Identification of hypertension subtypes using microRNA profiles and machine learning

Reel, Smarti; Reel, Parminder Singh; Van Kralingen, Josie; Larsen, Casper K; Robertson, Stacy; MacKenzie, Scott M

Published in:
European Journal of Endocrinology

DOI:
10.1093/ejendo/lvaf052

Publication date:
2025

Identification of hypertension subtypes using microRNA profiles and machine learning

*Smarti Reel[1] (PhD), *Parminder Singh Reel[1] (PhD), *Josie Van Kralingen[2] (PhD), Casper K Larsen[3] (PhD), Stacy Robertson[2] (PhD), Scott M. MacKenzie[2] (PhD), Alexandra Riddell[2] (PhD), John D. McClure[2] (PhD), Stelios Lamprou[2] (MSc), John M.C. Connell[4] (MD), Laurence Amar[3,5] (MD, PhD), Alessio Pecori[6] (MD), Martina Tetti[6] (MD), Christina Pamporaki[7] (MD, PhD), Marek Kabat[8] (MD, PhD), Filippo Ceccato[9] (MD, PhD), Matthias Kroiss[10–13] (MD, PhD), Michael C Dennedy[14] (MD), Anthony Stell[15] (PhD), Jaap Deinum[16] (MD, PhD), Paolo Mulatero[6] (MD), Martin Reincke[13] (MD), Anne-Paule Gimenez-Roqueplo[3,17] (MD, PhD), Guillaume Assié[18,19] (MD, PhD), Anne Blanchard[20] (MD, PhD), Felix Beuschlein[13,21] (MD), Gian-Paolo Rossi[22] (MD), Graeme Eisenhofer[7,23] (PhD), Maria Christina Zennaro[3,17] (MD, PhD), **Emily Jefferson[1] (PhD), **†Eleanor Davies[2] (PhD)

*SR, PSR and JCvK contributed equally to this work.

**ED and ERJ contributed equally to this work.

†ED is corresponding author

1. Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee DD2 4BF, UK.
2. School of Cardiovascular and Metabolic Health, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow G12 8TA, UK.
3. Université Paris Cité, Inserm, PARCC, Paris, France.
4. Molecular and Clinical Medicine, School of Medicine, University of Dundee, Dundee DD2 4BF, UK.
5. Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Unité Hypertension artérielle, Paris, France.

6. Division of Internal Medicine and Hypertension Unit, Department of Medical Sciences, University of Torino, Italy.
7. Department of Medicine III, Department of Endocrinology, University Hospital Carl Gustav Carus, TU Dresden, Germany.
8. Department of Hypertension, National Institute of Cardiology, Warsaw, Poland.
9. UOC Endocrinologia, Dipartimento di Medicina DIMED, Azienda Ospedaliera-Università di Padova, Padua, Italy.
10. Clinical Chemistry and Laboratory Medicine, Core Unit Clinical Mass Spectrometry, Universitätsklinikum Würzburg, Germany.
11. Schwerpunkt Endokrinologie/Diabetologie, Medizinische Klinik und Poliklinik I, Universitätsklinikum Würzburg, Germany.
12. Comprehensive Cancer Center Mainfranken, Universität Würzburg, Würzburg, Germany.
13. Medizinische Klinik und Poliklinik IV, LMU Klinikum, LMU München, Germany.
14. The Discipline of Pharmacology and Therapeutics, School of Medicine, National University of Ireland 33 Galway, Ireland.
15. School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia.
16. Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands.
17. Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Service de Génétique, Paris, France.
18. Université de Paris, Institut Cochin, INSERM, CNRS, F-75014 PARIS, France.
19. Department of Endocrinology, Center for Rare Adrenal Diseases, Assistance Publique–Hôpitaux de Paris, Hôpital Cochin, 75014 Paris, France.
20. Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Centre d'Investigations Cliniques 9201, Paris, France.
21. Klinik für Endokrinologie, Diabetologie und Klinische Ernährung, UniversitätsSpital Zürich (USZ) und Universität Zürich (UZH), Zurich, Switzerland.

22. Internal & Emergency Medicine-ESH Specialized Hypertension Center, Department of Medicine-DIMED, University of Padua, Padua, Italy.
23. Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Carl Gustav Carus, TU Dresden, Germany.

Short title: MicroRNA Profiles of Hypertension Subtypes

Corresponding author address: School of Cardiovascular and Metabolic Health, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow G12 8TA, UK; eleanor.davies@glasgow.ac.uk.

Keywords: MicroRNA
Hypertension
Machine learning
Diagnostics

Total Word Count: 3937

ABSTRACT

Objective: Hypertension is a major cardiovascular risk factor affecting about 1 in 3 adults. Although the majority of hypertension cases (~90%) are classified as 'primary hypertension' (PHT), endocrine hypertension (EHT) accounts for ~10% of cases and is caused by underlying conditions such as primary aldosteronism (PA), Cushing's syndrome (CS), pheochromocytoma or paraganglioma (PPGL). EHT is often misdiagnosed as PHT leading to delays in treatment for the underlying condition, reduced quality of life and costly, often ineffective, antihypertensive treatment. MicroRNA circulating in the plasma is emerging as an attractive potential biomarker for various clinical conditions due to its ease of sampling,

the accuracy of its measurement and the correlation of particular disease states with circulating levels of specific microRNAs.

Methods: This study systematically presents the most discriminating circulating microRNA features responsible for classifying and distinguishing EHT and its subtypes (PA, PPGL, CS) from PHT using 8 different supervised machine learning (ML) methods for the prediction.

Results: The trained models successfully classified PPGL, CS and EHT from PHT with AUC 0.9 and PA from PHT with AUC 0.8 from the test set. The most prominent circulating microRNA features for hypertension identification of different disease combinations were hsa-miR-15a-5p and hsa-miR-32-5p.

Conclusions: This study confirms the potential of circulating microRNAs to serve as diagnostic biomarkers for EHT and the viability of machine learning as a tool for identifying the most informative microRNA species.

SIGNIFICANCE STATEMENT

Identification of secondary forms of hypertension is key to targeted management and prevention of complications but stratification of patients is a challenging diagnostic process that can delay clinical treatment. Here, we present the development of a customised machine learning pipeline that uses circulating microRNA expression profiles to identify patients with endocrine forms of hypertension, including primary aldosteronism, Cushing's syndrome and pheochromocytoma/paraganglioma. The key outcome of this work includes trained classifiers that can predict hypertension subtypes with AUC of up to 0.9 on an unseen test dataset. It provides highly compelling evidence that machine learning can be used to identify the most discriminating diagnostic circulating miRNAs and, ultimately, translate this approach into future clinical diagnostics.

Nonstandard Abbreviations and Acronyms

100RR: 100 Random Repeats

CFS: Correlation-based Feature Selection

CS: Cushing's Syndrome

EHT: Endocrine Hypertension

LB: Logitboost

LMT: Logistic Model Tree

miRNA: MicroRNA

ML: Machine Learning

NB: Naïve Bayes

NGS: Next-generation Sequencing

PA: Primary Aldosteronism

PHT: Primary Hypertension

PPGL: Pheochromocytoma or Paraganglioma

RF: Random Forest

SL: Simple Logistic

SMO: Sequential minimal optimization

INTRODUCTION

Hypertension is a significant health problem affecting a third of adults.[1,2] Approximately 85–95%[3] of cases are defined as primary hypertension (PHT) while the remainder has a multiplicity of underlying causes, including such endocrine disorders as primary aldosteronism (PA), Cushing's syndrome (CS), pheochromocytoma or paraganglioma (PPGL). These are difficult to distinguish satisfactorily and, consequently, to treat effectively.[4] Therefore, the identification of novel diagnostic biomarkers is a worthwhile endeavour that might also lead to more effective and specific treatments for hypertension in its various forms.[5]

MicroRNAs (miRNAs) are small, endogenously expressed, non-coding RNAs that negatively regulate cellular gene expression at the post-transcriptional level. MiRNA action has significant regulatory effects in many tissues and biological processes, including tumorigenesis and regulation of endocrine systems. Our previous work demonstrated that specific miRNAs produced by adrenocortical cells influence their secretion of aldosterone and cortisol. Furthermore, we found miRNA expression profiles of aldosterone-producing adenoma samples differ significantly from those of normal adrenal tissue, indicating disruption of miRNA production as a feature of this condition.[6,7] We hypothesise such changes will be reflected in the levels of specific miRNAs present in the circulatory system. Such circulating miRNAs are released into the bloodstream from various tissues of the body, either actively or passively, in a process that is still poorly understood. Nevertheless, changes in circulating miRNA levels have proven worth as specific disease biomarkers and have been investigated in cases of hypertension, identifying certain miRNAs as having prognostic, diagnostic, and therapeutic value.[8,9] Specifically, hsa-miR-21 and hsa-miR-510 showed potential owing to their increased circulating levels in hypertensive patients.[10,11] Considering

the markedly distinct molecular pathogenesis of the various EHT subtypes, we hypothesise that patients with PA, PPGL and CS are likely to have distinctive circulating miRNA profiles. Recently, ML methods have gained popularity in next-generation biomarker discovery initiatives using high throughput next-generation sequencing (NGS) data.[12–14] This has provided the motivation to use ML algorithms for miRNA-disease association prediction.[15,16] Moreover, ML has also been successfully employed to classify lung adenocarcinoma,[17] melanoma,[18] and tumor origin[19] using miRNA profiles. Some studies have focused on associating individual miRNAs with specific endocrine hypertension subtypes; hsa-miR-483-5p, hsa-miR-101 and hsa-miR-183 were each reported to associate with pheochromocytoma[20,21] and Castro-Vega et al. have discussed the genomic features underlying PPGL, including associated changes in miRNA expression.[22] Vetrivel et al. reported changes in adrenal tissue[23] and circulating miRNA[24] levels associated with different forms of CS. However, no studies have explored the use of supervised ML approaches for a combined classifier solution.

Recently, a multiomics study of a retrospectively-collected cohort of EHT and PHT patients was conducted under the ENS@T-HT project.[25] This took a multiomic approach, combining numerous datasets, including steroids (plasma and urinary), catecholamines, small metabolites and plasma miRNAs, together with a ML pipeline in order to identify optimal diagnostic signatures that distinguish PPGL, PA and CS from PHT. This analysis showed circulating miRNAs to have particularly high balanced accuracies within the generated multiomic diagnostic signatures.[25] This indicated that circulating miRNAs may have value as diagnostic biomarkers in their own right, separate from the other elements of the multiomics signature. To test this, the present study uses an ML approach to analyse the circulating miRNA dataset in isolation from the other ENS@T-HT omics. As with that multiomic analysis, the key aim here was to stratify hypertensive patients amongst themselves. EHT (PA+PPGL+CS) and its subtypes (PA, PPGL, CS) were classified from PHT using various

supervised ML methods. The algorithm was also trained to classify different hypertensives into PPGL, PA, CS, and PHT groups with the ultimate objective of developing a circulating miRNA biomarker clinical test. The classification performance was evaluated through overall classification accuracy, specificity and sensitivity. The most prominent miRNA features responsible for distinguishing hypertension subtypes were also identified.

MATERIALS AND METHODS

Omic Dataset

EDTA-plasma samples were collected from 330 male or female patients aged between 11–78 years who had been diagnosed with one of the underlying four hypertension subtypes (PA, PPGL, CS, PHT). Samples were provided by ENS@T-HT Horizon 2020 project collaborators from previous study archives. No formal sample size calculation was conducted. Specific inclusion and exclusion criteria for each hypertension subtype, including diagnostic criteria and ethics approval details are provided in Supplemental Section 1.1 and 1.2, respectively. All research was conducted in compliance with the Declaration of Helsinki. Total RNA was isolated from 200µL EDTA-plasma using the miRNeasy Mini kit (QIAGEN, Manchester, UK) standard protocol. Samples were eluted in 30µL RNase-free water. cDNA was reverse-transcribed from 4µL of undiluted RNA in a 20µL reaction volume according to the standard protocol of the Universal cDNA synthesis kit II (Exiqon, Vedbaek, Denmark). For quality control purposes, plasma samples were spiked with UniSp2, UniSp4, and UniSp5 RNAs before RNA isolation and RNA samples were spiked with cel-miR-39-3p and UniSp6 cDNA during reverse transcription using components from the miRCURY LNA™ Universal microRNA PCR System RNA Spike-in kit (Exiqon). Selected plasma miRNAs were quantified using Serum/Plasma Focus microRNA PCR Panels (384-well, V4.M, Exiqon) according to their standard protocol, using ExiLENT SYBR® Green master mix (Exiqon) and ROX solution

(Thermo Fisher, Renfrew, UK) on a Quantstudio 12K Flex Real-time PCR System (Thermo Fisher). A full list of the 179 miRNAs and 13 controls analysed on the panel is provided in Supplemental Table S1.

Raw Ct data generated by qRT-PCR using the QuantStudio System were analysed using GenEx software (v.6, MultiD Analyses, Vedbaek, Denmark). Interplate calibration was performed using UniSp3 amplification results; this enabled cross-plate comparisons. Quality control checks for both RNA isolation and cDNA preparation were performed using the spike-in controls; where the amplification value (cycle threshold, Ct) exceeded predefined limits the sample was flagged. For RNA isolation spike-in assays, this limit was +/- 2 Ct from the mean across all samples, and for cDNA synthesis spike-in assays the limit was +/- 1 Ct from the mean. Next, if fewer than 90% of miRNAs amplified in a sample, that sample was flagged. If a patient sample was flagged in 2 or more categories, it was excluded from further analysis. If a miRNA had >50% missing data it was excluded from further analysis.

Disease combinations

Table 1 provides a breakdown of the study patients by subtype, age, and sex.

Results for 5 disease combinations were assessed:

1. ALL vs ALL (i.e. PA vs PPGL vs CS vs PHT)
2. EHT vs PHT
3. PA vs PHT
4. CS vs PHT
5. PPGL vs PHT

ML Analysis pipeline

The ML analysis pipeline consisted of three key steps (See Figure 1). The first step involved exclusion of extreme outliers followed by random splitting of the data into training (80%) and

testing (20%) sets in a stratified manner, according to established practice.[26] (See Supplementary Table S2). In the second step, the most suited feature selection method and classifiers were chosen for each disease combination using random sub-sampling validation (using training/validation set). The top selected features were then saved for each scenario. Lastly, in step 3, the final model training and testing was conducted. The top performing classifiers were trained using the training set (with only the reduced feature set). The trained classifiers were then used to predict the disease type of testing data. These results were then evaluated using various performance metrics. The details of outlier detection, features selection, classifiers and evaluation scenarios are provided in Supplemental Methods (See Supplemental Section 1.3).

RESULTS

Data generation

Samples from 346 patients were analysed. Following quality checking, as detailed above, data from 330 of these were subjected to further analysis. Mean non-normalised Ct figures and standard deviations across all 330 patients for each measured miRNA are shown in Table S1. Data normalisation was performed to enable direct comparison of the sample results. Normalisation used five miRNAs identified as being most stably-expressed across the dataset by Normfinder software:[27] hsa-miR-106a-5p, hsa-miR-425-5p, hsa-miR-222-3p, hsa-let-7g-5p and hsa-let-7i-5p. Normalisation was performed for each sample by subtracting the mean Ct value of the 5 normaliser miRNAs from each of the feature miRNA Ct values, generating the deltaCT values used for subsequent analysis. Non-detected miRNA values were imputed by assigning them the value (*max* + 1), where *max* was the maximum Ct detected for that miRNA across all samples. (Note that Ct value is inversely proportional to transcript quantity.) Imputation accounted for 1.44% of all data points.

Of the 179 unique human miRNAs measured, 5 were used for normalisation and 1 (hsa-miR-208a-3p) was excluded on quality grounds; 173 human miRNAs were therefore put forward for further analysis. A total of 175 features were extracted; these comprised measured circulating levels for the 173 individual human miRNA species (See Table S1) plus the age and sex of the patient. The final dataset was catalogued in RDMP Software[28] for systematic access.

ML pipeline analysis

Results of ML analysis for different disease combinations are described in this subsection. Firstly, the performance of including/excluding outliers on classifier performance is evaluated, followed by selection of the best classifiers and feature selection method. Using these parameters, the top discriminating features are then selected for different evaluation scenarios using random subsampling. Finally, these features are used to train and test final models.

Outliers, best classifiers and feature selection methods

Figure 2A shows balanced accuracy (i.e. the mean of specificity and sensitivity), sensitivity and specificity plots for 8 classifiers, comparing results for the ALL vs ALL disease combination using two datasets i.e. including and excluding outliers. Performance improved when outliers were removed before classifying any given disease combination. The best-performing four classifiers across the three performance metrics were LB, LMT, SL, and SMO. Therefore, outliers were removed for all further analysis.

Next, for the ALL vs ALL disease combination, two feature selection methods, wrapper (Boruta) and filter (CFS), were applied. The classification was also performed by using a complete feature list (i.e. no feature reduction). The balanced accuracies, sensitivities and

specificities were compared for all three feature selection methods using the four selected classifiers LB, LMT, SL, and SMO.

Figure 2B compares classification results for these feature selection methods using the top 4 classifiers for the ALL vs ALL disease combination. It was observed that using all features for classification provided best results, followed by the wrapper (Boruta) method, with the filter (CFS) method lowest. However, the wrapper method used only the top discriminating features rather than all 175 features, and mean accuracies were comparable. The wrapper-based feature selection method 'Boruta' improved the overall accuracy compared to using the filter selection method and was therefore used for all five disease combinations.

Evaluation scenarios

Classification performance

Figure 3 shows the various performance metrics for each of the 5 disease combinations, 3 scenarios, and 4 classifiers over 100 random repeats. In Scenario 1 (comparison of Set A vs Set B), for almost all disease combinations, both Sets A & B provided balanced accuracies with ~1% variation. The balanced accuracy was highest for PPGL vs PHT (~78%) and lowest for ALL vs ALL (~66%). Also, the sensitivities for CS vs PHT and ALL vs ALL were low (< 60%) compared to the corresponding specificities (> 84%). However, the opposite case applied for EHT vs PHT, where the sensitivities were higher (> 85%) compared to the corresponding specificities (< 57%). Similar trends were observed for F1, AUC and Kappa score. There was marginal difference between Sets A and B, indicating that age and sex are not highly differentiating features in circulating miRNA-based endocrine hypertension classification.

The non-uniform number of samples in the different Sets of Scenarios 2 & 3 does not validate their direct metric comparison. However, it was useful in evaluating the discriminating

features. The performance metrics were higher for Set D (female subset) than for Set C (male subset) in all disease combinations for the majority of classifiers, except for CS vs PHT where the specificity was lower for Set D. The lowest balanced accuracy (50%) and zero sensitivity (but 100% specificity) was observed for the male subset (Set C) during CS vs PHT classification due to an extremely low number of male samples compared to female samples (in Set D). Overall, Set D showed the highest balanced accuracy for PA vs PHT (83%) using the SMO classifier. Similar results were observed for the remaining three metrics.

The disease combinations were also compared on the basis of patient age, i.e. >=50 years (Set E) or <50 years (Set F). Better metrics were observed for Set E (older patients) for all disease combinations than for Set F except in CS vs PHT. The highest balanced accuracy of 80% was observed in EHT vs PHT (Set E) for all classifiers, except LB.

Significant Features

Figure 4A shows a list of the most significant features used during 100RR for the different disease classifications and scenarios. The list shows only the features used >50 times. The largest number of features were selected for the ALL vs ALL classification, while the fewest features were selected for CS vs PHT, several of which were exclusive to this classification (hsa-miR-495-3p, hsa-miR-485-3p, hsa-miR-186-5p, hsa-miR-1260a, hsa-miR-195-5p). Similarly, two separate sets of features (hsa-miR-139-5p, hsa-miR-326 and hsa-miR-223-3p, hsa-miR-133a-3p) were exclusively selected in PA vs PHT classification for Set D (the female subset) and Set E (elder patients) respectively.

Figure 4B shows the joint list of features in Set A of all five disease combinations. hsa-miR-15a-5p and hsa-miR-32-5p were common to all 5 disease classifications, while hsa-miR-424-5p and hsa-miR-574-3p were important features only for CS vs PHT and PA vs PHT, respectively. Age was observed to be a noteworthy feature for ALL vs ALL and PA vs PHT classification.

Final Model Training and Testing

In the final step of the ML pipeline, significant features of each disease combination were used to create a subset of the training dataset. Once trained, the chosen classifiers were tested on the testing set.

Table 2 shows the classification results for the top performing classifiers on test data. EHT vs PHT achieved the best classification (balanced accuracy: 89%, sensitivity: 95%, AUC: 0.9) using LMT classifier on 67 test samples (See Supplementary Figure S2 for confusion matrices). Similarly, for PPGL vs PHT, LB classifier provided 85% balanced accuracy with corresponding sensitivity and AUC of 87% and 0.9. However, performance metrics for ALL vs ALL, CS vs PHT and PA vs PHT were marginally lower, with balanced accuracy of 73%, 71% and 73% respectively.

DISCUSSION

In this study, a supervised ML pipeline was employed to classify and distinguish EHT and its subtypes (PA, PPGL, CS) from PHT using circulating miRNA data previously analysed in a multiomic context under the ENS@T-HT retrospective study.[25] In this singleomic analysis, the top discriminating miRNA biomarkers were identified for different classification scenarios and the classification performance was evaluated through overall classification balanced accuracy, specificity, sensitivity, AUC, F1 score and Kappa. The Boruta-based feature selection approach and exclusion of extreme outliers provided an overall better performance on miRNA data.

It was observed that circulating miRNA helped classify EHT and PPGL from PHT in the test set with balanced accuracy of 89% and 85%, respectively. Overall, hsa-miR-15a-5p and hsa-miR-32-5p were the most important biomarkers for classifying all disease combinations from each other. Also, two pairs of miRNAs (hsa-let7d-3p, hsa-miR-335-5p and hsa-miR-162-3p,

hsa-miR-15b-3p) were important in classifying all disease combinations except CS vs PHT and PPGL vs PHT. However, hsa-miR-424-5p, hsa-miR-495-3p, hsa-miR-186-5p, hsa-miR-1260a, hsa-miR-485-3p, hsa-miR-195-5p and hsa-miR-301a-3p were uniquely selected for CS vs PHT classification, while hsa-miR-629-5p, hsa-miR-92a-3p and hsa-miR-423-5p were used for both CS vs PHT and ALL vs ALL classification. Alongside other miRNAs, hsa-miR-30c-5p was one of the significant identifiers for ALL vs ALL and EHT vs PHT. Also, hsa-miR-148b-3p, hsa-miR-107, hsa-miR-27b-3p, hsa-miR-324-5p and hsa-miR-199a-5p were observed to be essential biomarkers for EHT vs PHT (selected for all 100 random repeats). Age appeared as a significant feature only in two classifications: ALL vs ALL and PA vs PHT. In the evaluation scenarios (Step 2 of the ML pipeline), it was also observed that female patients (Set D) and older patients (Set E) were better classified with higher balanced accuracies than male patients (Set C) and younger patients (Set F), except in the case of CS vs PHT where younger patients had higher classification accuracy than older patients. However, these trends are likely to be sensitive to the class imbalance, with more female CS patients than males.

The rationale for investigating circulating miRNA levels in the context of endocrine hypertension is twofold. Firstly, circulating miRNAs have shown promise in the diagnosis of several conditions where significant and consistent changes have proved sufficiently informative to justify their use for diagnostic purposes, if certain technological and methodological limitations can be overcome.[29,30] Secondly, we and others have demonstrated that miRNAs have significant roles in regulatory mechanisms that underpin endocrine homeostasis, such as corticosteroid secretion. Disruption of these contributes to PA and CS, with affected adrenal tissue or tumours showing marked changes in miRNA level relative to healthy tissue.[7] Although the mechanisms by which miRNAs enter the circulation remain unclear, it is reasonable to hypothesise that such changes at the tissue level might be reflected in the circulation, either as a result of altered miRNA expression within the

diseased tissue, or through its impact on the miRNA expression and secretion of other tissues.[31]

This study identifies several circulating miRNAs as potentially valuable discriminatory features in the diagnosis of endocrine hypertension (Figure 4A). Few of these have previously been highlighted by studies of these or similar conditions and this lack of correlation is probably due to several factors. Firstly, this work has focused on distinguishing various forms of endocrine hypertension from primary hypertension whereas previous studies attempted to discriminate between other combinations of diseased (and healthy) individuals,[32,33] or else concentrated on pathological or prognostic features of a single condition and/or its subtypes.[34–36] Methodological differences may also play a part, and it is probable that our investigation of free circulating miRNAs would have had different outcomes if we had chosen to examine serum or vesicle-associated miRNAs rather than plasma. Finally, the normalising of raw expression data together with the novel ML approach to identifying feature miRNAs must also be acknowledged as key differences between this and previous studies; these processes have been carefully designed to improve consistency of analysis and to aid its adoption in a multi-centre clinical setting as a common diagnostic tool.

Detailed investigation and description of possible tissue sources, genetic targets and pathophysiological roles of the feature miRNAs identified in this study lies beyond the scope of this particular investigation. However, it is worth noting that the key classifying biomarkers identified here have been flagged in previous studies of hypertension. For example, we observed a decrease in circulating levels of hsa-miR-15a-5p in primary hypertensive patients that was not seen in endocrine hypertensive patients. Variability in circulating levels of this miRNA with different forms of hypertension was previously noted by Nandakumar et al., who observed a rise in circulating hsa-miR-15a-5p in hypertensive patients relative to those with hypertension and chronic kidney disease.[33]

Previous studies examining adrenal effects on circulating miRNA levels have primarily been concerned with differentiating adrenocortical adenoma from adrenocortical carcinoma. These have consistently shown raised circulating levels of hsa-miR-483-5p to correlate with adrenocortical carcinoma[37] and, perhaps unsurprisingly, this does not emerge as one of our classifying features. However, it is interesting to note that one such study observed raised hsa-miR-210-3p levels in cortisol-producing adrenocortical carcinoma,[38] as this miRNA is one of the classifying features we identified for CS. We also found this miRNA to be one of the classifying features in our PPGL vs PHT comparison, consistent with previous studies associating hsa-miR-210-3p with PPGLs,[39] and identifying its serum levels as a potential biomarker of PPGL malignancy.[36] An investigation of phaeochromocytoma tissue identified 18 miRNAs as being differentially expressed between malignant and benign samples; in addition to the aforementioned hsa-miR-15a-5p,[40] these included hsa-miR-574-3p and hsa-miR-451a which we find to be classifiers of PPGL vs PHT comparison, but only within the female subset (Set D). A study by Vetrivel et al. of circulating miRNAs was primarily concerned with CS diagnosis and classification rather than its differentiation from primary and/or other endocrine forms of hypertension.[24] However, of the 8 miRNAs they identified by NGS as being differentially expressed between their 3 study groups (ACTH-dependent CS, ACTH-independent CS and non-CS controls), we have measured 3 and these include hsa-miR-629-5p, which we find to be one of our strongest classifiers of CS vs PHT. Vetrivel et al. were subsequently unable to validate the differential expression of circulating hsa-miR-629-5p between the two forms of CS using RT-PCR, but it should be noted that their protocol involved normalisation to a single 'housekeeping' circulating miRNA, hsa-miR-16-5p, that we find to be a significant classifier for PA vs PHT. The same investigators also used NGS to identify possibly one of our most discriminating miRNA (hsa-miR-486-5p) for CS vs PHT, as one of 5 miRNAs commonly downregulated in CS adrenal tissue relative to control but again they were unable to validate their NGS data using RT-PCR.[23] Finally, we previously identified

significant differences in specific miRNA levels between APAs and normal adrenal tissue,[6,7] but none of these emerged here as classifying features of PA in their circulating form, except for hsa-miR-22-3p (only in subset C: male).

Balanced accuracy and AUC were used as the primary metrics to assess performance of the ML models throughout, as they are unaffected by disease prevalence, unlike standard or conventional accuracy metrics. Nevertheless, the estimates derived using these performance metrics will inevitably be more imprecise where sample size is small. Indeed, although this study provided crucial insight into a complex multi-class problem, one key shortcoming was fewer CS samples due to the rarity of the disease. The ENS@T-HT study is capturing data prospectively from a larger population which should overcome this limitation in the future (ClinicalTrials.gov Identifier: NCT02772315).

CONCLUSION

This study predicted different subtypes of hypertension using circulating miRNA data. A ML approach using 5 disease combinations and 8 supervised ML classifiers was introduced and different scenarios were evaluated based on age and sex bifurcation. The ML approach provided promising classification results and a reduced set of features that have potential as formal biomarkers for detection of hypertension subtypes. Confirmation of our findings using the same methodology in a different study population is now required. To this end, we are currently conducting a separate prospective study of circulating miRNA levels in EHT patients specifically to test the reproducibility of our findings and confirm the diagnostic utility of these circulating miRNAs as accessible high-throughput biomarkers.

ACKNOWLEDGEMENTS

Data related to the results presented in this article can be obtained upon reasonable request by researchers who provide a methodologically sound proposal. This is subject to approval by the ENSAT-HT executive committee, which includes representatives of the biosamples collections and omics data generation centres. Researchers would be required to complete a Data Sharing Agreement. All requests should be directed by email to the corresponding author.

SOURCES OF FUNDING

This project was funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 633983. This work was supported by the Clinical Research Priority Program of the University of Zurich for the CRPP HYRENE and by the Deutsche Forschungsgemeinschaft (DFG) within the CRC/Transregio 205/1 "The Adrenal: Central Relay in Health and Disease" to F.B. C.P is funded by Deutsche Forschungsgemeinschaft (314061271-TRR/CRC 205- 1/2). S.L is a PhD student funded by the MRC Precision Medicine DTP programme (MR/N013166/1). For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

DISCLOSURES

Co-authors Guillaume Assie and Felix Beuschlein are on the editorial board of EJE. They were not involved in the review or editorial process for this paper, on which they are listed as authors.

REFERENCES

1. Mills, K. T. *et al.* Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries. *Circulation* **134**, 441–450 (2016).
2. Rahimi, K., Emdin, C. A. & MacMahon, S. The Epidemiology of Blood Pressure and Its Worldwide Management. *Circulation Research* **116**, 925–936 (2015).
3. Williams, B. *et al.* 2018 ESC/ESH Guidelines for the management of arterial hypertension. *Eur. Heart J.* **39**, 3021–3104 (2018).
4. Dosh, S. A. The diagnosis of essential and secondary hypertension in adults. *J Fam Pract* **50**, 707–712 (2001).
5. Touyz, R. M. & Burger, D. Biomarkers in Hypertension. in *Special Issues in Hypertension* 237–246 (Springer, Milano, 2012). doi:10.1007/978-88-470-2601-8_19.
6. Robertson, S. *et al.* Regulation of Corticosteroidogenic Genes by MicroRNAs. *International Journal of Endocrinology* **2017**, 2021903 (2017).
7. Robertson, S. *et al.* MicroRNA-24 is a novel regulator of aldosterone and cortisol production in the human adrenal cortex. *Hypertension* **62**, 572–578 (2013).
8. Li, W. Y. *et al.* Circulating microRNAs as potential non-invasive biomarkers for the early detection of hypertension-related stroke. *Journal of Human Hypertension* **28**, (2014).
9. Li, M. & Zhang, J. Circulating MicroRNAs: Potential and Emerging Biomarkers for Diagnosis of Cardiovascular and Cerebrovascular Diseases. *BioMed Research International* https://www.hindawi.com/journals/bmri/2015/730535/ (2015) doi:10.1155/2015/730535.
10. Sekar, D., Shilpa, B. R. & Das, A. J. Relevance of microRNA 21 in Different Types of Hypertension. *Curr Hypertens Rep* **19**, 57 (2017).
11. Krishnan, R., Mani, P., Sivakumar, P., Gopinath, V. & Sekar, D. Expression and methylation of circulating microRNA-510 in essential hypertension. *Hypertens. Res.* **40**, 361–363 (2017).

12. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* **49**, 107739 (2021).
13. Marceddu, G. *et al.* Analysis of machine learning algorithms as integrative tools for validation of next generation sequencing data. *Eur Rev Med Pharmacol Sci* **23**, 8139–8147 (2019).
14. Njage, P. M. K. *et al.* Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. *Risk Analysis* **39**, 1397–1413 (2019).
15. Chen, X., Gong, Y., Zhang, D.-H., You, Z.-H. & Li, Z.-W. DRMDA: deep representations-based miRNA–disease association prediction. *J. Cell. Mol. Med.* n/a-n/a (2017) doi:10.1111/jcmm.13336.
16. Pallez, D., Gardès, J. & Pasquier, C. Prediction of miRNA-disease Associations using an Evolutionary Tuned Latent Semantic Analysis. *Scientific Reports* **7**, 10548 (2017).
17. Tai, M. C. *et al.* Blood-borne miRNA profile-based diagnostic classifier for lung adenocarcinoma. *Sci Rep* **6**, (2016).
18. Hanniford, D. *et al.* A miRNA-based signature detected in primary melanoma tissue predicts development of brain metastasis. *Clin Cancer Res* **21**, 4903–4912 (2015).
19. Søkilde, R. *et al.* Efficient Identification of miRNAs for Classification of Tumor Origin. *The Journal of Molecular Diagnostics* **16**, 106–115 (2014).
20. Patterson, E. *et al.* The microRNA expression changes associated with malignancy and SDHB mutation in pheochromocytoma. *Endocr Relat Cancer* **19**, 157–166 (2012).
21. Li, Z., Yu, X. & Shen, J. The role of miRNAs in the pheochromocytomas. *Tumor Biol.* **37**, 4235–4239 (2016).
22. Castro-Vega, L. J., Lepoutre-Lussey, C., Gimenez-Roqueplo, A.-P. & Favier, J. Rethinking pheochromocytomas and paragangliomas from a genomic perspective. *Oncogene* **35**, 1080–1089 (2016).

23. Vetrivel, S. *et al.* Characterization of Adrenal miRNA-Based Dysregulations in Cushing's Syndrome. *Int J Mol Sci* **23**, 7676 (2022).
24. Vetrivel, S. *et al.* Circulating microRNA Expression in Cushing's Syndrome. *Frontiers in Endocrinology* **12**, (2021).
25. Reel, P. S. *et al.* Machine learning for classification of hypertension subtypes using multi-omics: A multi-centre, retrospective, data-driven study. *eBioMedicine* **84**, (2022).
26. Gholamy, A., Kreinovich, V. & Kosheleva, O. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. https://scholarworks.utep.edu/cs_techrep/1209 (2018).
27. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250 (2004).
28. Nind, T. *et al.* The Research Data Management Platform (RDMP): A novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data. *GigaScience* (2018) doi:10.1093/gigascience/giy060.
29. Szelenberger, R., Kacprzak, M., Saluk-Bijak, J., Zielinska, M. & Bijak, M. Plasma MicroRNA as a novel diagnostic. *Clinica Chimica Acta* **499**, 98–107 (2019).
30. MacKenzie, S. M. *et al.* Circulating microRNAs as Diagnostic Markers in Primary Aldosteronism. *Cancers* **13**, 5312 (2021).
31. Cui, C. & Cui, Q. The relationship of human tissue microRNAs with those from body fluids. *Scientific Reports* **10**, 5644 (2020).
32. Belaya, Z. *et al.* Circulating Plasma microRNA to Differentiate Cushing's Disease From Ectopic ACTH Syndrome. *Front. Endocrinol.* **11**, (2020).
33. Nandakumar, P. *et al.* MicroRNAs in the miR-17 and miR-15 families are downregulated in chronic kidney disease with hypertension. *PLOS ONE* **12**, e0176734 (2017).

34. Calsina, B. *et al.* Integrative multi-omics analysis identifies a prognostic miRNA signature and a targetable miR-21-3p/TSC2/mTOR axis in metastatic pheochromocytoma/paraganglioma. *Theranostics* **9**, 4946–4958 (2019).
35. Decmann, A. *et al.* Circulating miRNA Expression Profiling in Primary Aldosteronism. *Front Endocrinol (Lausanne)* **10**, (2019).
36. Ruff, S. M. *et al.* MicroRNA-210 May Be a Preoperative Biomarker of Malignant Pheochromocytomas and Paragangliomas. *J Surg Res* **243**, 1–7 (2019).
37. Decmann, A., Perge, P., Turai, P. I., Patócs, A. & Igaz, P. Non-Coding RNAs in Adrenocortical Cancer: From Pathogenesis to Diagnosis. *Cancers (Basel)* **12**, (2020).
38. Perge, P. *et al.* Evaluation and diagnostic potential of circulating extracellular vesicle-associated microRNAs in adrenocortical tumors. *Sci Rep* **7**, (2017).
39. Tsang, V. H. M. *et al.* Overexpression of miR-210 is associated with SDH-related pheochromocytomas, paragangliomas, and gastrointestinal stromal tumours. *Endocr Relat Cancer* **21**, 415–426 (2014).
40. Meyer-Rochow, G. Y. *et al.* MicroRNA profiling of benign and malignant pheochromocytomas identifies novel diagnostic and therapeutic targets. *Endocr Relat Cancer* **17**, 835–846 (2010).

FIGURE LEGENDS

Figure 1: The ML analysis pipeline with three steps. [A]All features selected, filter method (CFS: correlation-based feature selection) and wrapper method (Boruta). [B]J48, Naïve Bayes (NB), IBk, Logitboost (LB), Logistic Model Tree (LMT), Simple Logistic (SL), Random Forest (RF) and Sequential minimal optimization (SMO). [C]PA vs PPGL vs CS vs PHT (i.e. All vs All), EHT vs PHT, PA vs PHT, CS vs PHT and PPGL vs PHT. [D]Scenario 1:Set A (all features inc. age & sex) vs Set B (all features exc. age & sex),Scenario 2: Set C (Male patients) vs Set D (Female patients) and Scenario 3: Set E (Patient age >= 50) vs Set F (Patient age < 50).

Figure 2: Box plots comparing classification performance for ALL vs ALL disease combination (A) either excluding or including extreme outliers and (B) excluding outliers dataset for the top 4 classifiers using all features, filter and wrapper feature selection methods.

Figure 3: Heatmap comparing classification performance for Sets A-F using 4 classifiers for 5 disease combinations. The count in each box is a mean of 100 runs (random repeats).

Figure 4: Significant features selected during (A) 100RR for the different disease classifications and scenarios, and (B) final classifier training and testing. In Part A, the count in each box represents the number of times a particular feature was selected during 100RR and only features used >50 times are selected.

TABLE LEGENDS

Table 1: Patient data for all disease types namely Cushing's syndrome (CS), primary aldosteronism (PA), pheochromocytoma or paraganglioma (PPGL) and primary hypertension (PHT).

Table 2: Classification results on test set for five disease comparisons of top performing classifiers.

Disease	Patient Count (n=)	Gender		Age distribution	
		Male	Female	Patient age >= 50	Patient age < 50
CS	35	5	30	17	18
PA	109	58	51	44	65
PPGL	75	31	44	43	32
PHT	111	48	63	71	40

Disease Combination	Classifier	Performance Metrics (Test set)								
		Total Instances	Correctly Classified	Incorrectly Classified	Balanced Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	F1	Kappa
ALL – ALL	LMT	67	45	22	73	58	88	0.8	0.6	0.5
EHT – PHT	LMT	67	61	6	89	95	83	0.9	0.9	0.8
CS – PHT	LB	30	26	4	71	43	100	0.9	0.6	0.5
PA – PHT	LMT	45	33	12	73	64	83	0.8	0.7	0.5
PPGL – PHT	LB	38	32	6	85	87	83	0.9	0.8	0.7

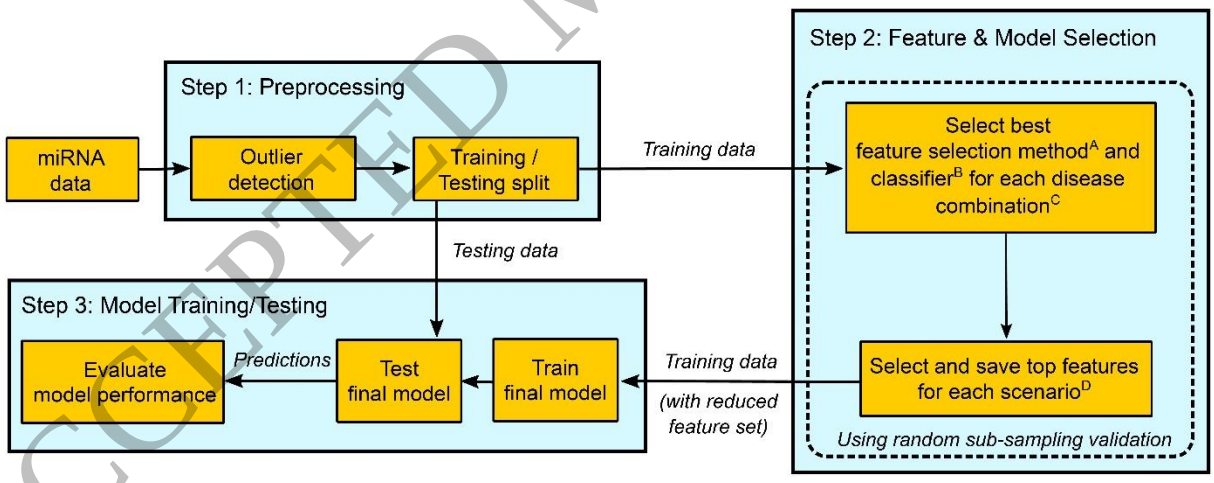

Figure 1

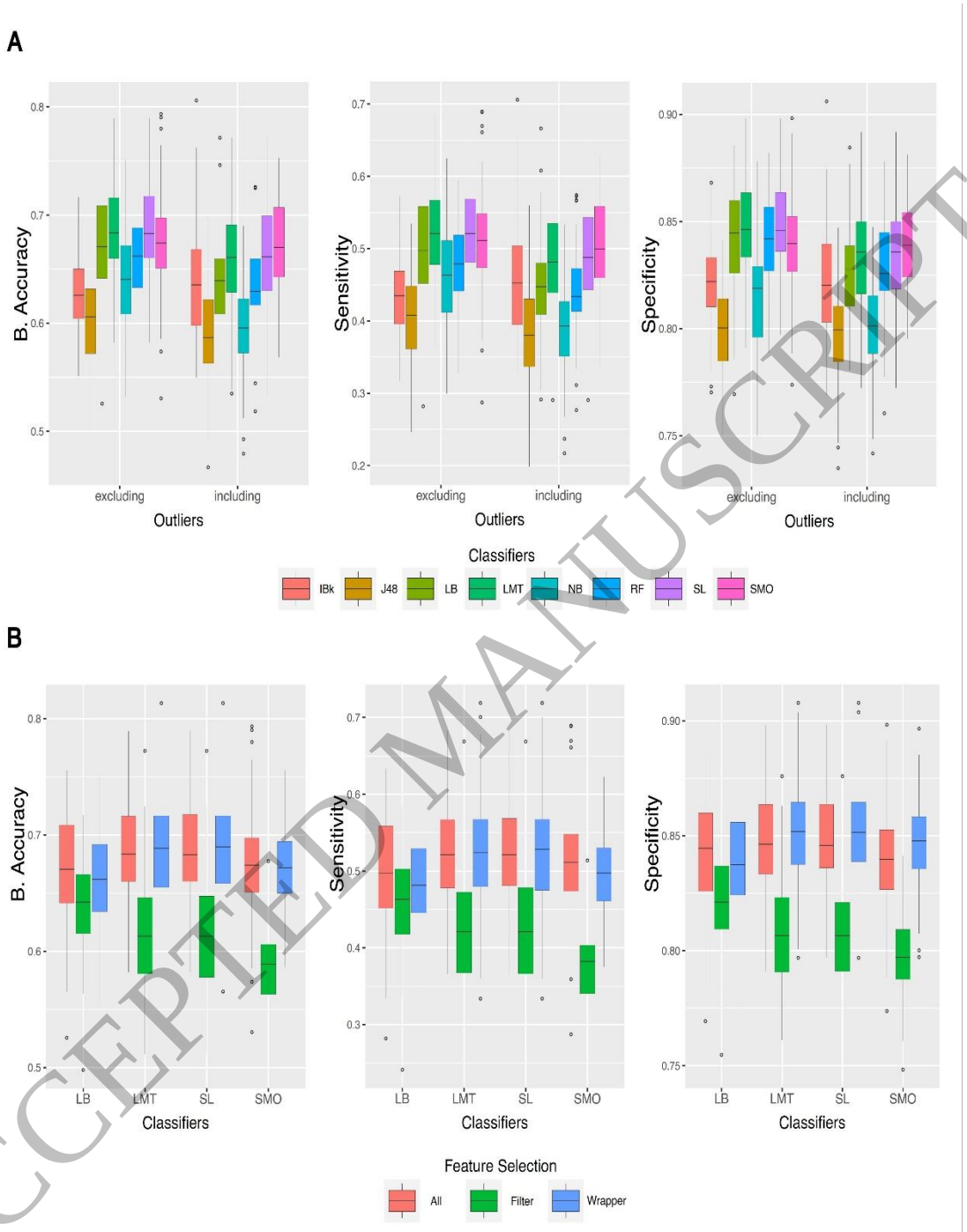

Figure 2
513x545 mm (x DPI)

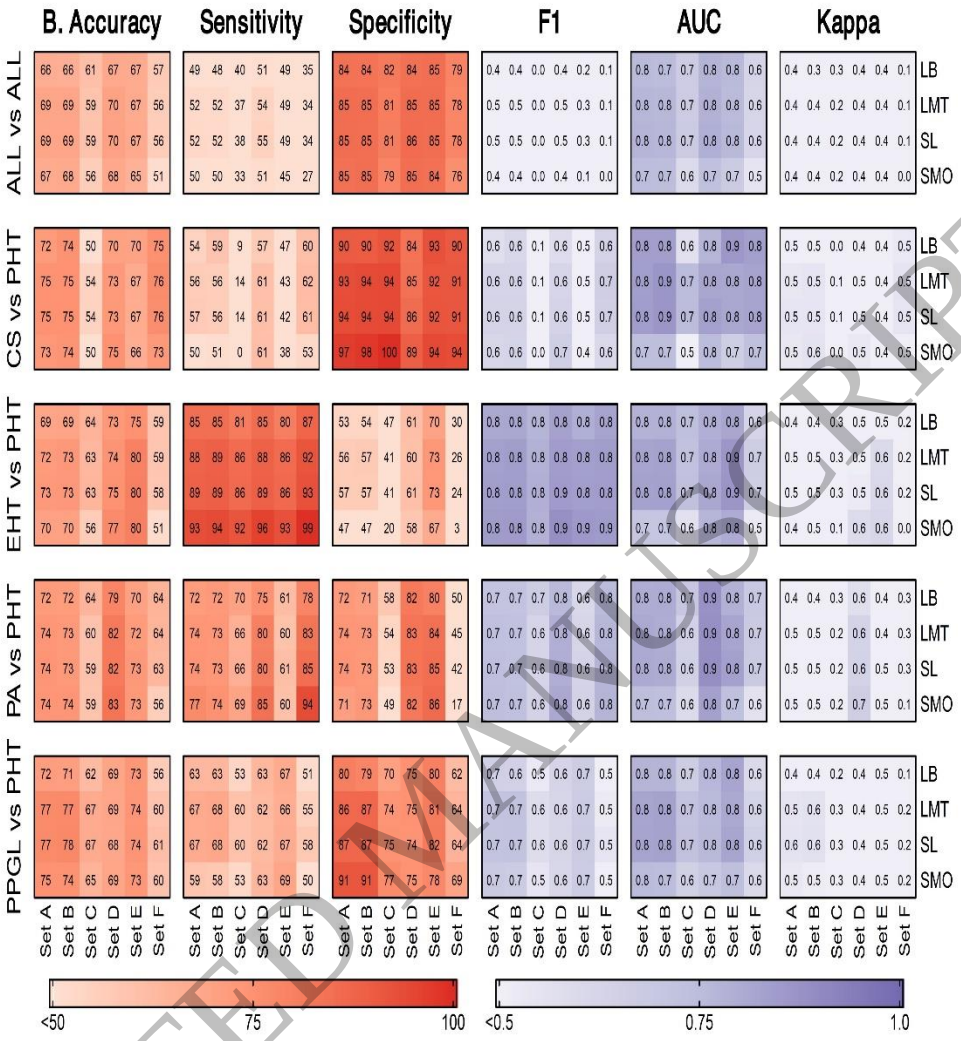

Figure 3
518x366 mm (x DPI)

Figure 4
401x559 mm (x DPI)