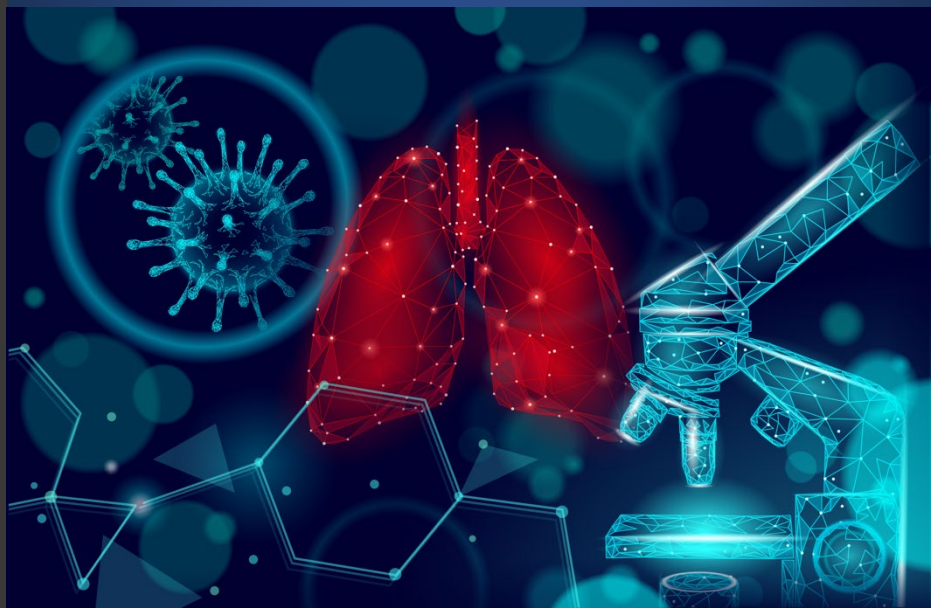
THREE PAPERS

Quinn Jackson, CSci FIScT

Pattern Computer Inc.

38 Yew Lane

Friday Harbor, WA 98250



Copyright © 2021 Pattern Computer Inc.

All Rights Reserved

References to specific products may be registered trademarks of their respective companies.

Abstract:

We set out to reproduce or improve upon the predictive accuracies of three recent published studies in the area of breast cancer tumor identification, fatal heart failure in cardiac patients, and behavioral determinants of cervical cancer. Our Pattern Discovery Engine™ was able to exceed the published, peer-reviewed accuracy results presented in these three papers and provide subsets of the covariates of the datasets that most informed its final predictive models. In addition, the Pattern Discovery Engine produced a human readable mathematical model that describes the relationships of the covariates to the final predictive model.

Study: Wisconsin Breast Cancer Dataset

This dataset, which we previously explored using Gilford Island, and which is discussed in a Pattern Computer whitepaper¹, comes to us from the [WisconsinBreastCancer] dataset, and is the subject of much prior study.² It consists of 30 continuous covariates mapped to a binary determination of the benign or malignant nature of the sample, across 569 observations, 212 of which map to malignant tumors and 357 of which map to benign tumors.

Though many papers have been published that examine this dataset against various machine-learning approaches, we shall compare our Pattern Discovery Engine's performance in this regard using the peer-reviewed results published in [Ak 2020]. The primary claim of interest put forth in [Ak 2020] is found in the abstract: "Results obtained with the logistic regression model with all features included showed the highest classification accuracy (98.1%), and the proposed approach revealed the enhancement in accuracy performances." This claim is then supported in the paper by the findings and analysis of those findings.

We undertook to reproduce these results with our Pattern Discovery Engine. The dataset, having already been prepared, was readily available to us and needed no further preparation. A script of twenty-five identical runs was used; in each of the runs, a training set of 80% of the total observations (selected at random) was used, and 20% were set aside as hold-out for validation.

¹ *q.v.* "The Power of Knowing Why," available at the URL:

<https://www.patterncomputer.com/publications/knowing-why/>

² *q.v.* [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

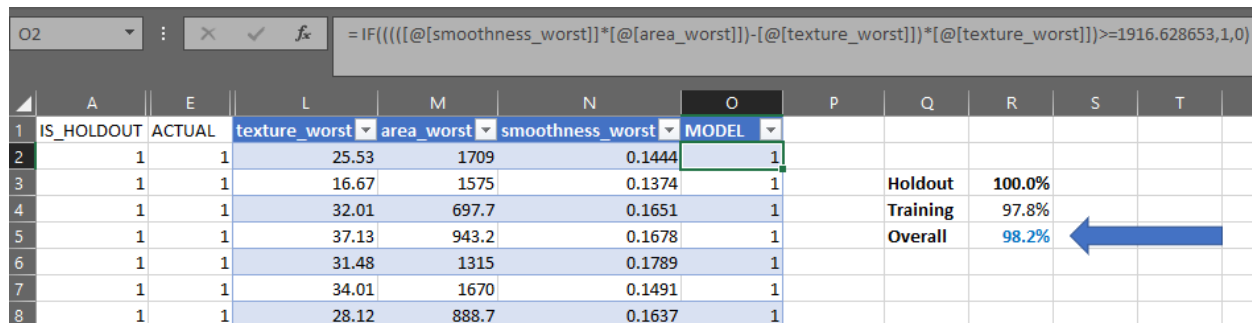
Of these runs, the best accuracy on held-out data was **100%**, with an accuracy of **97.8%** on training. Overall accuracy against the entire dataset was **98.2%**. We have exceeded the published performance obtained by [Ak 2020].

We know how our results came about, and which parameters and assumptions were made, and we have an exact mathematical equation in Microsoft® Excel® format to further validate our findings, as seen here:

$$= \text{IF}(((\text{smoothness_worst} * \text{area_worst}) - \text{texture_worst}) * \text{texture_worst}) \geq 1916.628653, 1, 0)$$

Figure 1: Wisconsin Breast Cancer dataset study, Excel formula

When entered into Excel® (using arrays) we can confirm its accuracy on the training, holdout, and overall set:



IS_HOLDOUT	ACTUAL	texture_worst	area_worst	smoothness_worst	MODEL
1	1	25.53	1709	0.1444	1
3	1	16.67	1575	0.1374	1
4	1	32.01	697.7	0.1651	1
5	1	37.13	943.2	0.1678	1
6	1	31.48	1315	0.1789	1
7	1	34.01	1670	0.1491	1
8	1	28.12	888.7	0.1637	1

Holdout	100.0%
Training	97.8%
Overall	98.2%

Figure 2: Wisconsin Breast Cancer dataset study, Excel formula

Study: Heart Failure Dataset

This dataset comes to us from [HeartFailureDataset], and results obtained from it have been peer-reviewed and put forth recently in [Chicco & Jurman 2020]. It is a small dataset, representing only 299 observations, each assigned 13 clinical attributes, against a binary outcome, with “1” signifying that the patient died before the next visit to a clinician (96 cases) and “0” signifying that the patient lived (203 cases). As such, it is an unbalanced dataset as regards the response variable. The primary claim made in [Chicco & Jurman 2020] is stated as: “We also carry out an analysis including the follow-up month of each patient: even in this case, serum creatinine and ejection fraction are the most predictive clinical features of the dataset and are sufficient to predict patients’ survival.”

A script of 10 identical runs was prepared, with 60% of the observations being selected at random for training, with the remaining 40% of the observations having been set aside as held-out validation data. Our first goal was to reproduce the results of the paper: that **serum_creatinine** and **ejection_fraction** alone were good overall predictors, as stated in the paper: “serum creatinine and ejection fraction are the most predictive clinical features of the dataset and are sufficient to predict patients’ survival.”

Our results across these 10 runs were as follows:

TEST RUN	TRAINING	HOLDOUT	TRAINING		HOLDOUT	
	ACCURACY	ACCURACY	TP_ACCURACY	TN_ACCURACY	TP_ACCURACY	TN_ACCURACY
1	77.70%	77.50%	71.90%	80.30%	69.20%	81.50%
2	76.00%	75.80%	73.40%	77.40%	68.80%	78.40%
3	77.10%	75.80%	68.00%	80.60%	67.40%	81.10%
4	79.30%	78.30%	71.90%	83.50%	65.60%	83.00%
5	74.30%	70.80%	71.90%	75.40%	69.20%	71.60%
6	78.20%	75.00%	75.00%	79.50%	65.90%	80.30%
7	71.50%	75.80%	66.70%	73.40%	82.20%	72.00%
8	77.70%	71.70%	77.40%	77.80%	79.10%	67.50%
9	71.50%	82.50%	66.10%	74.40%	82.40%	82.60%
10	72.60%	70.80%	70.00%	73.90%	88.90%	63.10%
Average	75.60%	75.40%	71.20%	77.60%	73.90%	76.10%

Figure 3: Heart failure study results – hypothesis exploration

In Figure 3 above, see that our average holdout accuracy on these runs was **75.4%**, with holdout true negative accuracy of **76.1%** and true positive accuracy of **73.9%**. We can see that the covariates **serum_creatinine** and **ejection_fraction** alone do indeed provide predictive features to predict patients’ survival. Against the held-out set, representing 40% of the observations, we achieved a predictive rate substantially better than chance; moreover, the predictive rates both in the patients who died before their next visit to a clinician and the patients who survived is fairly well balanced. We have more or less confirmed the published findings and exceeded the published accuracy measures in a matter of 10 runs that took a total of about 15 minutes to run.

But this is not the entire picture. We have exceeded the predictive results given in [Chicco & Jurman 2020], which reports a best average prediction³ using only those two covariates as being **58.5%**, and highly unbalanced true positive and true negative performance. However, the claim is worded “serum creatinine and ejection fraction are the most predictive clinical features of the dataset.” Are they truly the most predictive clinical features of the dataset in this regard? Up to this point, we do not know, since we have limited our runs by specifying the hypothesis of the published paper.

³ q.v. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5/tables/9>

We have not yet asked the Pattern Discovery Engine to make its own hypothesis from the dataset alone. We therefore prepared 10 more runs but did not specify a starting hypothesis; all covariates were available to our Engine for consideration.

This second set of runs yielded:

TEST	TRAINING	HOLDOUT	TRAINING		HOLDOUT		FINAL
RUN	ACCURACY	ACCURACY	TP_ACCURACY	TN_ACCURACY	TP_ACCURACY	TN_ACCURACY	ORDER
1	85.50%	83.30%	85.70%	85.40%	72.50%	88.80%	5
2	86.60%	79.20%	86.70%	86.60%	75.00%	81.00%	5
3	87.70%	84.20%	88.10%	87.50%	89.20%	81.90%	5
4	91.10%	72.50%	91.10%	91.10%	72.50%	72.50%	5
5	87.70%	81.70%	87.70%	87.70%	79.50%	82.70%	5
6	86.00%	83.30%	86.20%	86.00%	71.10%	89.00%	4
7	88.30%	81.70%	88.30%	88.20%	80.60%	82.10%	4
8	84.90%	83.30%	85.20%	84.70%	80.00%	84.70%	4
9	87.70%	79.20%	87.50%	87.80%	67.50%	85.00%	4
10	84.90%	85.80%	85.00%	84.90%	77.80%	89.30%	6
Average	87.00%	81.40%	87.20%	87.00%	76.60%	83.70%	5

Figure 4: Heart failure study results

This represents a substantial jump in overall predictive accuracy against held-out data. Our average accuracy on holdout is now **81.4%**, with **76.6%** true negative accuracy and **83.7%** true positive accuracy on held-out data. As noted in the **FINAL ORDER** column in Figure 4 above, the best predictor on holdout uses 6 covariates, shown below:

- **age**
- **serum_creatinine**
- **ejection_fraction**
- **smoking**
- **serum_sodium**
- **time**

We have demonstrated that when we take into account more than the two covariates proposed by the authors of the paper, we arrive at a significantly more performant model. We will not investigate those additional covariates further here, except to say that we arrived at this level of understanding of the dataset within 15 minutes of CPU time and the time it took to examine the results with the most ubiquitous of data analysis tools: SQL, Excel, and the human eye.

Study: Behavior Determinant Based Cervical Cancer

The final study we will consider comes to us from [Sobar *et al.* 2016]. This paper outlines a study of 72 participants, and the dataset represents 21 observations with cervical cancer and 51 controls with no cervical cancer diagnosis. The results reported, in summary, are stated as: “From the experimental result, both [Naïve Bayes] and [Logistic Regression] are promising as a classifier to detect [cervical cancer] risk based on behavior and its determinant with accuracy **91.67%** and **87.5%** respectively....” The authors support their claim with a 10-fold model cross-validation analysis.

Our study of this dataset occurred in two phases: initial exploration and final model production.

Phase 1: Initial Exploration

We first acquired and readied the [CervicalCancerDataset] used in the study, which took approximately 10 minutes to prepare. The dataset consists of 19 behavior determinants and one binary response variable signifying cancer “1” and no cancer “0.” Given that a 10-fold approach was used by the authors of the paper, the Pattern Discovery Engine was configured to execute 10 runs, each with 90% training and 10% holdout at random on the dataset.

Our results on this run are shown below:

TEST	TRAINING	HOLDOUT	TRAINING		HOLDOUT	
RUN	ACCURACY	ACCURACY	TP_ACCURACY	TN_ACCURACY	TP_ACCURACY	TN_ACCURACY
1	87.30%	100.00%	88.89%	86.67%	100.00%	100.00%
2	82.54%	88.89%	83.33%	82.22%	66.67%	100.00%
3	87.30%	100.00%	88.89%	86.67%	100.00%	100.00%
4	79.37%	88.89%	77.78%	80.00%	66.67%	100.00%
5	96.83%	88.89%	94.44%	97.78%	66.67%	100.00%
6	79.37%	100.00%	77.78%	80.00%	100.00%	100.00%
7	90.48%	88.89%	88.89%	91.11%	66.67%	100.00%
8	87.30%	88.89%	88.89%	86.67%	66.67%	100.00%
9	93.65%	88.89%	94.44%	93.33%	66.67%	100.00%
10	84.13%	77.78%	83.33%	84.44%	100.00%	66.67%
Average	86.83%	91.11%	86.67%	86.89%	80.00%	96.67%

Figure 5: Behavior determinant based cervical cancer study – first run results

Our initial results essentially reproduced those of the paper, with **91.11%** accuracy on holdout falling only slightly under the highest accuracy achieved by the authors. At this point, we switched approaches and continued further, to see if we might get even higher accuracy.

Phase 2: Final Model Generation

The 10-fold approach used in the paper and the 10-fold runs that we executed above all suffer from the same limitation: although k -fold validation is used in an attempt to generalize a model when presented a small sample size⁴, over the span of 10 such folds all observations in the dataset are ultimately seen by the model building phase. There is not truly held-out data over the span of the entire study.

To accommodate this, we approached the dataset differently for final model generation, asking this question: If only 48 observations had been originally available to us to train on, could we construct a model that gives good predictive accuracy on 24 subsequently supplied, entirely unseen observations? To begin to explore this question, we shuffled the entire 72 observation dataset with only one constraint: the training dataset (48 observations) had to have an equal proportion of positive cases as the test dataset (24 observations). This resulted in a training set of 14 positive cases and 48 observations total and a held-out test set of 7 positive cases and 24 observations total.

We trained on 100% of the 48 observations that had been set aside for training, across 20 runs. We then selected the two best models of that run, each having reached 100% accuracy on the training set.

Of those two best models, when then tested against the entirely held-out test dataset, the best yielded an accuracy of **95.83%** (surpassing the published result), and the next-best yielded an accuracy of **91.67%** (matching the published result).

The first of these two top performant models (95.83% accuracy on holdout, 97.22% accuracy overall) used the covariates **perception_severity**, **empowerment_desires**, **socialSupport_instrumental**, and **motivation_strength**. The second (91.67% accuracy on holdout, 94.44% accuracy overall) used the covariates **perception_severity**, **empowerment_desires**, and **motivation_strength**. In summary: we arrived at one final model that matched to the published paper's result on holdout, with the added distinction that we did so with only 3 algorithmically determined covariates out of 19 possible covariates, and another model that exceeded the published paper's results, all on training against only 48 of the total of 72 observations. Despite this small training set size, the overall accuracies of both predictors on the entire dataset surpassed the published paper's results.

⁴ See <https://machinelearningmastery.com/k-fold-cross-validation/> for more on this.

Again, we can easily verify these findings in Excel, using the formula supplied by the Pattern Discovery Engine, as seen below:

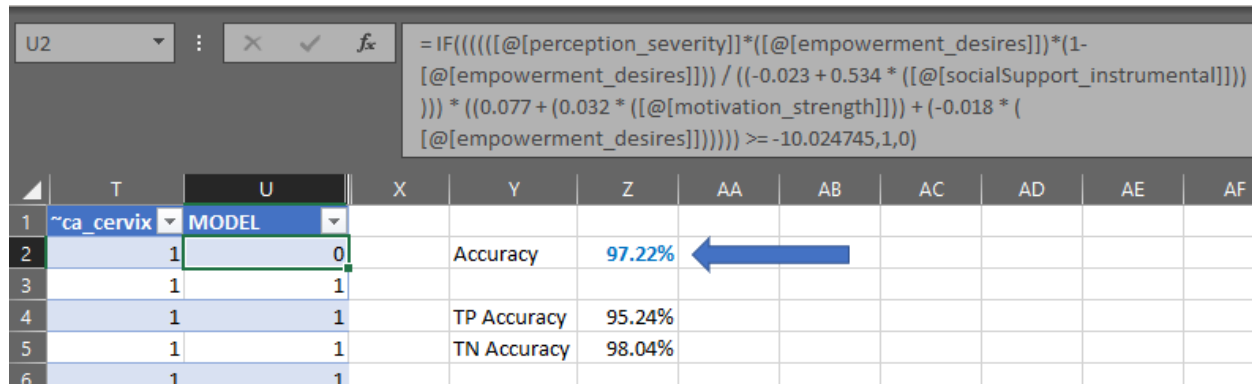


Figure 6: Behavior determinant based cervical cancer study – Excel formula

In Summary

Our goal was to reproduce or improve upon peer-reviewed published results on three publicly available datasets with our Pattern Discovery Engine, and in each case, we not only exceeded the accuracies of these published results, but we also derived models expressed as simple mathematical equations that allow us to further explore these datasets. In one case, we were able to do this by training on as few as 48 observations. One must not lose sight of the fact that on datasets of this kind, the percentage points of the predictors map to either cancer diagnoses or patient death: every percentage of accuracy has real-world significance. As such, understanding how and why a predictor in these domains predicts an outcome is of utmost interest. Our Pattern Discovery Engine not only provides the highest accuracy, but also provides a perspective to guide researchers, data scientists and subject matter experts throughout that investigation, by supplying interpretable mathematical models to support that inquiry.

References

[Ak 2020] Muhammet Fatih Ak, “A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications,” *Healthcare (Basel)*, Vol. 8 (2), June 2020, available at URL: <https://dx.doi.org/10.3390%2Fhealthcare8020111>

[CervicalCancerDataset] “Behavior Determinant Based Early Detection dataset,” available at URL: <https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk>

[Chicco & Jurman 2020] Davide Chicco & Giuseppe Jurman: “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Medical Informatics and Decision Making*, Vol. 20 (16), 2020, available at URL: <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5>

[HeartFailureDataset] “Heart failure clinical records dataset,” available at URL:

<https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>

[Sobar *et al.* 2016] Sobar *et al.*, “Behavior Determinant Based Early Detection and Machine Learning Algorithm,” *Advanced Science Letters*, Vol. 22, pp. 3120-3123, 1 October 2016, available at URL:

https://www.researchgate.net/publication/318009235_Behavior_determinant_based_cervical_cancer_early_detection_with_machine_learning_algorithm

[WisconsinBreastCancer] “Breast Cancer Wisconsin (Diagnostic) dataset,” available at URL:

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))