scientific reports

Check for updates

OPEN

On effectively predicting autism spectrum disorder therapy using an ensemble of classifiers

Bhekisipho Twala[1]✉ & Eamon Molloy[2]

An ensemble of classifiers combines several single classifiers to deliver a final prediction or classification decision. An increasingly provoking question is whether such an ensemble can outperform the single best classifier. If so, what form of ensemble learning system (also known as multiple classifier learning systems) yields the most significant benefits in the size or diversity of the ensemble? In this paper, the ability of ensemble learning to predict and identify factors that influence or contribute to autism spectrum disorder therapy (ASDT) for intervention purposes is investigated. Given that most interventions are typically short-term in nature, henceforth, developing a robotic system that will provide the best outcome and measurement of ASDT therapy has never been so critical. In this paper, the performance of five single classifiers against several multiple classifier learning systems in exploring and predicting ASDT is investigated using a dataset of behavioural data and robot-enhanced therapy against standard human treatment based on 3000 sessions and 300 h, recorded from 61 autistic children. Experimental results show statistically significant differences in performance among the single classifiers for ASDT prediction with decision trees as the more accurate classifier. The results further show multiple classifier learning systems (MCLS) achieving better performance for ASDT prediction (especially those ensembles with three core classifiers). Additionally, the results show bagging and boosting ensemble learning as robust when predicting ASDT with multi-stage design as the most dominant architecture. It also appears that eye contact and social interaction are the most critical contributing factors to the ASDT problem among children.

World autism awareness month is celebrated in April worldwide by people and organisations alike. This entire month is dedicated to raising awareness, sharing understanding, and shedding light on a global and South African health crisis, which parents have been battling for years. Autism Spectrum Disorder (ASD) can be defined as a developmental disability that affects social interaction, communication and learning skills (where the spectrum reflects a wide range of symptoms that the child can present). Autism therapies are interventions that attempt to lessen the deficits and problem behaviours associated with ASD.

About 1 in 160 children worldwide are diagnosed with ASD, with a higher rate of 1 in 68 children in the United States (Centers for Disease Control[1]). Abnormalities characterise themselves in social interactions and patterns of communication and a restricted, stereotyped, repetitive repertoire of interests and activities (WHO[2]). However, the treatment and intervention services for ASD are tricky since there are time-consuming treatments conducted. The ASD symptoms typically appear in the first 2 years of a child's life, developing in a specific period. There are many types of treatments available including behavioural, developmental, educational, social-relational, pharmacological and psychological. One South African group, Quest School (Sowetan LIVE[3]) has proposed that children be diagnosed when they are young, not at 5 years of age as is the norm.

We cannot overemphasise the impact ASD has on adults and children. The common signs in adults include finding it hard to understand what others are thinking or feeling and getting very anxious about social situations. For young children it includes not responding to their name, avoiding eye contact, and repetitive movements such as flapping their hands and flicking their fingers or rocking their bodies. Signs of ASD for older children include not seeming to understand what others are thinking or feeling, unusual speech, such as repeating phrases and talking "at" others or finding it hard to say how they feel. ASD can sometimes be different in girls (women) and boys (mean) with it being harder to spot especially in girls (women). Women may have learnt to hide signs of

[1]Office of the Deputy Vice-Chancellor (Digital Transformation), Tshwane University of Technology, Private Bag x680, Pretoria 001, South Africa. [2]Waterford Institute of Technology, School of Science & Computing, Waterford, Ireland. ✉email: twalab@tut.ac.za

ASD to "fit in" by coping with people who do not have ASD while girls may hide some signs of ASD by copying how other children behave and play[4].

Recently, there has been an explosion in ASD cases worldwide, and they have been increasing at an alarming rate (Centres for Disease and Control[1]). The World Health Organisation (WHO[5]; Wolff and Piven[6,7] have argued that 1 out of every 160 children has ASD. A certain percentage of people with this disorder have been shown to live independently, whilst others would require life-long care and support. ASD traits are difficult to trace due to tests and diagnoses requiring significant time and cost. However, several treatments, therapies and interventions can help children with ASD improve their abilities and reduce their symptoms.

The topic of ASD therapy enhancement has been of interest to researchers for decades as the effects of a robot-enhanced intervention for children with ASD. There has been a lot of research work in the spheres of machine learning (ML) and statistical pattern recognition (SPC), where communities were discussing how to combine models or model predictions[8]. Furthermore, much research work in these communities has shown that an ensemble learning of classifiers is an effective technique for improving predictive accuracy (due to its variance reduction benefit). Ensemble learning of classifiers' development and successful fielding have significantly lagged behind bio-medical and health science research activities, yet it has been prominent in other fields. A central concern of these applications is the need to increase the predictive accuracy of early ASD diagnosis and test decisions.

The basic idea behind ensemble learning is to train multiple classifier learning systems to achieve the same objective and then combine their predictions. There are different ways ensembles can be developed and the resulting output combined to classify new instances. The popular approaches to creating ensembles include changing the cases used for training through techniques such as bagging[9], boosting[10], stacking[11], changing the features used in training[12], and introducing randomness in the classifier itself[13].

Due to the nature of ASD and its impact on societies, an improvement in predictive therapy enhancement accuracy or even a fraction of a per cent translates into significant future savings in time, costs, and even deaths[14–16]. Furthermore, the economic effect of ASD on individuals with the disorder, their families, and society as a whole has been poorly understood and has not been updated in light of recent ASD prediction and detection findings. This enormous effect on families warrants better therapeutic, prediction and detection methods from machine learning and statistical pattern recognition communities.

Supporting the development of a child with ASD is a multi-profile therapeutic work on disturbed areas, especially understanding and linguistic expression used in social communication, development of mutual social contacts and functional or symbolic play. In recent years, robot learning[17] and robot-assisted ASD therapy (RET) have grown in popularity. The key research findings on RET have shown its effectiveness for children with ASD in particular: communication (common attention, imitation, undertaking communication behaviours, recognizing and understanding emotions and developing sensitivity to physical contact[18–22], Chernyak et al.[23]).

Other tools that are being embraced by therapists, counsellors, teachers, parents and their children to help those with ASD to better communicate and connect with others are virtual reality (VR) and augmented reality (AR). Several research studies examined suggest promising findings about the effectiveness of virtual and augmented reality-based treatments for the promotion, support, and protection of health and well-being in children and adolescents with ASD. VR and AR have also been used to help those without ASD understand what living with the condition means[24–26].

Using a variety of machine learning techniques, one could analyze the parent's age, socio-economic status and medications to predict a child's ASD diagnosis. Predictive algorithms could also be useful for identifying factors that may contribute to ASD. For example, machine learning algorithms helped find an association between ASD and a parent's use of substances such as caffeine and certain antidepressants[27]. Machine learning has also been used to better understand (or classify) why ASD traits vary in their nature and severity from person to person[28]. This was after another machine learning study by Stevens et al.[29] analyzed behavioural data and found two overarching behavioural profiles of ASD, each with its subgroups based on the severity of different traits. Other scholars have investigated the clinical applications of robots in the diagnosis and treatment of ASD (Diehl et al.[30]) while others have created machine-learning algorithms that could help robots understand when an autistic child needs help[31,32].

Several other studies for predicting ASD traits in an individual have been carried out by the ASD research and data (science) analytics community using several machine learning (ML) and statistical modelling techniques. These include screening detection, identification, classification and prediction of ASD traits in an individual.

For screening detection, alternating and functional decision trees[33–35], support vector machines[36] and "red flags"[37] have been used while support vector machines have been used for both detection and identification[38]. Kosmicki et al.[33] investigated logistic model trees and logistic regression for detecting non-ASD against ASD among children.

To predict ASD traits, a support vector machine, a naïve Bayes classifier, and the random forest have been further applied by[39]. Prediction of ASDT response from baselines fMRI using random forests and tree bagging was proposed by Dvornek et al.[40] with their learning pipeline method achieving higher accuracy compared with standard methods. Bala et al.[41] investigated the identification of ASD among toddlers, children, adolescents and adults using several machine learning algorithms (K-Star, classification and regression trees, *k*-nearest neighbour, support vector machine, bagging and random tree). SVM achieved the best performance for the prediction of ASD at different age levels.

Deep learning and neural networks have been used by Heinsfeld et al.[38] to predict ASD patients using imaging of the brain with a follow-up research work on classification and hemodynamic fluctuations by Xu et al.[42]. An empirical comparison of Adaboost, flexible discriminant analysis (FDA), decision tree (C5.0), boosted generalised linear model (GLMboost), linear discriminant analysis (LDA), mixture discriminant analysis (MDA), penalised discriminant analysis (PDA), support vector machines (SVM) and classification and regression trees

(CART) for early stage detection of ASD for toddlers, children and adults. Good performances were observed for SVM (toddlers), Adaboost (children) and GLMboost (adults)[43].

Recently, Kanchana et al.[44] predicted early phases of ASD in adults using naïve Bayes, logistic regression, random forest and random tree with random forest achieving the highest predictive accuracy.

From the review, it is evident that all the researchers have used single classifiers for detecting, predicting or classifying ASD in general yet ensemble models have been more stable and, more importantly, shown to predict better than single classifiers. They have also been known to reduce model bias and variance. Some studies have assessed the ethical and social implications of translating embodied artificial intelligence (AI) applications into mental health care across the fields of Psychiatry, Psychology and Psychotherapy[45,46]. Furthermore, despite the limitations of single classifiers which include them not being able to make predictions on new data that they have not seen before because all the historical information must be provided in advance and the overfitting problem (i.e. overfitting the data they are looking at).

None of the research studies have looked at predicting ASD therapy (ASDT) and ways to improve ASDT predictive accuracy. Yet, proactive corrective actions can be taken well in advance if the prediction of ASDT is even more accurate. A slight increase in therapy predictive accuracy will not only have a positive impact on foreseeing ASD in toddlers but also improve outcomes for children with ASD thereby reducing symptoms due to early behavioural interventions.

This research work proposes modelling using an ensemble of classifiers approach to help show their effectiveness when predicting ASDT in terms of a robot-assisted intervention group (Robot-Enhanced-Therapy) and a control group receiving intervention by humans only (Standard-Human-Treatment) conditions, respectively. The investigation aims to find out if the use of artificial intelligence algorithms can help identify a reliable method for identifying ASD children most likely to benefit from a specific intervention program in advance and a solid foundation for establishing a personalised intervention program recommendation system for ASD children.

Such an ensemble learning approach is used to overcome precariousness in predictions and to enhance the accuracy and efficiency of the predictions. The MCL systems architecture and resampling processes are also considered. In other words, this research work focuses on predicting the effectiveness of ASD-enhanced therapy using a social robot and a human being.

To this end:

1. The first significant contribution of the paper is the investigation of five single-classifier learning systems to identify the best performing in terms of predicting therapy enhancement for autistic children using a social robot the type of therapy (on the one hand) and standard human treatment (on the other hand).
2. The second contribution is the proposal of a multiple-classifier learning system (or ensemble learning) approach to predict ASD therapy enhancement. The idea is to assess if using such a multiple classifier learning systems (MCLS) approach will be worthwhile to overcome the limitations of a single classifier learning system (SCLS) in terms of predictive accuracy due to their inability to handle more complex tracking situations with high accuracies. To analyse the performance of MCLS over SCLS, the unique models must be accurate individually and they need to be sufficiently diverse. For this reason, all possible combinations of the number of classifiers per ensemble are explored (i.e. from two classifiers per ensemble to five classifiers).
3. Finally, feature selection through a decision tree-based approach is used to identify which physical characteristics are most significant in ASDT treatment.

To the best of our knowledge, this study is one of the few if not the first study that investigates the application of artificial intelligence in ASDT.

The rest of the paper is organised as follows: Sect. "Single-classifier learning systems" gives a background on single-classifier learning systems used for ASDT prediction. Then multi-classifier learning systems are examined from the intelligibility viewpoint to improve the effectiveness of ASDT predictive accuracy (Sect. "Multiple classifier learning systems"). Section "Experiments" presents the experimental design in set-up and results drawn from a DREAM dataset, supporting a data-driven study of ASD and robot-enhanced therapy. Finally, the paper is concluded with critical research findings and remarks in Sect. "Remarks and conclusion".

Single-classifier learning systems

There are several approaches to single-classifier learning. However, only five base methods of classifier construction are considered in this paper (i.e. a mixture of regression and tree-based, nets, instance-based and Bayesian-related). These include artificial neural network (ANN), decision tree (DT), *k*-nearest neighbour (*k*-NN), logistic discrimination (LgD) and the Naïve Bayes classifier (NBC). A brief description of the five classifiers and their use for classification or prediction tasks is now given.

Logistic discrimination

Logistic discrimination (LgD) is a supervised learning classification algorithm used to predict the probability of the target variable (for example, a class). It was initially developed by Cox[47] and later modified by Day and Kerridge[48]. LgD is related to logistical regression due to the dependent variable being dichotomous. In other words, only two possible values can be taken (for example, either 0 for non-detection of ASD or 1 for detecting ASD). For LgD, the probability density functions for the classes are not modelled like most supervised Learning Classification Methods but rather the ratios between them (i.e. it is partially parametric).

An unknown instance is a new element classified using a cut-off point score where the error rate is lowest for the cut-off point = 0.5[49]. The slope of the cumulative logistic probability function is steepest $\pi_i = 0.5$ $\pi_i \geq 0.5$ $\pi_i < 0.5$ at the halfway point [i.e. the logit function transforms continuous values to the range (0, 1)], which is

necessary since probabilities must be between 0 and 1. The LgD approach can be generalised to more than two classes (also called multinomial logit models). Multinomial Logit Models (MLMs) are derived similarly to the LgD models. For more details about MLMs and other modified versins of LgD, the interested reader is referred to Jolliffe[50] and, Hosmer and Lameshow[51] referred to the interested reader.

k-nearest neighbour

The *k*-nearest neighbour (*k*-NN) or instance-based learning approach is one of the most venerable and easy-to-implement machine learning algorithms for supervised and sometimes unsupervised learning[52]. *k*-NN can solve both classification and regression (prediction) problems by assuming that similar things exist near each other. Thus, the *k*-NN hinges on this assumption being true enough for the algorithm to be valid. Essentially, *k*-NN works by assigning the classification (or regression prediction) of the nearest set of previously classified (predicted) occurrences to an unknown instance. The memory is the storage for the entire training set.

To classify a new example, a distance measure (such as Hamming, Cosine similarity, Chebychev, Euclidean, Manhattan or Minkowski) is computed between the trained and unknown instances. For this paper, the Cosine similarity distance measure is used. Each stored training and the unknown instance are assigned the class of that nearest neighbouring instance. These Nearest Neighbours are first computed, and then the new example is given the most frequent class among the *k* neighbours. To select the value of *k* that is right for your data, the algorithm is run several times with different values of *k*. It reduces the number of errors encountered while maintaining the algorithms' ability to make predictions when given data not seen before accurately. For the paper, the process of supervised learning will be focused on.

Artificial neural networks

Like most state-of-the-art classification methods, Neural Networks or artificial neural networks (ANNs) are non-parametric (i.e. no assumptions about the data are made, as is the case with models such as linear regression). Instead, they are computational models inspired by an animal's nervous system. These are represented by connections (layers) between many simple computing processors or elements ("neurons"). They have been used for various classification and regression problems in economics, forensics, and pattern recognition[53]. The ANN is trained by supplying it with many numerical observations or the patterns to be trained (input data pattern) whose corresponding classifications (desired output) are known. The final sum-of-squares error (SSE) over the validation data for the network is calculated when training the network. This SSE value is then used to select the optimum number of hidden nodes resulting in a trained neural network.

A new unknown instance is carried out by sending its attribute values to the network's input nodes, where weights are applied to those values. Finally, the values of the output unit activations are computed. The weights and biases can be optimised by running the network multiple times. Its most significant output unit activation determines the classification of the new instance.

Decision trees

A Decision Tree (DT) classifier is a supervised machine learning algorithm used for regression and classification tasks. It starts with a single node (subsequently, a series of decisions) and branches into possible outcomes, giving it a tree-like diagram[54,55]. When training a DT, the best attribute is selected (using the information gain measure) from the total attributes list of the data for the root node, internal node and leaf or terminal nodes. A DT classifier is simple to understand, interpret and visualise.

According to Safavian and Landgrebe[56], a DT classifier has four primary objectives. These are (1) Classifying the training sample correctly as much as possible, (2) Generalising beyond the training sample so that unseen samples could be classified with high accuracy; (3) quickly updating the DT as more training samples become available (which is similar to incremental learning), and (4) Having a simple DT structure as possible. Despite the DT classifier strengths, Objective (1) is highly debatable and, to some extent, conflicts with Objective (2). Also, not all DT classifiers are concerned with objective (3). DTs are non-parametric and valuable to represent the logic embodied in software routines.

A DT takes as input a case or example described by a set of attribute values and outputs a Boolean multi-valued "decision," making it easy to build automated predictive models. For this paper, the Boolean case is considered. Classifying an unknown instance is easy once the tree has been constructed. Starting from the root node of the DT and applying certain test conditions would eventually lead you to a leaf node with a class label associated with it. The class label associated with the leaf or terminal node is assigned to the instance.

Naïve Bayes classifier

The Naïve Bayes classifier (NBC) is perhaps the most superficial and widely studied supervised probabilistic machine learning (ML) method that uses Bayes' theorem with strong independence assumptions between the features to procure results. The NBC assumes that each input attribute variable is independent of training the data. This can be considered a naïve assumption about real-world data. Then, the conditional probability of each attribute A_i, given the class label C is learnt from the training data[57,58].

The strength of the NBC lies in its ability to handle an arbitrary number of independent numerical or categorical attribute features. The solid but often controversial primary assumption (due to its "naivety") is that all the attributes A_i are independent given the value of class C. For classification, the Bayes rule is applied to determine the class of the unknown instances by computing the probability of C given $A_1, \ldots, A_n$ and then selecting the class with the highest posterior probability. The "naive" assumption of conditional independence of a collection of random variables is very important for the above result. Otherwise, it would be impossible to estimate all the parameters without such an assumption. This is a relatively strong assumption that is often not applicable.

However, any bias in estimating probabilities may not make a difference in practice – it is the order of the probabilities and not the exact values that determine the probabilities.

Nonetheless, NBC has been shown to solve many complex real-world problems and to do so effectively. Also, it requires a small amount of training data to estimate the parameter. A frequency table is created for each attribute against the target (class) to calculate the posterior probability of classifying an unknown instance. Then, the NBC is used to calculate the posterior distribution. Once again, the prediction outcome is the class with the highest posterior probability.

Multiple classifier learning systems

A multiple-classifier learning system (MCLS) can be defined as a set of classifiers whose individual predictions are combined in some way to classify new examples to produce one optimal predictive model. The most common type of MCLS includes an ensemble of classifiers that function for a parallel classifier input combination. Furthermore, a significant number of methods have been used to create and combine such individual classifiers, including ensemble methods, committee, classifier fusion, combination, aggregation, etc.

Once an MCLS is built and an aggregation determined, one has to design the MCLS architecture. There are three types of MCLS architectures, namely—static parallel (SP); multi-stage (MS) design; and three dynamic classifier selection (DCS)[59–61].

One of the most famous MCLS architectures is Static Parallel by Zhu et al.[62]. For SP, two or more classifiers are developed independently and executed in parallel. The outputs generated by all base classifiers are then combined to determine a final classification decision (selected from a set of possible class labels). Many combination functions are available for this architecture, including majority voting, weighted majority voting, the product or sum of model outputs, the minimum rule, the maximum rule and Bayesian methods. Averaging is mainly used for regression problems, while voting is used for classification problems. There are two categories of SP-related MCLS: a single ML algorithm is used as base learning (homogenous parallel), and multiple ML algorithms are used as base learning (heterogeneous parallel). For the paper, the former category has been used.

The second type of MCLS architecture is MS design, where the classifiers (usually with no overlaps) are organised into multiple groups and then iteratively constructed in stages. At each iteration, the parameter estimation process depends on the classification properties of the classifiers from previous stages. As with SP, this design benefits from processing inputs in parallel. It ensures that labels are assigned using only the necessary features. In addition, the number and composition of stages used by the model have proven to have a significant impact on overall performance. Some MS approaches have been used to generate models applied in parallel using the same combination rules used for SP methods. For example, most boosting strategies have been shown to create weak classifiers, but they tend to form stronger ones[63].

A dynamic classifier selection (DCS) is an ensemble learning architecture developed and applied to different regions within the problem domain. The technique involves training MCLS on the dataset and selecting the best prediction models. The *k*-NN approach is sometimes used to determine instances that are closely related to the unknown instance to be predicted (see Sect. "Single-classifier learning systems"). While one classifier may be shown to outperform all others based on global performance measures, it may not necessarily dominate all other classifiers entirely. Weaker competitors will sometimes beat the best across some regions (Kittler[64]). Research has shown DCS performs better than single classifiers and even better than combining all the base classifiers. Furthermore, Kuncheva[65] approached DCS problems from a global and local accuracy perspective with promising results.

Ensemble learning of classifiers can be classified into three stages: (1) generation, (2) selection, and (3) integration. The objective of the first stage is to obtain a pool of models, followed by a selection of a single classifier or a subset of the best classifiers. Finally, the base models are combined to obtain the prediction for new or unknown instances. The aspect of multiple classifier systems is determining the number of component classifiers in the final ensemble (also known as ensemble size or cardinality) is the most important. The impact of ensemble size on efficiency in time and memory and predictive performance makes its determination a critical problem[65–68], Li et al.[69]).

Furthermore, one should assume that diversity among component classifiers should be another influential factor in an accurate ensemble. However, no explanatory theory reveals how and why diversity among components contributes to overall ensemble accuracy. Therefore, all possible ensemble sizes and their respective diversities are considered for this paper.

Recently, Multi-Classifier-Based Boosting was introduced, where clustering and classifier training are performed jointly[70,71]. These methods have been applied to object detection, where the entire training set is available from the beginning. Other related works include multiple instance learning[72,73] and multiple deep learning architectures[74]. The former algorithm learns with bags of examples, which only need to contain at least one positive example in the positive case. Thus, the training data does not have to be aligned. Mellema et al.[74] developed the system using anatomical and functional features to diagnose a subject as autistic or healthy.

Ensemble methods offer several advantages over single models, such as improved accuracy and performance, especially for complex and noisy problems. They can also reduce the risk of overfitting and underfitting by balancing the trade-off between bias and variance, and by using different subsets and features of the data. Furthermore, they can provide more confidence and reliability by measuring the diversity and agreement of the base models, and by providing confidence intervals and error estimates for the predictions. Despite their pros, ensemble methods have some drawbacks and challenges such as being computationally expensive and time-consuming due to the need for training and storing multiple models, and combining their outputs. Additionally, they can be difficult to interpret and explain, as they involve multiple layers of abstraction and aggregation, which can obscure the logic and reasoning behind the predictions.

Experiments

Experimental set-up

The main aim of this randomized controlled experiment is to evaluate the effectiveness of five machine learning algorithms for predicting ASDT for a robot-assisted intervention group of autistic children and control receiving intervention by a human only. We further investigate how ASDT predictive accuracy could be improved using ensemble learning.

The investigations are conducted using a dataset of behavioural data and robot-enhanced therapy recorded from 61 children diagnosed with ASD[75]. The dataset covers 3000 therapy sessions and more than 300 h of treatment. Half of the children interacted with the social robot supervised by a therapist, while the other half was used as a control group (i.e. interacting directly with the therapist). In other words, the class attribute is the type of intervention – social robot-enhanced therapy condition (i.e. the interaction of an autistic child with a social robot) or social human therapy condition (i.e. the interaction of an autistic child with a human). The attributes are as follows:

(1) The ability of the child to wait for his or her turn;
(2) Social interaction and communication outcomes (engagement, eye contact, and verbal utterances);
(3) Behavioural outcomes (stereotype behaviours, maladaptive behaviours, and adaptive behaviours); and
(4) Emotional outcomes (functional and dysfunctional negative emotions, and positive emotions).

Furthermore, both groups followed the applied behaviour analysis (ABA) protocol. ABA uses scientific observations and principles of behaviour to improve and change behaviours of social interest[76]. In both the RET and SHT sessions, the children participated in a randomised manner to avoid ordering effects[32]. Participants in both groups went through a protocol of initial assessment, eight interventions, and a final assessment. The effect of the treatment was assessed using the Autism Diagnostic Observation Schedule (ADOS), in terms of the difference between the initial and final assessments[77].

All therapy sessions were recorded using the same sensorized therapy (Fang et al.[78]). Each session was recorded with three red–green–blue (RGB) cameras and two Red–Green–Blue-Depth (RGBD) Kinect cameras, providing detailed information on children's behaviour during therapy; the dataset comprises body motion, head position and orientation, and eye gaze variables, all specified as 3D data in a joint frame of reference. Other metadata attributes include participant age, participant gender numeric ID, target ability or task, therapy condition (response elaboration training or substitutive hormonal therapy) and date of therapy. A complete list of sensor primitives and associated methods is provided in Table 1.

This public release of the dataset does not include any footage of children. Instead, processed features of the recorded data are provided. According to the source of the data, informed consent was obtained from all subjects and/or their legal guardian(s) when the data was collected. The experimental protocols were approved by the University of Skövde in Sweden which is the main source of the data. For more information, the reader can contact Billing et al.[75].

In addition, metadata including participant's ID, age, gender, and ASD diagnosis variables (3D skeleton comprising joint positions of the upper body; 3D head position and orientation; 3D eye gaze vectors; therapy condition; therapy task including joint attention, imitation and turn-taking; data and time of recording and initial ADOS scores) are included. As this was secondary data, no ethics committee had to approve the study within our environment. Furthermore, all methods were carried out following relevant guidelines and regulations.

For the simulations, five base classifiers were modelled using default hyper-parameters for each respective classifier. Each approach utilises a different form of parametric estimation or learning. For example, they generate various forms in linear models, density estimation, trees, and networks. These classifiers are among the top 10 most influential and popular algorithms in data mining[87]. They are all practically applicable to ASD, with known examples of their application within the robotics-enhanced therapy industry.

First, each state-of-the-art classification method (base classifier) was constructed using MATrix LABoratory or MATLAB software[88,89]. These base classifiers were later used and assessed as a benchmark against various MCL

Sensor primitive	Interpretation method
Relative eye-gaze	Two-eye model-based gaze estimation based on RGBD[79]
Head pose	Pose from orthography and scaling with iterations (POSIT)[80]
Gaze estimation	A 3D gaze vector is achieved by combining the relative eye gaze with the calculated head pose (Fang et al.[78])
Face detection	Boosted cascade face detector[81]
Facial features	Supervised descent method proposed by[82]
Face expressions	Frontalised local binary patterns (LBP) are classified using SVM[83]
3D skeleton	Microsoft kinect SDK
Action recognition	3D joints moving trend method based on skeleton data[84,85]
Object tracking	GM-PHD tracker[86]
Sound direction	Microsoft kinect SDK

Table 1. Sensor primitives extracted by the sensorized intervention table.

systems. It was evident that the benefits of using ensembles could not be achieved by simply copying an individual model and combining the individual predictions. For this reason, all possible combinations of the number of classifiers per ensemble were explored (i.e. from two classifiers per ensemble to five classifiers). These ensembles are defined as Multiple Classifier Learning Systems 2 (MCL 2) (for two classifiers per ensemble),Multiple Classifier Learning Systems 3 (MCL 3) (for three classifiers per ensemble); Multiple Classifier Learning Systems (MCL 4) (for four classifiers per ensemble), and Multiple Classifier Learning systems 5 (MCL 5) (for all five classifiers in the ensemble).

To assess the performance of the base classifiers, the training set—validation set—test set methodology is employed. First, the dataset was split randomly into a 60% training set for each run, a 30% validation set and a 10% testing set. To test the effectiveness of the classifiers, the dataset was further split randomly into 5-folds. The smoothed error rate (i.e. smoothing the normal error count using estimates of posterior probabilities and the posterior probabilities using Bayesian estimation with conjugate priors) was used as a performance measure in all the experiments. This rate was used primarily for its variance reduction benefit and for dealing effectively with a tie between two competing classes[90]. The F-measure (score) was also used as a performance measure for the single classifier empirical comparison experiments. The benefit of the F-measure is that it considers the models' ability over two class attributes, which makes it a robust gauge of model performance.

Feature (factor) ranking and selection methods have been implemented with two basic steps of a general architecture for our experiments: subset generation and subset evaluation for the ranking of each feature in every dataset. Then, the filter method is used to evaluate each subset. Overall, a mutual information-based approach on the single classifier that exhibits the lowest error rate is utilised for this task. Mutual information calculates the reduction in entropy from the transformation of a dataset. The technique is summarised below.

A DT classifier has implicit feature selection during the model-building process. It identifies and ranks the features (factors) that significantly impact or contribute to ASD. The set of features available forms the input to the algorithm with a DT as output. The purpose of this technique was to discard irrelevant or redundant features (factors) from a given vector. For the paper, feature (factor) selection was used by evaluating the mutual information gain of each variable in the context of the target variable (robot-child against human-child therapy).

The fixed-effect model (Kirk[91]) is used to test for statistical significance of the main effects (i.e. the five single classifiers; twenty-three multiple classifier systems, three multiple classifier architectures and five resampling procedures) versus their respective interactions. Each experiment is randomly replicated five times (5-fold) making it a total of $5 \times 23 \times 3 \times 5 \times 5 = 8625$ experiments.

Experimental results

Experimental results on the ASDT predictive performance of single classifiers (on the one hand) and MCLS (on the other hand) are described. The behaviour of multiple classifiers is explored for different MCLS architectures and resampling procedures.

The results are presented in three parts.

The first part compares the performance and robustness of five single-classifier learning systems in predicting ASDT in autistic children. The second part investigates the performance of MCLS (i.e. ensembles, resampling procedures, and architectures) to determine if there is an improvement in ASD therapy predictive accuracy. These overall results are for each MCL system. They are averaged for all ensemble learning combinations about resampling procedures and architectures. Then, the experimental comparison of MCL systems (for all possible ensemble combinations) is presented. Finally, the behavioural factors (in ranking order) that contribute to and are critical when addressing the ASD problem have been identified.

Figures 1, 2, 3, 4, 5, 6, 7, 8 and 9 plot the smoothed error of the instances learned on the target domain, averaged over five-fold cross-validation runs by each one of the methods. The same folds were used to evaluate each method. All the main effects (i.e. base or single classifier systems, MCL systems, resampling procedures, and MCL systems' architectures) were significant at the 5% level, with F-ratios of 131.7, 71.4, 513.6 1132.6, respectively.

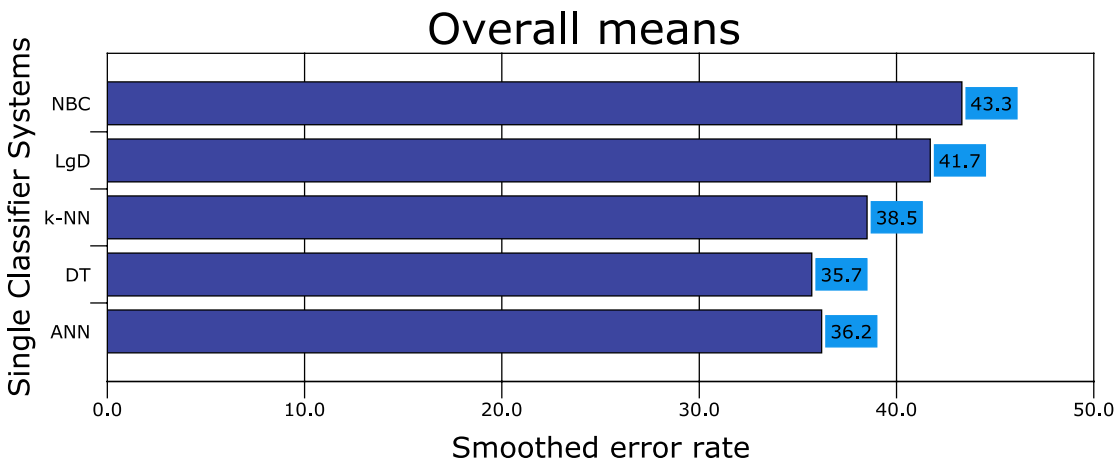


Figure 1. Single classifier systems.

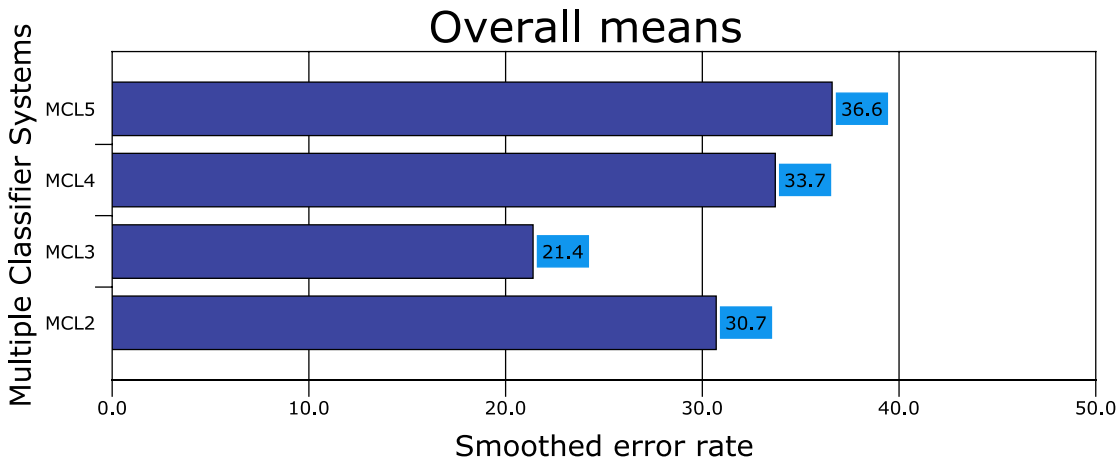


Figure 2. Multiple classifier systems.

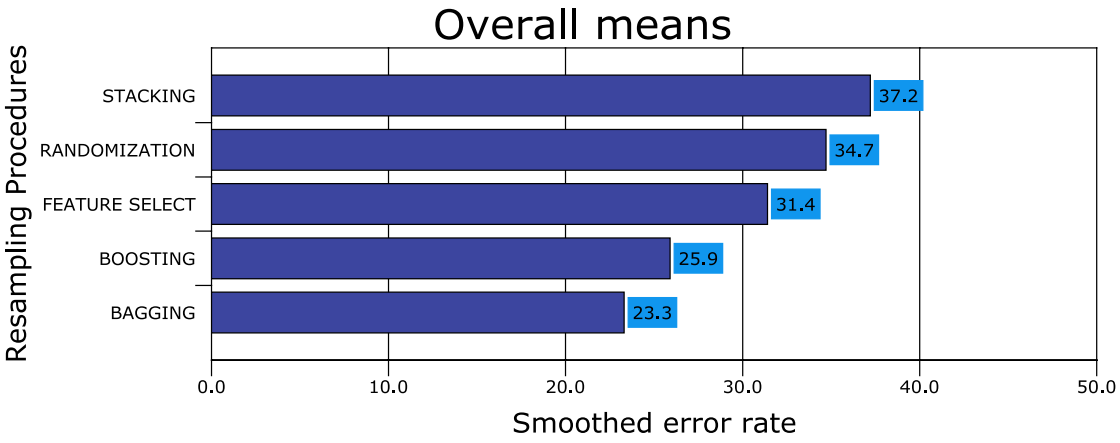


Figure 3. Resampling procedures.

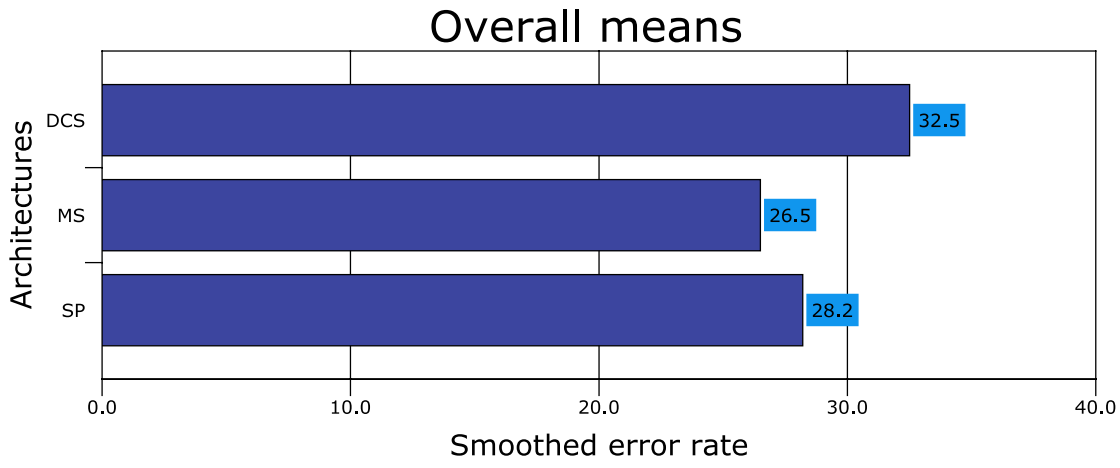


Figure 4. Multiple classifier systems architectures.

From Fig. 1, it follows that DT is the best base classifier, exhibiting a smoothed error rate of 35.7% ± 1.7% (or 0.643 accuracies; F-measure of 0.631). The second-best base classifier is ANN, followed by *k*-NN and LgD with smoothed error rates of 36.2% ± 2.2% (0.625 accuracies; F-measure 0.6035), 38.5% ± 1.6% (0.618 accuracies; F-measure 0.603), and 41.7% ± 1.9% (0.583 accuracies; F-measure 0.575). Finally, the worst performance is by the NBC with a smooth error rate increase of 43.3% ± 1.4% (0.567 accuracies; F-measure 0.531). The most

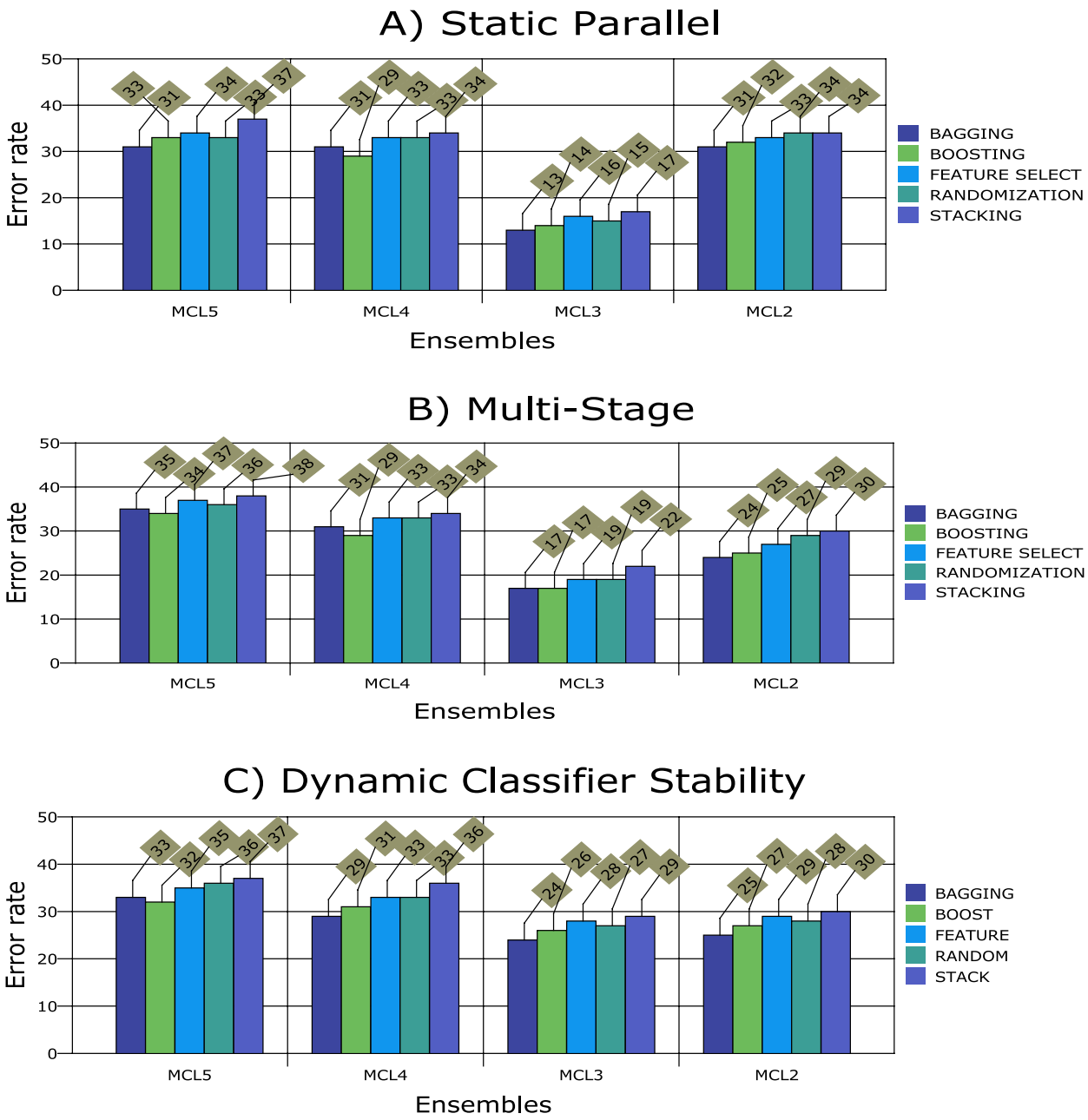


Figure 5. Multiple classifier learning systems (overall results).

relevant attributes to predicting ASD therapy are communication (non-verbal), eye contact and social interaction for the single classifiers.

From Fig. 2, the performance of the single baseline model is taken as the reference point. It appears that the performance of all the MCL systems is statistically significant at the 95% confidence level compared to single-classifier learning systems. The MCL systems when the ensemble size is composed of only three classifiers achieve the least smoothed error rate of 21.4% ± 1.9%. All ensembles with only two classifiers exhibit the second-best performance (a smoothed error rate of 30.7% ± 1.8%), while those with only four classifiers take the third spot. The worst performance is when the ensemble comprises all five classifiers (with a smoothed error rate of 36.5% ± 2.2%). The difference in performances between the four ensembles is statistically significant at the 5% significance level. Eye contact and social interaction were the most relevant features when predicting ASD (using multiple classifier systems).

All ensemble classifiers with bagging achieve the lowest error rate (23.3% ± 1.9%), followed by boosting (25.9% ± 1.5%), feature selection (31.4% ± 1.7%) and randomization (34.7% ± 1.5%), respectively. Stacking ensemble classifiers achieve the lowest accuracy rate (37.2% ± 2.1%). From the accuracy point of view, the performance differences of the ensemble classifiers were statistically significant with a 0.95 degree of confidence (Fig. 3).

From Fig. 4, it appears that all the multiple classifier systems have a more significant robust effect when the multi-stage design is used as an architecture (accuracy rate of 73.5% ± 1.8%), followed by static-parallel and

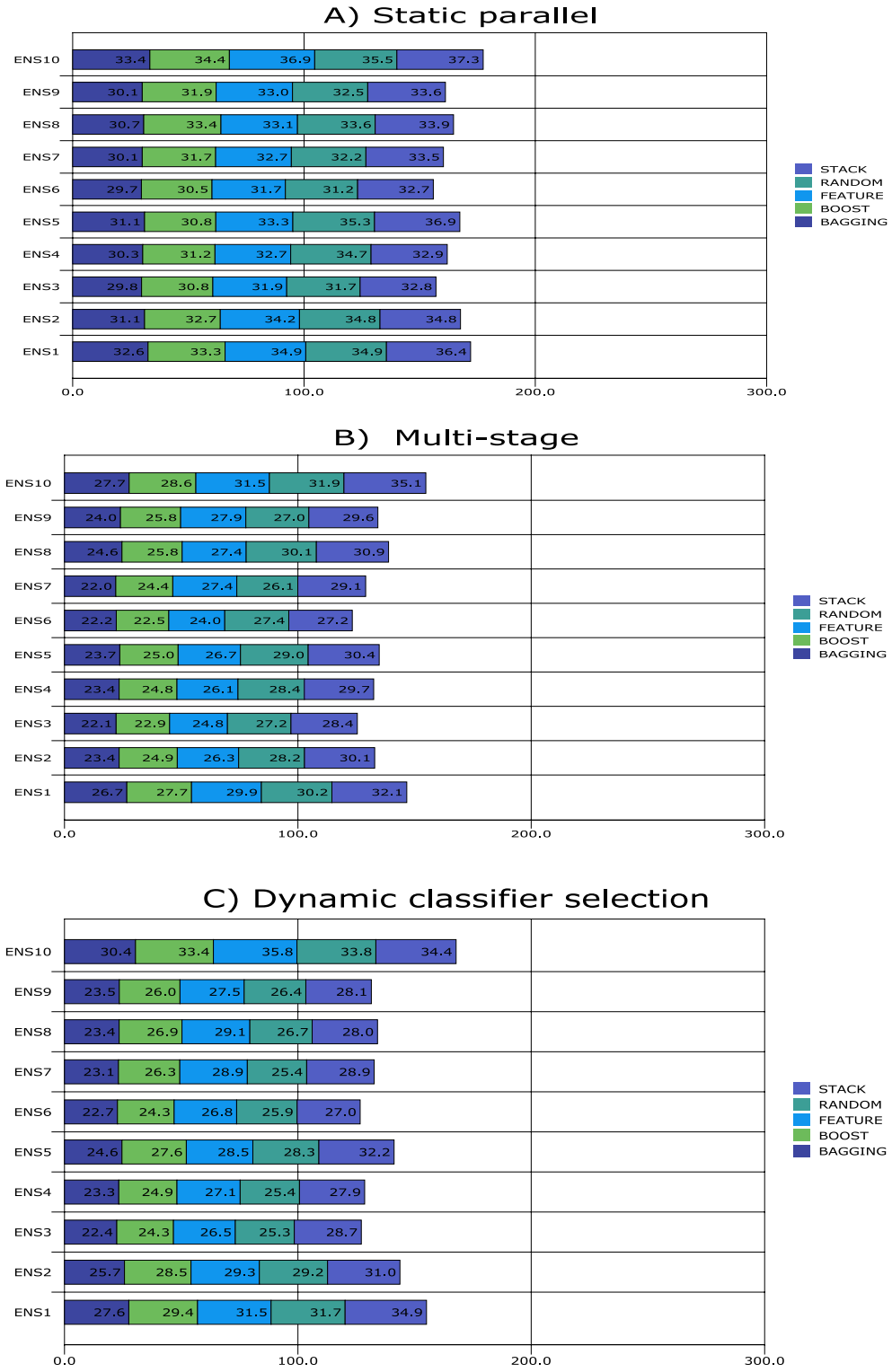


Figure 6. Multiple classifier learning systems **2**.

dynamic classifier selection with accuracy rates of 71.8% ± 1.3% and 67.5% ± 1.7%, respectively. The difference in performance between the architectures was significant at the 5% level.

The results presented in Fig. 5 show all the MCLS performing worse under the dynamic classifier selection (an error rate of 32.5% ± 1.7%) compared with single parallel and multi-stage. On the other hand, the MCLS performs slightly better when the multi-stage architecture design is used (26.5% ± 1.5%) than a single parallel (28.2% ± 2.3%). Thus, the difference in performance between the three architectures was significant at the 5% significance level (following a similar pattern to Fig. 4 results).

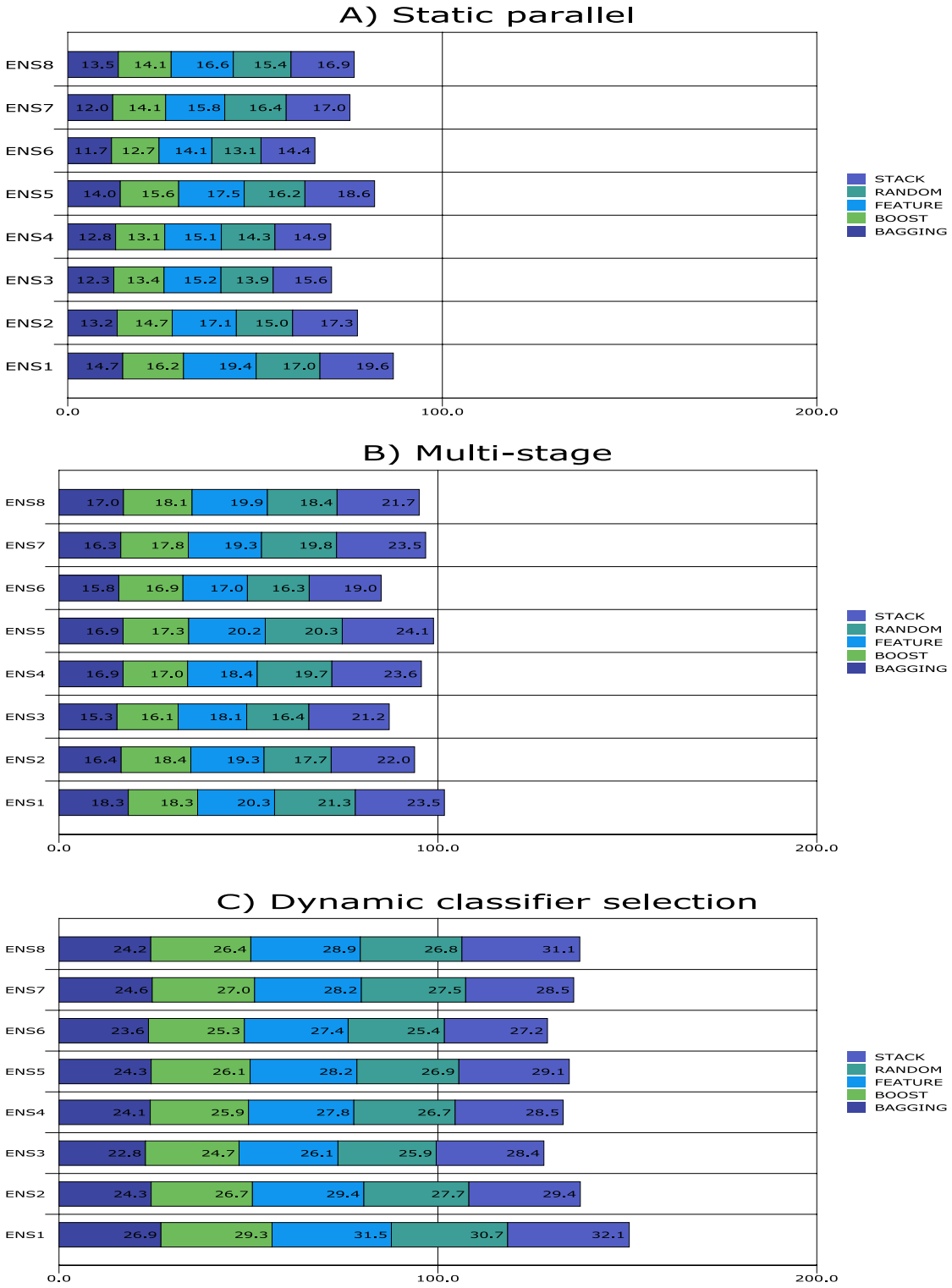


Figure 7. Multiple classifier learning systems **3**.

The three-way interaction effect between multiple classifier learning systems, architectures and resampling procedures was found to be statistically significant at the 5% level. This means that the interaction between two attributes is different across the levels of the third attribute. In other words, there was a two-way interaction between resampling methods and multiple classifier learning systems varying across architectures; a two-way interaction between architectures and resampling methods varying across multiple classifier learning systems; and a two-way interaction between architectures and multiple classifier learning systems varying across resampling methods. The results are summarised in Fig. 5.

It follows that all MCLS perform differently from each other when predicting ASD therapy, with significant error rate increases observed for ensembles with five or four classifiers compared to those with three or two

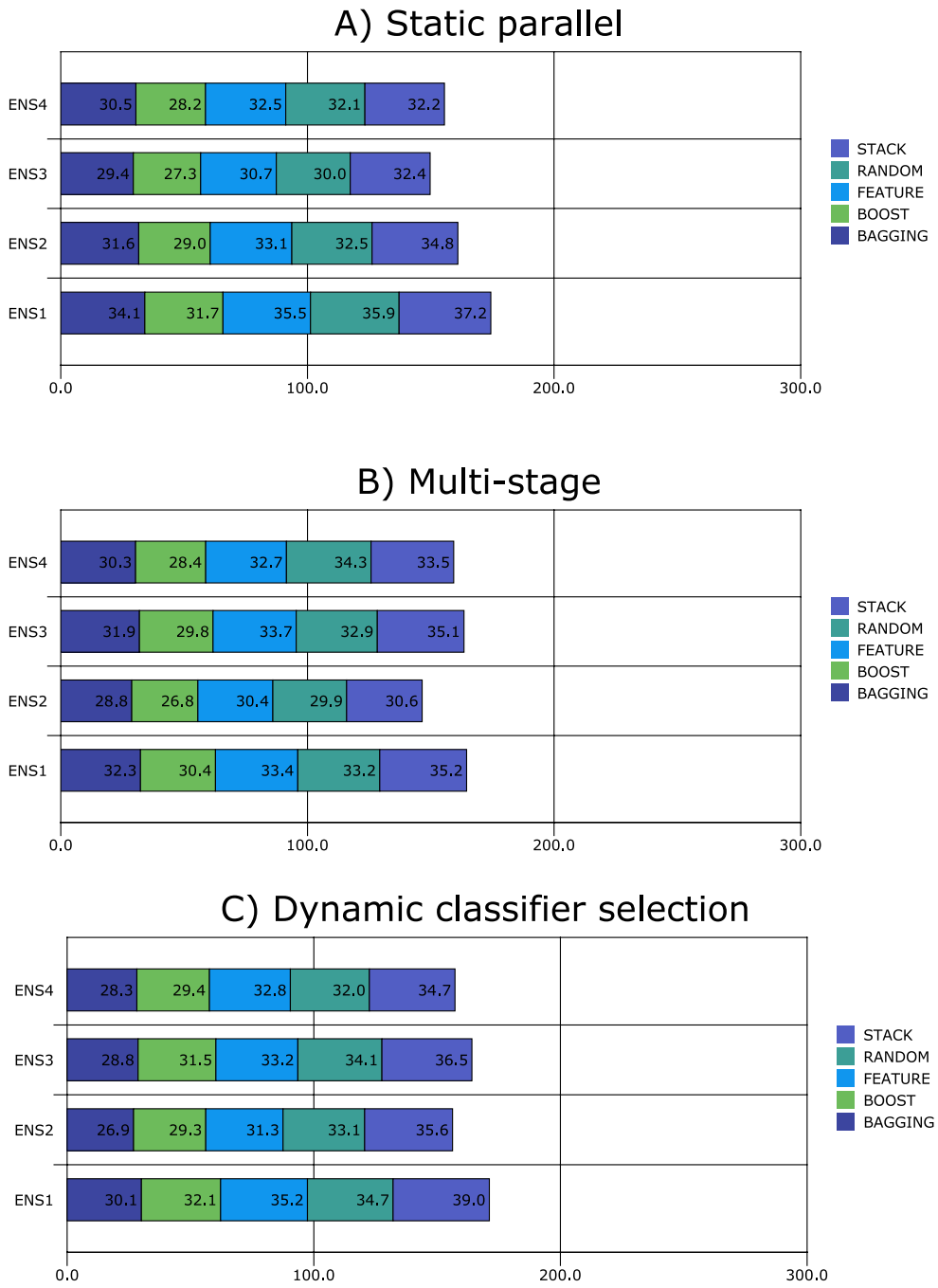


Figure 8. Multiple classifier learning system **4**.

classifiers per ensemble. Ensembles of three single classifiers achieve the highest accuracy rates with ensembles of five single classifiers achieving the lowest accuracies. Ensembling with boosting outperforms the other resampling methods with ensembling with stacking achieving the lowest accuracy rate. This is the case across all three architectures.

Static parallel multiple classifier learning for ASD therapy prediction achieves the highest accuracy, followed by multi-stage and dynamic classifier stability multiple classifier systems, respectively. Once again ensembling with bagging achieves the highest accuracy rates with poor performance for ensembling with stacking. This is the case across all multiple classifier learning systems.

The performance of all multiple classifier systems in terms of predicting ASDT is significantly different across all three architectures. Major differences are noticeable for multi-stage design against single parallel, with minor differences observed for multi-stage design against dynamic classifier selection. Once again, the results show that MCLS built through bagging is the best technique for predicting ASDT, followed by boosting, feature selection, randomisation and stacking, respectively.

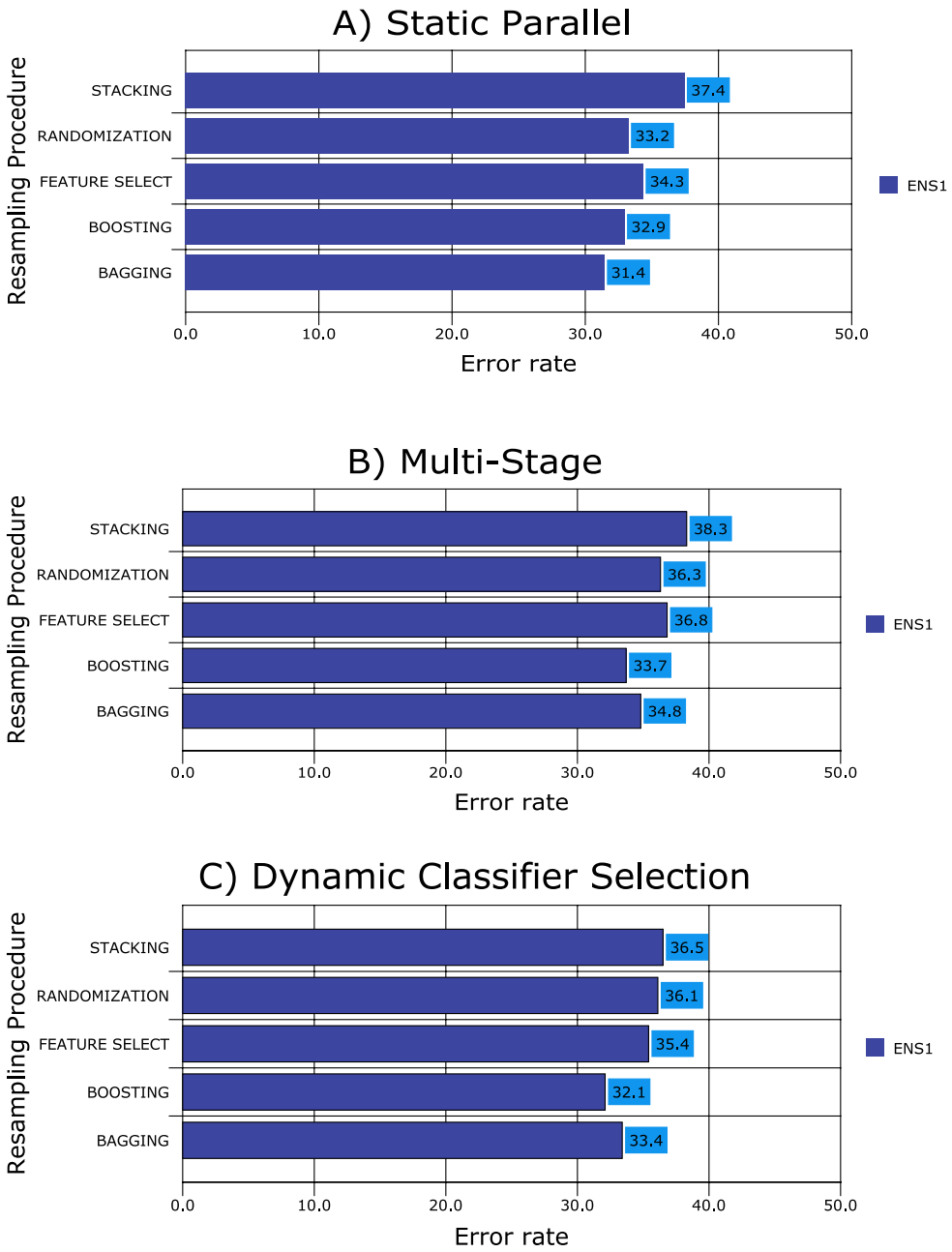


Figure 9. Multiple classifier learning systems **5**.

From Fig. 6A, the effect of the resampling procedures multiple classifier learning system 2 (MCLS 2) is transparent. MCLS 2 exhibits the worst performance for stacking, closely followed by feature selection and randomisation. The best overall performance for a static parallel architecture comes about when bagging is used. In contrast, the best performance is observed when decision trees and logistic discrimination are the two components of the ensemble. The ensembles of artificial neural networks and decision trees and logistic discrimination and naïve Bayes classifiers exhibit the worst performances.

From Fig. 6B, bagging exhibits minor error rate increases (with tight competition from boosting) for MCLS 2 and when the multi-stage architecture is used. One striking outcome is the artificial neural networks and logistic discrimination ensemble performance, which compares favourably with a Decision tree and logistic discrimination ensemble. However, the ensembles with artificial neural networks decision trees and logistic discrimination and naïve Bayes classifiers exhibit one of the worst performances for MCLS 2. Another poor performance is when the *k*-nearest neighbour and naïve Bayes classifiers are used as ensemble components, primarily when randomisation is used.

The dynamic classifier selection system is observed. At the same time, stacking continues to struggle and achieves the worst performance, especially when the artificial neural network and a decision tree (on the one hand) and logistic discrimination and naïve Bayes classifier (on the other hand) are components of the ensemble

(Fig. 6C). The best-performing ensembles are artificial neural network and logistic discrimination, and the decision tree and logistic discrimination are components.

The performance of methods for multiple classifier learning (MCL3) follows a similar pattern to the one observed for MCL2 (Fig. 7).

Figure 7A also shows smaller increases in error rates for all the resampling procedures for static parallel than the same architecture for MCL2. The best-performing ensemble is when Decision Trees, *k*-nearest neighbour and logistic discrimination are components. On the other hand, poor performances are observed when the Artificial Neural Network, Decision Trees k-nearest neighbour and Artificial Neural Network, and *k*-nearest neighbour and naïve Byes classifiers are components of the ensemble. This is the case for the feature selection and stacking resampling procedures.

The methods for multi-stage design (Fig. 7B) are nearly identical to those observed for MCLS2, with all ensembles achieving higher accuracy rates when bagging and boosting are used. Otherwise, on average, the performance of all the methods worsens when stacking is used. The best-performing ensemble is the decision tree, *k*-nearest neighbour, and logistic discrimination (primarily feature selection, randomisation and stacking). For stacking, the ensemble method composed of an Artificial Neural Network, *k*-nearest neighbour, and naïve Bayes classifiers proves to be the worst-performing.

The impact of MCLS 3 on predictive accuracy is shown in Fig. 7C. Once again, bagging yields the best performance, closely followed by boosting with severe competition from randomisation. Once again, the best-performing ensemble is when the artificial neural network, decision trees and naïve Bayes classifiers are components. The ensemble with the *k*-nearest neighbour, logistic discrimination and naïve Bayes classifier drops from being the third-best performing (when stacking and multi-stage design is used) to being one of the worst (when stacking and dynamic classifier election are used).

Overall, all the MCLS 3 systems perform better when static parallel is used, followed by multi-stage and dynamic classifier selection.

Figure 8A follows that when using static parallel to build an MCLS 4, boosting is the best technique for dealing with the ASD spectrum disorder problem, with an Artificial Neural Network, *k*-nearest neighbour, logistic discrimination, and naïve Bayes classifiers as components for the ensemble. On the other hand, the ensemble with an artificial neural network, decision trees, *k*-nearest neighbour and logistic discrimination as components achieves the worst performance. This is the case at all resampling procedure levels (i.e. bagging, boosting, feature selection, randomisation and stacking).

It follows from Fig. 8B that the best technique for handling ASD Spectrum Disorder for a multi-stage design and across various resampling procedures is boosting, closely followed by bagging. However, poor performances are observed for feature selection, randomisation, and stacking methods. Also, the ensemble with an artificial neural network, decision trees, *k*-nearest neighbour and logistic 0 discrimination as components exhibit the worst performance.

For MCLS 4, bagging using dynamic classifier selection shows superior performances to the other resampling procedures (Fig. 8C). The best-performing ensemble (across bagging, boosting, and feature selection) is where components of artificial neural network, decision trees, k-nearest neighbour and naïve Bayes. On the other hand, randomisation and stacking of an ensemble with an Artificial Neural Network, k-nearest neighbour, logistic discrimination, and naïve Bayes perform best.

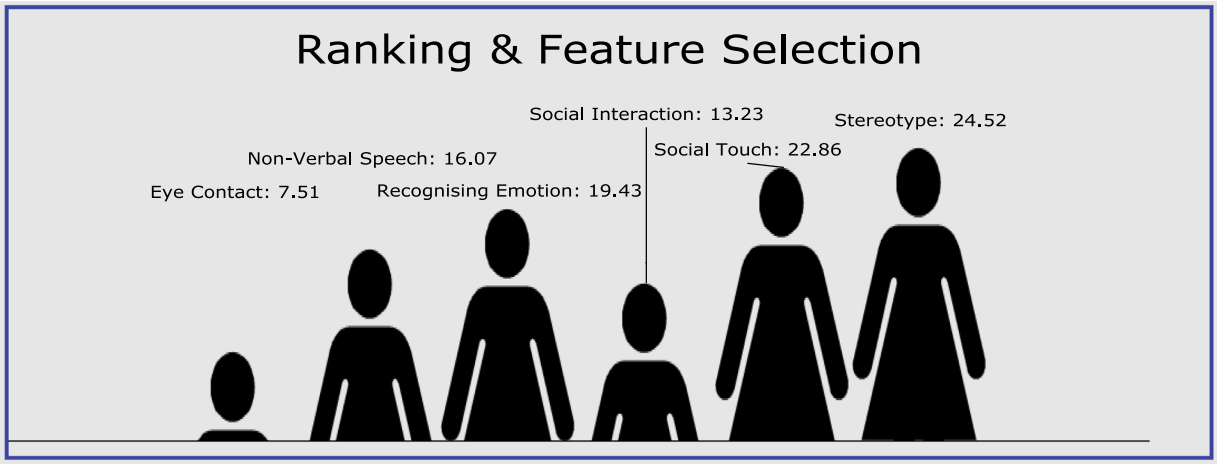
For this kind of problem, it seems that building an MCLS 5 using a static parallel architecture performs better compared with other architecture such as dynamic classifier selection and multi-stage design (Fig. 9A–C). Additionally, ensembling learning with boosting appears to be more effective especially when dynamic classifier selection and multi-stage design are used while bagging appears to be more effective when static parallel is used, outperforming resampling methods like feature selection, randomisation and stacking in some situations. Another good performance is when static parallel and multi-stage design ensemble learning is used with randomisation. Overall, ensemble learning with stacking is the worst-performing method across all three architectures.

Social difficulties are a core of ASD with one of the many psychological factors being the lack or low levels of joint attention with the interaction partners. Given the attention the use of social robots has received in ASD interventions, it was important to investigate the most significant attributes that contribute to ASD therapy (i.e. RAAT vs. SHT) and rank them accordingly. Such ranking will help investigate if RET produces similar patterns in comparison with SHT.

Feature selection is one of several ranking approaches that have been used for dealing with the high dimensionality of data and improving classification accuracy[92]. One of the goals of feature selection in machine learning is to find the best features to build applicable models of a studied phenomenon (for example, removing non-informative or redundant ASD predictors from the model). There are many feature selection algorithms including filtering, encapsulation and embedded ones (Tang and Pen[93–95]. Many feature selection techniques are classified into supervised (wrapper filter, intrinsic, embedded) and unsupervised learning (for unlabelled data).

The goal of feature selection techniques in artificial intelligence is to find the best set of features that allows one to build optimised models of a studied phenomenon (ASDT in this case). There are many feature selection algorithms but for this paper, we use the classic methods for constructing a decision tree which is the same process of feature selection. The decision tree algorithm (a supervised learning and embedded approach) was used to select the features in ranking order and according to the mutual information criterion whereby node impurities in the decision tree are utilised. The strengths of the decision tree algorithm are high classification accuracy and strong robustness. The feature selection process algorithm results modelled to obtain features considered most relevant to ASD therapy enhancement and their merit value ranks each feature are analysed and summarised in Fig. 10.

In terms of ranking, the results show *eye contact* yielding the slightest cross-validation error (7.51% ± 0.14%) followed by *social interaction* (13.23% ± 0.34%) and *non-verbal speech* (16.07% ± 0.28%), respectively. Otherwise, *social touch* and *stereotype* are the two features exhibiting error rates of more than 20%, i.e. (22.86% ± 0.19%)



Feature	Cross validation error
Eye contact	**7.51**% ± 0.14%
Non-verbal speech	16.07% ± 0.28%
Recognising emotion	19.43% ± 0.12%
Social interaction	13.23% ± 0.34%
Social touch	22.86% ±0.19%
Stereotype	24.52% ± 1.45%

Figure 10. ASD therapy features sorted by relevance.

and (24.52% ± 1.45%), respectively. In addition, all the features were significantly different at the 5% level of significance. In other words, eye contact for autistic children appears to have more impact on ASD-enhanced therapy compared to, say, stereotypes or social touch.

Remarks and conclusion

In this paper, novel research is performed regarding the exploration and prediction intervention use in autistic children using ASD-specific characteristics. Open questions related to predicting with confidence addressed include: How can ASDT data be utilised effectively to achieve more efficient confidence-based predictions using ensemble classifiers? To this end, the significant contributions of the paper include showing the robustness of single classifiers for predicting ASDT enhancement using a social robot against a human (therapist). Additionally, it shows how MCLS provides therapy enhancement performance improvements over the single base classifier (including the best-performing one). Further, a tree-based approach is used to quantitatively determine the importance of each physical attribute (according to mutual information-based ranking). Additionally, the conclusions are that single training classifiers can obtain influential ensembles in several different ways. Still, that high average individual accuracy or much diversity would generate influential ensembles.

There are several notable takeaways from this work.

First, ensembles are built with a combination of three classifiers and using bagging to achieve the perfect fit. The good performance of these ensembles could be attributed to the stable nature of nearest neighbour and linear threshold algorithms when they were core components of the ensemble. Ensembles built with dynamic classifier selection by segmenting the population into several sub-regions consistently perform poorly. However, the performance of most static parallel and multi-stage combination ensemble strategies provides statistically significant improvements over the single best classifier. We understand that in very large datasets, randomisation is expected to do better than, say, bagging or boosting but given the size of the ASDT data, bagging achieved the best results.

Eye contact and interactive communication appear to be the critical behavioural factors to be considered when dealing with ASD therapy. However, this can be argued because of the inability of children with this disorder to communicate and use language, which depends heavily on their intellectual and social development. In other words, some children with ASD may not be able to communicate using speech or language, and some may have minimal speaking skills. Therefore, joint attention in children could be another factor that needs consideration when dealing with ASDT.

Previous studies did not provide a clear conclusion about the predictive accuracy of multiple classifier systems for intervention use in autistic children. This study was the first of its kind focusing on predictive intervention use of ASDT using single classifiers and multiple classifier systems. When creating confidence-based predictors using conformal prediction, several open questions regarding how knowledge can be extracted from data using ensemble learning. The study has utility for researchers, clinicians and parents alike. It affords the potential to learn and become socially fluent no matter how strong the autism impairments may be. Although a cure for ASD has not been found yet, accurate prediction of ASDT could lead to improved outcomes or even a complete cure. Additionally, the study paves the way for investigating if an Artificial Intelligence device could be programmed

to notice and react to verbal and non-verbal responses. These could include facial expressions, body movements, and vocal and physiologic reactions from an autistic child (i.e. could artificial intelligence replace a therapist?). This assertion is based on our study where the application of AI for ASDT prediction shows promising results.

The study is based on children over the age of 3 years. In a subsequent study, the datasets of children of all ages will be critically analysed to train the therapeutic prediction model. The focus will be to collect more data from various sources and age groups and further improve the proposed ML classifier to enhance its accuracy. Furthermore, a more state-of-the-art classification method (including single classifiers like support vector machines that were not considered in these experiments) will also be considered. The focus of the study was on children with ASD, our next research will be on both autistic children against autistic adults. Additionally, our study was purely focused on behaviours. Future work will investigate specific cognitive mechanisms that might be targeted or affected by robot vs. human interactions.

In sum, this research provides an effective and efficient approach to predicting and detecting ASD traits for children above three years. This is because tests and diagnoses of ASD traits are costly and lengthy. The difficulty of detecting ASD in children and adolescents does not help another cause for the delay in diagnosis. Thus, with the help of accurate ASD spectrum disorder predictive accuracy, an individual can be guided early to prevent the situation from getting any worse and reduce costs associated with such delay.

Data availability

Many thanks are extended to the Swedish Data National Service for making the dataset available. The data that support the findings of this study are available from the University of Skövde, Skövde, Sweden but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University of Skövde, Skövde, Sweden. [Contact: Prof Erik Billing < erik.billing@his.se > . There are no biomedical financial conflicts of interest to disclose. Also, this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Received: 13 January 2023; Accepted: 31 October 2023
Published online: 15 November 2023

References

1. Centres for Disease Control and Prevention. *Autism Spectrum Disorder (ASD): Data & Statistics. [Online]*. https://www.cdc.gov/ncbddd/ASD/data.html (Accessed 24 March 2021)
2. World Health Organization *International Classification of Diseases for Mortality and Morbidity Statistics* (11th Revision) (2018).
3. Sowetan LIVE *Children with ASD are excluded from school* [Online]. https://www.sowetanlive.co.za/news/south-africa/2022-04-05-children-with-autism-excluded-from-schools/, (Accessed April 2022) (2022).
4. National Health Service *What is autism? [Online]*. https://www.nhs.uk/conditions/ASD/ (Accessed March 2022) (2022).
5. World Health Organization *Autism Spectrum Disorders* [Online]. http://www.who.int/news-room/fact-sheets/detail/ASD-spectrum-disorder (Accessed 22 August 2020) (2017).
6. Wolff, J. J. & Piven, J. Predicting autism in infancy. *J. Am. Acad. Child Adolesc. Psychiatry* **60**(8), 958–967 (2020).
7. Hong, S.-K. *et al.* Toward neuro subtypes in autism. *Biol. Psychiatry* **88**(1), 111.128 (2020).
8. Leroy, G. Irmscher, A. & Charlop-Christy, M.H. Data mining techniques to study therapy success with autistic children. In *International Conference on Data Mining*, 26-29 June 2006, Monte Carlo Resort (2006).
9. Breiman, L. Bagging predictors. *Mach. Learn.* **26**(2), 123–140 (1996).
10. Freund, Y. & Schapire, R. A decision-theoretic generalisation of online learning and an application to boosting. *J. Comput. Syst.* **55**, 119–139 (1996).
11. Wolpert, D. Stacked generalisation. *Neural Netw.* **5**(2), 241–259 (1992).
12. Ho, TK. Random decision forests. In *Proc. of the 3rd International Conference on Document Analysis and Recognition*, 278–282 (1995).
13. Dietterich, T. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomisation. *Mach. Learn.* **40**(2), 139–158 (2000).
14. Buescher, A. V. S., Cidav, Z., Knapp, M. & Mandell, D. S. Costs of autism spectrum disorder in the United Kingdom and the United States. *J. Am. Med. Assoc. Paediatr.* **168**(8), 721–728 (2014).
15. Dawson, G., Rieder, A. D. & Johnson, M. H. Prediction of autism in infants: Progress and challenges. *Lancet Neurol.* **22**(3), 244–254 (2023).
16. Soul, J. S. & Spence, S. J. Predicting autism spectrum disorder in very preterm infants. *Paediatrics* **146**(4), e2020019448 (2020).
17. Klingspor, V., Morik, K. & Rieger, A. Learning concepts from sensor data of a mobile robot. *Mach. Learn. Spec. Issue Robot Learn.* **23**(2–3), 305–332 (1995).
18. Anzalone, S. M., Boucenna, S., Ivaldi, S. & Chetouani, M. Evaluating the engagement with social robots. *Int. J. Soc. Robot.* **7**(4), 465–478 (2015).
19. Bharatharaj, J., Huang, L., Al-Jumaily, A., Elara, M. R. & Krägeloh, C. Investigating the effects of robot-assisted therapy among children with autism spectrum disorder using bio-markers. *IOP Conf. Ser. Mater. Sci. Eng.* **234**, 012017 (2017).
20. Kim, E. S. *et al.* Social robots as embedded reinforcers of social behavior in children with ASD. *J. ASD Dev. Disord.* **43**, 1038–1049 (2013).
21. Zhang, Y. *et al.* Theory of robot mind: False belief attribution to social robots in children with and without autism. *Front. Psychol.* **10**, 1732 (2019).
22. Zhang, Y. *et al.* Could social robots facilitate children with autism spectrum disorders in learning distrust and deception?. *Comput. Hum. Behav.* **98**, 140–149 (2019).
23. Chernyak, N. & Gary, H. E. Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. *Early Educ. Dev.* **27**, 1175–1189 (2016).
24. Berenguer, C., Baixauli, I., Gómez, S., Andrés, M. D. E. P. & De Stasio, S. Exploring the impact of augmented reality in children and adolescents with autism spectrum disorder: A systematic review. *Int. J. Environ. Res. Public Health* **17**, 6143 (2020).
25. Dechsling, A. *et al.* Virtual and augmented reality in social skills interventions for individuals with autism spectrum disorder: A scoping review. *J. Autism Dev. Disord.* **52**(11), 4692–4707 (2022).
26. Zhang, M., Ding, H., Naumceska, M. & Zhang, Y. Virtual reality technology as an educational and intervention tool for children with autism spectrum disorder: Current perspectives and future directions. *Behav. Sci. (Basel)* **12**(5), 138 (2022).

27. Arnevik, E. A. & Helverschou, S. B. Autism spectrum disorder and co-occurring substance use disorder—A systematic review. *Subst. Abuse* **10**, 69–75 (2016).
28. Chand, G. B. *et al.* Two distinct neuroanatomical subtypes of schizophrenia were revealed using machine learning. *Brain* **143**(3), 1027–1038 (2020).
29. Stevens, E. *et al.* Identification and analysis of behavioural phenotypes in autism spectrum disorder via unsupervised machine learning. *Int. J. Med. Inform.* **129**, 29–36 (2019).
30. Diehl, J. J., Schmitt, L. M., Villano, M. & Crowell, C. R. The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Res. ASD Spectr. Disord.* **6**(1), 249–262 (2012).
31. Jain, S., Thiagarajan, B., Shi, Z., Clabaugh, C. & Matarić, M. J. Modelling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Sci. Robot.* https://doi.org/10.1126/scirobotics.aaz3791 (2020).
32. Kumazaki, H. *et al.* The impact of robotic intervention on joint attention in children with autism spectrum disorders. *Mol. ASD* **9**, 46 (2018).
33. Kosmicki, J. A., Sochat, V., Duda, M. & Wall, D. P. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Transl. Psychiatry* **5**(2), e514 (2015).
34. Vakadkar, K., Purkayastha, D. & Krishnan, D. Detection of autism spectrum disorder in children using machine learning techniques. *SN Comput. Sci.* **22**(5), 386 (2021).
35. Wall, D., Kosmicki, J., Deluca, T., Hastard, E. & Fusaro, V. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl. Psychiatry* **2**(4), e100 (2012).
36. Bone, B. *et al.* Use of machine learning to improve Autism screening and diagnostic instruments: Effectiveness efficiency and multi-instrument fusion. *J. Child Psychol. Psychiatry* **57**, 927–937 (2016).
37. Allison, C., Auyeung, B. & Baron-Cohen, S. Toward brief "red flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1000 cases and 3000 controls. *J. Am. Acad. Child Adolesc. Psychiatry* **51**(3), 338 (2012).
38. Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A. & Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage Clin.* **17**, 16–23 (2018).
39. Dewi, E. S. & Imah, E. M. Comparison of machine learning algorithms for autism spectrum disorder classification. In *International Joint Conference on Science and Engineering (IJCSE 2020)* (eds Dewi, E. S. & Imah, E. M.) (Atlantis Press, 2020).
40. Dvornek, N.C., Yang, D., Venkataraman, A., Ventola, P., Staib, L.H., Pelphrey, K.A., and Duncan, J.S. Prediction of Autism Treatment Response from Baseline fRMI using Random Forests and Tree Bagging. https://arXiv.org/arXiv:1805.09799 (2018).
41. Bala, M., Ali, M. H., Satu, M. S., Hasan, K. F. & Moni, M. A. Efficient machine learning models for early stage detection of autism spectrum disorder. *Algorithms* **15**, 166 (2022).
42. Xu, L., Geng, X., He, X., Li, J. & Yu, J. Prediction in autism by deep learning short-time spontaneous hemodynamic fluctuations. *Front. Neurosci.* **13**, 1120 (2019).
43. Akter, T. *et al.* Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access* **7**, 166509–166527 (2019).
44. Kanchana A., and Khilar, R. Prediction of autism spectrum disorder using random forest classifier in adults. *IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Goa, India, 242–249 (2022).
45. Fiske, A., Henningsen, P. & Buyx, A. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J. Med. Internet Res.* **21**(5), e13216 (2019).
46. Luxton, D. *Artificial Intelligence in Behavioral and Mental Health Care* 308 (Academic Press, 2015).
47. Cox, D. R. Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman* (ed. David, F. N.) 55–71 (Wiley, 1966).
48. Day, N. E. & Kerridge, D. F. A general maximum likelihood discriminant. *Biometrics* **23**(2), 313–323 (1967).
49. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation. In *Parallel Distributed Processing* Vol. 1 (eds Rumelhart, D. E. & McClelland, J. L.) 318–362 (MIT Press, 1986).
50. Jolliffe, I. *Principal Component Analysis* (Springer Verlag, 1986).
51. Hosmer, D. W. & Lameshow, S. *Applied Logistic Regression* (Wiley, 1989).
52. Aha, D. W., Kibbler, D. W. & Albert, M. K. Instance-based learning algorithms. *Mach. Learn.* **6**(37), 37–66 (1991).
53. Ripley, B. D. *Pattern Recognition and Neural Networks* (Wiley, 1992).
54. Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and Regression Trees* (Wadsworth, 1984).
55. Quinlan, J. R. *C.4.5: Programs for Machine Learning* (Morgan Kauffman Publishers Inc., 1993).
56. Safavian, S. R. & Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybernet.* **21**, 660–674 (1991).
57. Duda, R. O. & Hart, P. E. *Pattern Classification* 2nd edn. (Wiley, 1973).
58. Kononenko, I. Semi-naïve Bayesian classifier. In *Proceedings of the European Conference on Artificial Intelligence* (ed. Kodratoff, Y.) 206–219 (Springer, Berlin Heidelberg, 1991).
59. Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *Eur. J. Oper. Res.* **210**(2), 368–378 (2011).
60. Twala, B. Toward accurate software effort prediction using multiple classifier systems. In *Computational Intelligence and Quantitative Software Engineering* (eds Pedrycz, W. *et al.*) 135–151 (Springer-Verlag, 2016).
61. Twala, B. Multiple classifier learning to credit risk assessment. *Expert Syst. Appl.* **37**(2010), 3326–3336 (2009).
62. Zhu, H., Beling, P. A. & Overstreet, G. A. A study in the combination of two consumer credit scores. *J. Oper. Res. Soc.* **52**, 2543–2559 (2001).
63. Schapire, R., Freund, Y., Bartlett, P. & Lee, W. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proc. of International Conference on Machine Learning*, Morgan Kaufmann, 322–330 (1997).
64. Kittler, J., Hatef, M., Duin, R. P. W. & Matas, J. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 226–239 (1998).
65. Kuncheva, L. I. Switching between selection and fusion in combining classifiers: An experiment. *IEEE Trans. Syst. Man Cybern. B Cybern.* **32**, 146–156 (2002).
66. Hernandez-Lobato, D., Martınez-Munoz, G. & Suarez, A. How large should ensembles of classifiers be?. *Pattern Recognit.* **46**(5), 1323–1336 (2013).
67. Li, N. & Zhou, Z. H. Selective ensemble of classifier chains. *Proc. Int. Workshop Mult. Classif. Syst.* **2013**, 146–156 (2013).
68. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33**(1), 1–39 (2010).
69. Li, N., Jiang, Y. & Zhou, Z. H. Multi-label selective ensemble. *Proc. Int. Workshop Mult. Classif. Syst.* **2015**, 146–156 (2013).
70. Babenko, B., Yang, M-H. and Belongie, S. Visual tracking with online multiple instance learning. In. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)* (2009).
71. Kim, T. K. & Cipolla, R. Multiple classifier boosting and tree-structured classifiers. In *Machine Learning for Computer Vision. Studies in Computational Intelligence* Vol. 411 (eds Cipolla, R. *et al.*) (Springer Berlin, Heidelberg, 2013).
72. Jackowski, K. New diversity measure for data stream classification ensembles. *Eng. Appl. Artif. Intell.* **74**, 23–34 (2018).
73. Viola, P. Platt, J. C. & Zhang, C. Multiple instances boosting for object detection. *In 12th Annual Conference on Neural Information Processing Systems (NeurIPS 06)*, Vancouver, Canada, 5–9 December 2006, 1417–1426 (2006).

74. Mellema, C., Treacher, A., Nguyen, K., & Montillo, A. Multiple deep learning architectures achieve superior performance diagnosing ASD spectrum disorder using features previously extracted from structural and functional MRI. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1891–1895 (2019).
75. Billing, E. *et al.* The DREAM dataset: Supporting a data-driven study of Autism spectrum disorder and robot-enhanced therapy. *PLoS One* **15**(8), e0236939 (2020).
76. Cooper, J. O., Heron, T. E. & Heward, W. L. *Applied Behaviour Analysis* 2nd edn. (Pearson, 2007).
77. Gotham, K., Pickles, A. & Lord, C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J. Autism Dev. Disord.* **39**(5), 693–705 (2009).
78. Cai, H. *et al.* Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study. *IEEE Sens. J.* **9**(4), 1508–1518 (2019).
79. Zhou, X., Cai, H., Li, Y. & Liu, H. Two-eye model-based gaze estimation from a Kinect sensor. In *IEEE International Conference on Robotics and Automation* (eds Zhou, X. *et al.*) 1646–1653 (IEEE, 2017).
80. Dementhon, D. F. & Davis, L. S. Model-based object pose in 25 lines of code. *Int. J. Comput. Vis.* **15**(1–2), 123–141 (1995).
81. Viola, P. & Jones, M. J. Robust real-time face detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2004).
82. Xiong, X., De La Torre, F. Supervised descent method and its applications to face alignment. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 532–539 (2013).
83. Wang, Y., Yu, H., Dong, J., Stevens, B. & Liu, H. Facial expression-aware face fractalization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 10113 (eds Wang, Y. *et al.*) 375–388 (Springer Verlag, 2017).
84. Liu, T. *et al.* Toward fast 3D human activity recognition: A refined feature based on minimum joint freedom model. *J. Manuf. Syst.* **66**, 127–141 (2023).
85. Liu B., Yu H., Zhou X., Tang D., and Liu, H. Combining 3D joints Moving Trend and Geometry property for human action recognition. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics*, SMC 2016—Conference Proceedings. Institute of Electrical and Electronics Engineers Inc., 332–337 (2017).
86. Zhou, X., Yu, H., Liu, H. & Li, Y. Tracking multiple video targets with an improved GM-PHD tracker. *Sensors* **15**(12), 30240–30260 (2015).
87. Wu, X. *et al.* Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37 (2008).
88. Gilat, A. *MATLAB: An Introduction with Applications* 2nd edn. (Wiley, 2004).
89. MATLAB. *Version 9.6 (R2019a)* (The MathWorks Inc., 2019).
90. Twala, B. *Effective Techniques for Dealing with Incomplete Data when Using Decision Trees*. PhD thesis, Open University, Milton Keynes (2005).
91. Kirk, E. E. *Experimental Design* 2nd edn. (Brooks, Cole Publishing Company, 1982).
92. Zhou, H. F., Zhang, J. W., Zhou, Y. Q., Guo, X. & Ma, Y. M. A feature selection algorithm of a decision tree based on feature weight. *Expert Syst. Appl.* **164**, 113842 (2021).
93. Tang, P. & Peng, Y. Exploiting distinctive topological constraints of local feature matching for logo image recognition. *Neurocomputing* **236**, 113–122 (2017).
94. Gao, W., Hu, L., Zhang, P. & He, J. Feature selection considering the composition of feature relevancy. *Pattern Recognit. Lett.* **112**, 70–74 (2018).
95. Gao, W., Hu, L., Zhang, P. & Wang, F. Feature selection by integrating two groups of feature evaluation criteria. *Expert Syst. Appl.* **110**, 11–19 (2018).

Author contributions

Conception or design of the work (B.T.). Data collection (FROM DOMAIN EXPERT). Data analysis and interpretation (B.T. & E.M.). Drafting the article (B.T. & E.M.). Critical revision of the article (B.T. & E.M.). Final approval of the version to be published (B.T. & E.M.).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.