Data-Driven Decision Support for Autism Diagnosis using Machine Learning

SOTIRIS BATSAKIS*, Technical University of Crete, Greece
MARIOS ADAMOU, South West Yorkshire Partnership NHS Foundation Trust, UK
ILIAS TACHMAZIDIS, University of Huddersfield, UK
GRIGORIS ANTONIOU, University of Huddersfield, UK
THANASIS KEHAGIAS, Aristotle University, Greece

This paper describes work in progress about using AI technologies to support diagnostic decision making. In particular, we analyse clinical data of past cases to develop a data-driven prediction model for future cases. To do so, we use a versatile AutoML platform that applies a multitude of machine learning algorithms and their configurations. Our results show initial promise, but also point to limitations of currently available data, opening up avenues for further research.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; **Machine learning**.

Additional Key Words and Phrases: Autism, Clinical decision support

ACM Reference Format:
Sotiris Batsakis, Marios Adamou, Ilias Tachmazidis, Grigoris Antoniou, and Thanasis Kehagias. 2021. Data-Driven Decision Support for Autism Diagnosis using Machine Learning . In *International Conference on Management of Digital EcoSystems (MEDES '21), November 1–3, 2021, Virtual Event, Tunisia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3444757.3485101

1 INTRODUCTION

Autism spectrum disorder (ASD) is characterized by pervasive difficulties in reciprocal social cognition, alongside apparent strict repetitive behaviors and interests. Currently, no biomarker for diagnosing ASD exists. Because of this, the diagnostic process tends to be time consuming and costly for health services. The recommendation is that diagnosis of ASD in adulthood is reached on a consensus of expert opinion from observations by multidisciplinary teams, which include observations of current behaviours and cognitive abilities, alongside detailed history taking.

The process of diagnosing ASD in adulthood can be complex for a variety of reasons, which can lead to underdiagnosis and missed treatment opportunities. Ideally, information from a variety of sources is required, and if the contribution of information from a primary caregiver is not available, it may be difficult to build an accurate interpretation, as self-insight from patients themselves may be unreliable. Further, it requires a high level of specialisation by professionals, as ASD symptoms can overlap with other disorders.

With pressure on health services to deliver efficient and effective care for patients, employing screening measures can facilitate a timely and economical system for specialist services to identify

[*] This author is also affiliated with the University of Huddersfield, UK

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MEDES '21, November 1–3, 2021, Virtual Event, Tunisia

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8314-1/21/11…$15.00
https://doi.org/10.1145/3444757.3485101

those who are more likely to have the condition in question. Whilst a varied collection of ASD screening measures is available for both developmental and adulthood populations [7], for ASD in adulthood, the most generally used screening measures for ASD is the Autism Questionnaire presented in [2], which forms the basis of the analysis in this work. The objective of this work is to apply machine learning for analysing Autism Questionnaire results and investigating the components of the assessment, in relation to diagnostic outcome in a clinical setting. Analysis results in turn can offer insights for decision support for Autism diagnosis.

The remainder of this paper is organised as follows. The assessment data is presented in Section 2. The analysis procedure and results are presented in Section 3. Conclusions and directions of future work are presented in Section 4.

2 DATA DESCRIPTION

The dataset consists of autism assessment results for 192 patients, from Adult ADHD and Autism Service, South West Yorkshire Partnership NHS Foundation Trust, in the South and West Yorkshire geographical area, between 2017 and 2018. The Adult ADHD and Autism Service is a specialist Service in diagnosing ADHD and Autism in adulthood. Patients are referred to the Service by health care professionals, whom deem it appropriate based on patient's history and current difficulties. Inclusion criteria dictated that participants were over the age of 18 years (no cut off), had a good comprehension of the English language, and IQ within normal range. The assessment is designed to identify adults who may benefit from a full diagnostic assessment for autism spectrum disorder.

The assessment procedure adopts the procedure proposed in [2] and consists of two parts. The first part consists of a test that the examined individual completes based on AAA AQ and AAA EQ parts. The second part (AAA RQ score) is the result of answers of persons familiar with the examined individual, typically close relatives. Related to the diagnosis are social aspects, communication, imagination and obsessions of the examined individual (these are features CLASS SOCIAL, CLASS COMMUNICATION, CLASS IMAGINATION and CLASS OBSESSIONS, respectively) and they are defined from responses to AAA AQ, EQ and RQ and clinician's input. These parts of the AAA examination in turn are the Autism-Spectrum Quotient (AQ) score [3] and the Empathy Quotient (EQ) score [1], in addition to Relatives Quotient (RQ). Given the AAA AQ, AAA EQ and AAA RQ responses clinicians confirm answers (Yes=1), which count towards CLASS classification. Thus, CLASS classification is a function of AAA responses and clinician's assessment. The last feature of the dataset is the diagnostic outcome which is a binary categorical feature that the machine learning model has to predict. Overall the dataset is unbalanced with 28 of the examined patients out of 192 (14.58%) being diagnosed with autism after a full assessment is completed. Thus, in total the dataset consists of seven numerical input features (three consisting solely of questionnaire's results and four based on questionnaire's results and clinician's input) and an output categorical feature.

3 DATA ANALYSIS

The objective of data analysis is to create a model for predicting the diagnostic outcome given the AAA test data [2] as input. Specifically the input data are AAA test results consisting of AAA AQ, AAA EQ and AAA RQ scores. In addition the input data include the features CLASS SOCIAL, CLASS OBSESSIONS, CLASS COMMUNICATION and CLASS IMAGINATION derived from AAA test responses as defined in [2]. The dataset consists of exam results of 192 individuals, with 85.42% of diagnostic outcomes being negative. In this work, various classification methods have been used for the analysis.

Table 1. Classification Results using non interpetable algorithms of Weka

Model	Total Positive Rate	ROC Area
Multilayer Perceptron	**0.885**	0.805
SMO	0.854	0.500
Random Forest	0.859	**0.870**

3.1 Analysis using WEKA

The fist part of the analysis consisted of the application of six machine learning algorithms using Weka [6] over the dataset. Three of the algorithms are non interpretable and three are interpretable. The non-interpretable algorithms are Myltilayer Perceptron (the Neural Network implementation in Weka), SMO (Sequential Minimal Optimization algorithm for training a Support Vector Classifier) and Random Forest. The interpetable algorithms are the Decision Tree (J48), Logistic Regression and Semantic Artificial Neural Networks (SANN) [4]. SANN is a variant of Neural Networks with labeled hidden layer nodes which can be interpreted as logistic regression over each layer given the previous one. In all experiments, pre-processing has been applied by replacing missing values with the average value, while performance estimation and model selection was based on 10-fold cross validation.

The results of experiments using the non-interpetable classification algorithms of Weka and the default hyperparameters are presented in Table 1 (optimal values as marked in bold). Although Table 1 presents some basic results using the non-interpetable algorithms, the imbalance of the dataset and the relative importance of the different diagnostic outcomes and corresponding consequences makes the overall precision of algorithms one (but not the only) factor to take into account in the analysis. Thus, a detailed examination is required in order to assess the true usability of a data driven analysis in the decision process. Specifically, the cost of error varies given its type, typically it is a more serious error to predict a negative diagnostic outcome when it is actually positive (namely, a false negative) resulting in the patient not receiving the needed treatment, compared to predicting a positive diagnosis when in fact it is negative (namely, a false positive) with the cost being that of conducting a full assessment that eventually leads to a negative diagnosis. This observation in turn changes the use of a machine learning model in practice.

Typically, when each class is considered equally important and having similar costs for all types of errors a classifier selects the class having the higher probability. However, when classes have different importance and also different costs in case of classification errors, then the selection threshold of an algorithm must be adjusted accordingly. Data driven analysis may help making such policies more accurate and efficient. In practice, up to a certain degree, it is better to make an additional assessment of positive diagnosis to the patient rather than to select a negative diagnostic outcome (which could actually be positive).

After taking the above observations into account the detailed results for each algorithm are the following: SMO actually assigns all instances as having negative diagnostic outcome where the total positive rate is 0.854 (percentage of instances with negative diagnostic outcome) and the Receiver Operating Characteristic (ROC) curve (or Area Under the Curve - AUC) is 0.500 corresponding to a random classification, thus this model cannot be used in practice. Random forest achieved better results with total positive rate 0.859 and the ROC curve is 0.870. In this case, the classifier can be useful in practice. For example, given a policy that assigns much higher cost to a false negative error than to a false positive, the diagnostic outcome can be classified as positive even if the probability is low, in order to avoid false negative errors. Subsequently, if an assessment result is positive even if the probability of such outcome is according to classifier just 1% then all

Table 2. Classification Results using interpetable algorithms of Weka

Model	Total Positive Rate	ROC Area
Logistic Regression	0.844	0.814
Decision Tree (J48)	0.870	0.775
SANN	**0.875**	**0.870**

28 positive cases will be classified correctly and so are 47 of the negative ones, with the cost of having to provide full assessment in the 117 remaining negative cases. Thus the classifier can be used for making a decision for filtering out some cases, but also providing full assessment to all cases that have a positive diagnosis. By increasing the threshold to 2% the classification is correct for 26 out of the 28 positive cases and 69 out of 164 negative cases (still 95 negative cases will have full assessment). Thus reduction of false positives is combined with increase of false negatives and the relative cost of errors is used for defining the proper threshold and decision policy rather than the threshold value that maximizes classification accuracy, that is reported in Table 1. In case of Multilayer Perceptron (Neural Network) the total positive rate is 0.885 and the ROC curve is 0.805, thus offering the possibility of implementing a selection policy minimizing the cost of errors, but without creating an interpetable model.

Even though non-interpretable algorithms can assist in decision making by producing models that can predict the probability (given the results of an assessment) of a specific diagnostic outcome, thus facilitating the definition of a decision policy given the relative costs of errors, interpetability of the prediction model is often an important issue. Compliance to legal requirements and regulations means that specific rules have been taken into account when applying an AI-based system and this in turn means that the system's functionality is transparent and interpretable. A proposed approach is to employ interpretable machine learning algorithms, such as logistic regression and decision trees [5]. These algorithms are often efficient but not always as performing as non-interpretable ones, such as Support Vector Machines (SVM) and neural networks.

In the case of Neural Networks, using existing knowledge for building neural networks was first proposed in [8] and further developed in [9], introducing the Knowledge-Based Artificial Neural Networks (KBANN). These networks are constructed based on knowledge represented using logic rules, and in [4] a variant of KBANN called SANNs is proposed. SANNs are neural networks with labeled hidden layer nodes as KBANNs, but the construction of such neural networks is based on knowledge graphs rather than rules. In this work the interpetable algorithms applied to the autism assessment dataset are: Logistic regression, J48 decision tree and SANN. The SANN is constructed by introducing to the hidden layers nodes representing the AAA score (combining AAA AQ, AAA EQ and AAA RQ scores) and the CLASS score (combining the CLASS SOCIAL, CLASS OBSESSIONS, CLASS COMMUNICATION and CLASS IMAGINATION scores). The resulting network is presented in Figure 1.

The results using the interpetable algorithms of Weka are presented in Table 2 (optimal values as marked in bold). In medical diagnosis, interpreting the models is significant for decision making, thus we select to present the two categories of algorithms separately, since in case that interpretability is not an option but a strict requirement then only the corresponding algorithms can be used. Decision Tree (J48) achieved total positive rate of 0.870 and ROC curve of 0.775.

In the case of logistic regression the coefficients for predicting a negative diagnosis result are AAA AQ: 0.0381, AAA EQ: -0.0064, AAA RQ: -0.1282, CLASS SOCIAL: -0.585, CLASS OBSESSIONS: -0.2791, CLASS COMMUNICATION: -0.371, CLASS IMAGINATION: -0.6105 and Intercept: 7.344.

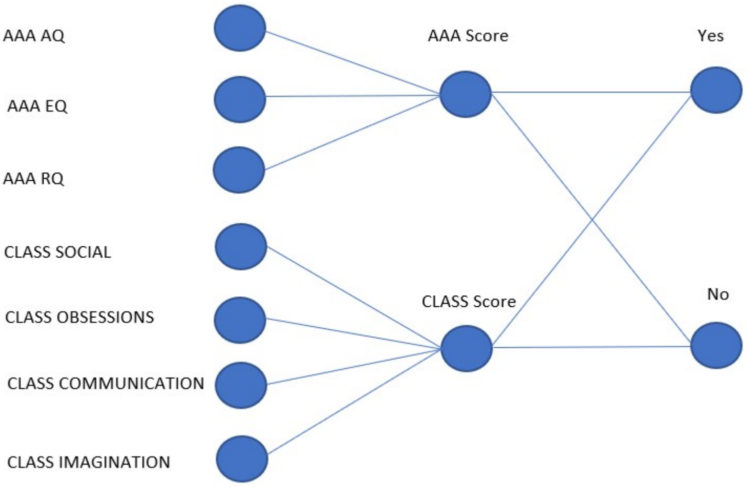

Fig. 1. SANN for classification on autism dataset.

These coefficients indicate factors correlated positively or negatively with negative diagnosis and the degree of this correlation (with CLASS features and AAA RQ having more weight).

The third algorithm, SANN, (using the network of Figure 1) achieved total positive rate 0.875 and ROC curve of 0.870 outperforming the other two interpretable algorithms. There are two hidden layer nodes in the SANN, the AAA Score node representing the cumulative AAA score and CLASS Score node representing cumulative CLASS score. The output node representing negative diagnostic output has weights 3.21 at input from the AAA Score Node and 4.84 at input from CLASS Score node, while the corresponding weights at positive diagnostic outcome node are -3.21 and -4.48, respectively. Thus, the positive diagnostic outcome has lower probability when cumulative AAA and CLASS scores are higher. The AAA Score in turn has weights 5.07 from AAA AQ input, -10.10 from AAA EQ and -12.39 from AAA RQ indicating that overall the higher the AAA AQ the lower the probability of a positive diagnosis and that lower AAA EQ and AAA RQ scores increase the probability of positive diagnostic outcome. Furthermore AAA EQ and AAA RQ scores have more weight than AAA AQ. The corresponding weights for the cumulative CLASS Score are for CLASS SOCIAL: -12.70, CLASS OBSESSIONS: -3.24, CLASS COMMUNICATION: -3.81 and CLASS IMAGINATION: -2.81 indicating that lower CLASS scores increase probability of positive diagnostic outcome.

Depending on the relative cost of classification errors, by setting a low threshold for accepting a positive diagnosis, the created model can be used to filter out cases which have a negative diagnostic outcome with very high probability. For example setting a threshold for classifying a case as positive to 1% then 26 out of 28 positive cases are classified correctly and so are 86 out of 164 negative cases (so a full assessment is applied for 78 negative cases). Thus, practically more than half of negative cases can be exempted from further examination while keeping almost all of positive cases. This is actually similar to the clinical assessment practice. For example in this dataset, out of the 192 cases, 28 are positive and 164 negative. In the screening process 125 cases went through full assessment and 67 did not. Out of these 125 cases, finally 26 were positive and 99 negative. Out of the 67 cases not further assessed, 65 were negative and 2 positive. Thus the policy adopted in clinical practice corresponds to that of applying a low threshold classifier, minimizing false negatives for the positive diagnosis class. Notice that, although SANN achieved high performance

Table 3. Area Under the Curve (AUC) results using JAD Bio

	Interpetability required		Interpetability not required	
	Feature Selection	No Feature Selection	Feature Selection	No Feature Selection
Preliminary	0.756	0.794	0.750	**0.833**
Typical	0.778	**0.807**	0.798	0.830
Extensive	**0.794**	0.806	**0.833**	0.823

and is interpretable, a disadvantage of this method is that the construction of network topology must be done manually, thus this algorithm is incompatible with a fully automated data analysis process.

3.2 Analysis using JAD Bio

Even though tools such as Weka can be used whether interpretability is required or not, when using a tool such as Weka there are two disadvantages; first the user must be familiar with machine learning which is not always the case in an environment such as the medical domain and second the analyst must apply various algorithms and also has to tune their hyperparamets in order to achieve optimal results. Overall this is a time consuming process, and in addition to this it is also uncertain, especially in case of a large search space for hyperparameter's values, with respect to the optimal selection of hyperparameters. This is the reason why systems automating machine learning are very important for wide scale adoption of machine learning for data analysis and decision support in the medical domain.

In this work in addition to the analysis done manually using Weka, the automated analysis tool called JAD Bio [10] has been used as well. By using JAD Bio, users simply upload their data and provide their preferences, subsequently the system selects the optimal model. In an application domain such as medical diagnosis where expertise on machine learning may not be available and a series of trials with many algorithms and their hyperparameters may not be an option due to limitations over resources such as time, the use of tools that automate machine learning tasks is expected to be widespread. JAD Bio allows for setting user preferences related to feature selection (optional or required), interpetability (optional or required) and time preference (preliminary, typical and extensive). Results using the above preferences are summarized in Table 3.

When using the JAD Bio system, in case that interpetability is not required, a Support Vector Machines (SVM) is the optimal model selected when combined with feature selection (and extensive time preference) and Classification Random Forests training 100 trees is the optimal algorithm when feature selection is not applied. In case the algorithm must be interpretable then Ridge Logistic Regression is the best performing algorithm when combined with feature selection (and extensive time preference) and without feature selection (and typical time preference). Feature selection, pre-processing and hyperparameter selection is performed automatically by the JAD Bio system.

Specifically, after examining various possible settings the JAD Bio system applied in preprocessing is constant removal and standardization. Then in feature selection the algorithm applied is Statistically Equivalent Signature (SES) algorithm with hyper-parameters: maxK = 2, and alpha = 0.1. JAD Bio selected 3 out of the total number of features in the original dataset: CLASS SOCIAL, AAA RQ and CLASS COMMUNICATION. Performance when using all features instead of only these three remained almost identical. The feature selection was applied by estimating the performance decrease when the feature was removed.

The best predictive model was Support Vector Machines (SVM) of type C-SVC with Polynomial Kernel and hyper-parameters: cost = 0.001, gamma = 10.0, degree = 3 having an Area Under the Curve

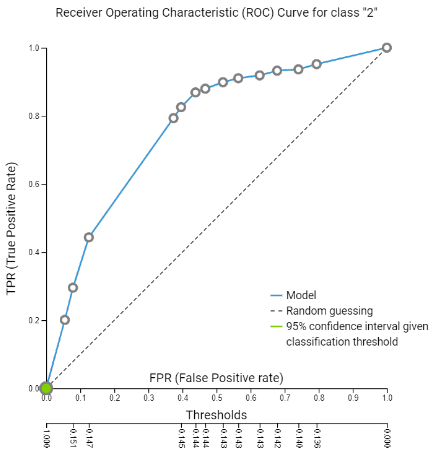

Fig. 2. ROC Curve of best performing model using JAD Bio.

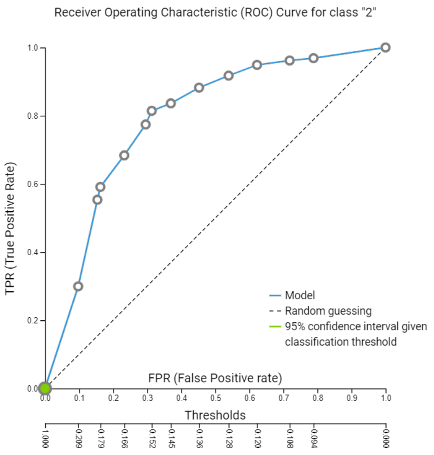

Fig. 3. ROC Curve of best performing interpretable model with feature selection using JAD Bio.

(AUC) of 0.833. Notice that the corresponding algorithm using Weka (SMO) has lower performance because of the different hyperparameter selection. The ROC curve of the best performing model using JAD Bio is presented in Figure 2. Using the diagram the user can specify the true positive rate for a specific class (in the case its class 2 indicating a positive diagnostic outcome) given the threshold selected.

The best interpretable model with feature selection was Ridge Logistic Regression with penalty hyper-parameter lambda = 100.0, with AUC (ROC) 0.794. The ROC curve for Ridge Logistic Regression is presented in Figure 3. Based on the curve, we can see that when setting the threshold to 9.4%, the true positive rate for the positive diagnostic outcome class is 0.969 and false negatives rate is 0.005. Taking into account the trade-off between false positive error rate and false negative error rate and the corresponding costs the optimal threshold can be defined for cost minimization.

Notice that JAD Bio adopts the bootstrap corrected cross validation performance estimation protocol presented in [11]. The objective of bootstrap corrected cross validation is to overcome the optimistic bias of cross validation, that is the typical method for performance estimation and model selection in machine learning (notice that 10-fold cross validation was used as performance metric in the experiments using Weka). The performance estimation is a task both difficult and critical, especially in medical applications were the reliability of the prediction model is a crucial parameter in decision making. This means that the performance metric of JAD Bio is less optimistic than this of Weka, but also this stricter performance evaluation is desirable in critical applications.

Overall the JAD Bio system produced models (including interpretable models) that offered high performance in addition to fully automating the analysis process which is a great advantage over traditional systems such as Weka. Although the dataset was not balanced and the two classes were difficult to separate (this is illustrated by the poor performance of SMO algorithm using Weka), by selecting carefully the threshold value of the classification model, after taking into account corresponding costs, the performed analysis can assist the decision making process. Notice also that depending on the cost estimation, a cost benefit analysis, when combined with an examination of the classification models, may lead to a decision to revise the assessment or even discontinue it in case there is no benefit of applying this assessment before the full assessment. This for example can be the case when the cost of making a false negative prediction regarding the diagnostic outcome is far greater than this of false positives.

4 CONCLUSIONS AND FUTURE WORK

This paper presented a data driven analysis over a dataset for autism assessment. Preliminary results showed that various algorithms achieved high performance although the diagnostic outcome classification was not an easy task because of the dataset characteristics (unbalanced, having some features that were not useful and not easily separable i.e. in a linear way). Furthermore, when applying such an analysis in practice, there are other crucial factors besides the total performance, such as the requirement of interpretability and automation of the analysis process, in addition to optimal performance for specific classes and the relative cost of various types of errors when specifying the decision process.

Future work will proceed in various directions. A particular direction will be to consider richer clinical data; there are even ideas to capture neurological data and/or facial expressions through video. Another interesting idea is to expand the AI technologies used by capturing and representing explicitly, through declarative rules, medical knowledge about how clinical data should be interpreted. Such a knowledge model could be used in conjunction with a machine learning model as discussed in this paper, thus deploying a hybrid AI approach.

REFERENCES

[1] Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders* 34, 2 (2004), 163–175.

[2] Simon Baron-Cohen, Sally Wheelwright, Janine Robinson, and Marc Woodbury-Smith. 2005. The adult Asperger assessment (AAA): a diagnostic method. *Journal of autism and developmental disorders* 35, 6 (2005), 807.

[3] Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. 2001. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders* 31, 1 (2001), 5–17.

[4] Sotirios Batsakis, Ilias Tachmazidis, George Baryannis, and Grigoris Antoniou. 2020. Semantic Artificial Neural Networks. In *European Semantic Web Conference*. Springer, 39–44.

[5] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[7] Tanja Sappok, Manuel Heinrich, and Lisa Underwood. 2015. Screening tools for autism spectrum disorders. *Advances in Autism* (2015).

[8] Jude W Shavlik and Geoffrey G Towell. 1991. An approach to combining explanation-based and neural learning algorithms. In *Applications Of Learning And Planning Methods*. World Scientific, 71–98.

[9] Geoffrey G Towell and Jude W Shavlik. 1994. Knowledge-based artificial neural networks. *Artificial intelligence* 70, 1-2 (1994), 119–165.

[10] Ioannis Tsamardinos, Paulos Charonyktakis, Kleanthi Lakiotaki, Giorgos Borboudakis, Jean Claude Zenklusen, Hartmut Juhl, Ekaterini Chatzaki, and Vincenzo Lagani. 2020. Just add data: Automated predictive modeling and biosignature discovery. *BioRxiv* (2020).

[11] Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. 2018. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning* 107, 12 (2018), 1895–1922.