University of Glasgow

MSc Bioinformatics

Evaluating Multi-Class Classification Metrics for microRNA Data of Hypertension Subtypes

Student: Mark Gallacher (2466057)

Supervisor: Dr John D. McClure

August 2024

A report submitted in partial fulfilment of the requirements for the

MSc Bioinformatics Degree at The University of Glasgow.

Summary

Background

High blood pressure (hypertension) is a key component of cardiovascular and cerebrovascular diseases with an estimated burden of 211.8 million disability adjusted life-years (DALYs). Many forms of hypertension exist, including endocrine hypertension subtypes which are rarely screened due to economical factors or invasive procedures. MicroRNA are likely implicated in hypertension and could offer a cost-effective high-throughput technique to diagnose the subtypes. Machine learning has been successfully applied to improve diagnosis in other diseases, due to the growing capabilities especially in handling high-dimensional data. Therefore, the diagnostic potential of machine learning with microRNA data is worth investigating.

Aims

Given the number of machine learning algorithms, effectively evaluating the models and selecting suitable hyperparameters is critical. Typically models tend to favour the largest class which can result in considerably poorer performance in the smaller class. In the context of multiclass classification, when more than two classes exist, these biases can be exaggerated. This project evaluated multiple metrics for imbalance data by training six types of models. Averaging techniques (Micro, Macro and Weighted) used in many metrics to extend to multiclass classification are also investigated. Additionally, a Dummy (naive) classifier is used to highlight the underlying tendencies of the metrics.

Results

Micro-averaged metrics were identical to Accuracy, limiting information for model diagnostics and overestimating model performance. Generally, Gradient Boosted Trees

was the better model compared to Random Forest, Support Vector Machines and Logistic Regression. K-Nearest Neighbours and Naive Bayes were markedly worse in handling the imbalance data. After selecting the top ten sets of hyperparameters for each metric, across all types of models, the class-specific metrics are compared to their averaged versions to evaluate the form of aggregation. Here, Micro- and Weighted-averaged metrics appeared insensitive to weak outcomes on the less frequent labels unlike Macro-averaged metrics. To estimate the consequences of using a single metric, a collection of “best” models was obtained by extracting the best score for each metric. These models are compared via their Recall of individual classes. Class-agnostic metrics preferred models with strong performances in the majority class, even if the minority class had inadequate scores. Whereas Balanced Accuracy selected a model with slightly worse performance in the largest class, but with better scores for the smaller classes.

Conclusion

The choice of metric seems to disproportionately affect the minority class(es), suggesting the importance of these classes is essential to selecting the most appropriate metric. Micro-averaging and Weighted-averaging often provided no additional information compared to Accuracy. Whilst the strictness of MCC and Cohen’s Kappa are appropriate in the medical setting, the interpretation of their values is not obvious. Balanced Accuracy (Macro-Recall) and Macro-averaged appeared to handle the imbalanced data better and seem the most appropriate choice.

Acknowledgements

This body of work would not be completed without the encouragement and wisdom of my supervisor. I cannot thank Dr John McClure enough for all his time and guidance he has given me throughout this project.

Table of Contents

Abbreviations

PHT - Primary (Essential) Hypertension

PA - Primary Aldosteronism

PPGL - Pheochromocytoma/catecholamine-producing Paraganglioma

CS - Cushing Syndrome

RF - Random Forest

GB - Gradient Boosted Trees

SVM - Support Vector Machine

KNN - K-Nearest Neighbours

LG - Logistic Regression

GNB - Gaussian Naive Bayes

MCC - Matthews Correlation Coefficient

TP - True Positive

FP - False Positive

TN - True Negative

FN - False Negative

TPR - True Positive Rate

Introduction

Hypertension (high systemic blood pressure) impacts a significant proportion of the adult population and carries an estimated global burden of 211.8 million disability-adjusted life-years (DALYs)(GBD 2015 Risk Factors Collaborators, 2016; Mills, Stefanescu and He, 2020). Furthermore, it is the largest modifiable factor of multiple common diseases, like myocardial infarction, ischemic stroke, coronary heart disease and chronic kidney disease (Oparil *et al.*, 2018; Koch, Papadopoulou-Marketou and Chrousos, 2020). With a wide range of aetiologies, hypertension is considerably heterogeneous and some forms can prove resistant to standard antihypertensive interventions (Sudano, Beuschlein and Lüscher, 2018; Williams *et al.*, 2018). Around 90-95% of hypertensive patients are categorised with primary (essential) hypertension (PHT), where no single pathophysiology is obvious, likely stemming from a combination of lifestyle, dietary and environmental factors (Staessen *et al.*, 2003; Sudano, Beuschlein and Lüscher, 2018). Conversely, secondary hypertension is categorised by an apparent pathophysiology into two main groups, renal and endocrine (Koch, Papadopoulou-Marketou and Chrousos, 2020; Sudano, Suter and Beuschlein, 2023). Secondary hypertension can have specific therapeutic interventions that may correct the underlying condition. Therefore, it is of significant importance to be able to effectively diagnose and distinguish between the types of hypertension.

Whilst identifying high blood pressure with a sphygmomanometer is convenient and offers reliable results, the various forms of hypertension are not as easily determined (Koch, Papadopoulou-Marketou and Chrousos, 2020). Moreover, endocrine hypertension, characterised by hormonal imbalances, can be further divided into specific subtypes which equally require to be differentiated from each other. The most common subtypes present with excessive mineralocorticoids (primary aldosteronism,

PA), glucocorticoids (Cushing syndrome, CS) or catecholamines (pheochromocytoma/catecholamine- producing paraganglioma, PPGL) or the imbalance of thyroid, seen in hyper- and hypo-thyroidism (Williams *et al.*, 2018; Koch, Papadopoulou-Marketou and Chrousos, 2020). If undetected at an early stage, chronic secondary hypertension can result in cardiovascular restructuring that limits the effectiveness of medication (Williams *et al.*, 2018). Unfortunately, the required screening process to diagnose specific endocrine hypertension is often costly, time-consuming and invasive.

Circulating microRNA may offer a potential cost-effective solution. MicroRNA are small, 22-26 nucleotide fragments reported to have inhibitory effects on transcription and can be found in serum, urine and plasma (Parveen *et al.*, 2019). Moreover, microRNA fragments have been observed influencing multiple biological processes, including functions relevant to cardiovascular physiology, such as cardiomyocyte proliferation, cardiac conductivity and cardiac hypertrophy (Shi *et al.*, 2015; Zhang *et al.*, 2018). In addition, microRNA have also been considered as convenient biomarkers for the early detection of cancers (Azari *et al.*, 2023), cardiovascular (Baghdadi *et al.*, 2023) and cerebrovascular diseases (Li *et al.*, 2014), including predicting myocardial infarction (Samadishadlou *et al.*, 2023). Specifically, some microRNA fragments were found to correlate with cardiac troponin (CtnT), a clinical marker for acute myocardial infarction (Song *et al.*, 2015). Furthermore, microRNA appear involved in vascular homeostasis, through regulating vasodilation during shear stress by increasing nitric oxide levels (Weber *et al.*, 2010), and the regulation of genes associated with corticosteroid synthesis (Robertson *et al.*, 2017). Overall, microRNA plays a notable role in a diverse range of biological processes and may be a pragmatic option to differentiate the subtypes of endocrine hypertension.

With the growing capabilities of machine learning, especially in deep learning, efforts have been made to apply machine learning for medical diagnosis (Yu, Beam and Kohane, 2018; Topol, 2019; Ahsan, Luna and Siddique, 2022). These developments complement the large quantity of data obtainable with next generation sequencing (NGS). One advantage of NGS is the quantification of microRNA, which can supply machine learning models data to help detect lung adenocarcinoma (Tai *et al.*, 2016) or predict brain metastases (Hanniford *et al.*, 2015). Baghdadi and colleagues (2023) implemented an ensemble method that achieved an F1-score of 92.3% in detecting the early stages of cardiovascular diseases. Furthermore, a multi-omics approach used serum microRNA and metabolomics to employ machine learning to effectively differentiate subtypes of hypertension with a specificity of 96% (Reel *et al.*, 2022). Using microRNA with modern classifiers may become a convenient method to improve diagnoses of many conditions, including endocrine hypertension.

However, imbalanced datasets are problematic in machine learning as algorithms tend to favour the largest group in a dataset, to the point where a minority class may be ignored (Galar *et al.*, 2012; Gaudreault, Branco and Gama, 2021). Within the medical setting, this can have serious implications as the prevalence of diseases are typically nonuniform (Williams *et al.*, 2018; Araf, Idri and Chairi, 2024). Imbalanced datasets can generate bias models which may obtain overoptimistic metric scores (Blagus and Lusa, 2010; Gaudreault, Branco and Gama, 2021).

Additionally, there are multiple state-of-the-art algorithms available for a given problem, which makes it more difficult to select the best model (Rainio, Teuho and Klén, 2024). Machine learning algorithms often have tunable hyperparameters that drastically impact performance and, like model selection, require some chosen metric(s) to find the optimal choice. A metric is a statistic that aims to estimate the performance and generalisability of a model to enable model selection and hyperparameter tuning, but

they are not without biases (Cawley and Talbot, 2010). The selected metric is important to ensure this comparison is robust and the chosen model has the optimal performance (Demšar, 2006).

One of the most common metrics to gauge performance of a classifier is Accuracy; the total number of correct labels divided by the total number of labels. For example, two recent microRNA binding sites predictors, MBSTAR and TarPmiR, achieved high Accuracy whilst also obtaining very poor scores in other metrics. TarPmiR celebrated an Accuracy of above 80% but achieved a Precision less than 10% (Ding, Li and Hu, 2016), whilst MBSTAR achieved an Accuracy of 72.02% with an F1-score of 34% (Bandyopadhyay *et al.*, 2015). Moreover, Brown (2018) illustrates how metrics, like Accuracy, are more likely to overestimate performance when the data becomes more imbalanced. Stehman and Foody (2019) highlighted that Accuracy does not inappropriately weigh the classes, it simply is class-agnostic and offers no information about them. Whereas, a class-sensitive metric could preserve the importance of the rare labels. Therefore, the choice of metric can strongly influence how capable the model's predictions are, especially for the less frequent labels.

Furthermore, within multiclass classification problems (Number of Classes > 2), the issues with imbalanced datasets may be worsened as multiple minority classes can be present. Some metrics do not naturally extend to the multiclass setting, such as Recall, Precision and F1 scores (Grandini, Bagli and Visani, 2020). For these metrics, an average score is required so that we return a single value. There are two main modes of averaging scores, Macro and Micro. Macro-level metrics calculate a given metric for each class, then take the average. Consequently, each class has equal significance on the final value, regardless of relative size (Grandini, Bagli and Visani, 2020; Rainio, Teuho and Klén, 2024). Micro-averaging sums the confusion matrix, for example, sums all True Positive values, then uses these totals for the metric calculation.

Micro-averaging provides more weight to the more frequent labels and has identical properties to Accuracy (Farhadpour, Warner and Maxwell, 2024). There is no consensus which metric offers the most reliable estimation of model performance, especially for imbalanced multiclass problems. Additionally, reliable model evaluation is likely to become increasingly important with the growing capabilities and use of artificial intelligence potentially overtaking trained medical professionals (Rainio, Teuho and Klén, 2024).

This project explores the behaviour of multiple metrics and evaluates the averaging strategies across a range of popular machine learning classifiers to judge the effectiveness of diagnosing the subtypes of endocrine hypertension. Two main groups of models were trained; linear models consisting of Logistic Regression and Support Vector Machines, and ensemble methods such as Random Forest and Gradient Boosted Trees. Additionally, the probability-based Gaussian Naive Bayes and the proximity-based K-Nearest Neighbours were also included.

Methods

The Data

Originally sourced from the 2022 multi-omics study on endocrine hypertension by Reel and colleagues (2022), the data consists of 173 plasma microRNA fragments and their estimated expression as determined by quantitative polymerase chain reaction (qPCR) from the Quantstudio 12K Flex Real-time qPCR System. The cohort consists of 465 patients with 133 healthy volunteers (HV), 111 for primary hypertension (PHT), 109 for primary aldosteronism (PA), 76 for PPGL and 36 with Cushing syndrome (CS). The HV group was not used as this project aimed to differentiate between subtypes of hypertension. To obtain normalised delta cycle thresholds (delta-Ct) from the qPCR results, the average of five consistent expression levels were used. This provides a reliable relative quantification of expression that is widely used to process real-time qPCR data (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008). Patient demographics, such as gender and age, were also excluded.

The Metrics

The metrics used include Accuracy, Balanced Accuracy, Precision, Recall, F1-Score, Cohen's Kappa and Matthews Correlation Coefficient. Most metrics are based on the confusion matrix, which is *N x N* in size, where *N* is the number of classes. For binary classification (*N* = 2) the matrix has the form:

	Positive	Negative
Predicted Positive	*True Positive (TP)*	*False Positive (FP)*
Predicted Negative	*False Negative (FN)*	*True Negative (TN)*

Accuracy is the sum of the correct labels divided by the total size of the dataset, therefore it is class-agnostic and often too simplistic for multiclass problems. Whereas, Balanced Accuracy is the mean of the accuracy for each class. For brevity, only the formula for the binary classification shown below.

$$Accuracy = \frac{Number\ of\ Correct\ Labels}{Total\ Number\ of\ Labels}$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Recall (also known as Sensitivity or True Positive Rate (TPR)) and Precision (often termed Positive Predictive Value (PPV)) are the model's ability to identify the positive class and the probability a positive label is actually correct, respectively. This pair of metrics is connected, as it is possible to achieve a very high Recall by sacrificing Precision, and vice-versa. To counter this, the F-beta score combines both metrics into a single value. However, beta is assigned the value 1 in the vast majority of cases. Consequently, achieving a high F-beta score requires both Precision and Recall to be high, preventing the problematic asymmetry.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F_{\beta}\ Score = (1 + \beta^2)\frac{Precision\ .\ Recall}{(\beta^2.\ Precision) + Recall}$$

$$F_1\ Score = 2\ .\frac{Precision\ .\ Recall}{Precision + Recall}$$

In the multiclass setting, the metrics Precision, Recall and consequently F1-Score require a form of aggregation to result in a single value. Three forms of averaging were explored, Macro, Micro and Weighted. Macro-averaging obtains the metric for each class, then calculates the mean while Micro-averaging collects the values from the confusion matrix before passing them to the formula. Weighted is the weighted sum of Macro-averaging, where the weights are proportional to the frequency of the labels.

Cohen's Kappa and Matthews Correlation Coefficient (MCC) are similar metrics, as they both aim to indicate the argument between the set of true labels and predicted labels, and range from -1 to +1. However, negative values are generally only achieved if the labels are misconfigured during the training, therefore their practical range is 0 to +1. Additionally, they use all four values in the 2 x 2 confusion matrix which renders them more strict and robust metrics. Whilst they extend to the multiclass setting, their binary formulas are shown for brevity.

$$Cohen's\ Kappa\ (\kappa) = \frac{2\ .\ (TP\ .\ TN - FN\ .\ FP)}{(TP+FP)(FP+TN)(TP+FN)(FN+TN)}$$

$$MCC = \frac{TP\ .\ TN - FN\ .\ FP}{\sqrt{(TP+FP)(FP+TN)(TP+FN)(FN+TN)}}$$

In the results, the metrics are grouped by their averaging method, (Macro, Micro and Weighted) with an additional group "None" to capture metrics like Accuracy and MCC, which naturally translate to the multiclass setting. It is worth noting this grouping is solely for presentation.

The Models

In order to evaluate the selected metrics for the imbalanced dataset, they were initially compared using a Dummy (sometimes termed Naive) classifier which always guessed the largest class. For metrics that achieved larger values in the Dummy model, this would indicate the metric is more sensitive to the size of the majority class. Lower values for this Dummy model would suggest a more reliable metric, especially as multiple classes were ignored. **Appendix C** illustrates the difference in scores between the genuine models and the Dummy classifier.

Secondly, the metrics were compared across six models, where each were trained over a wide range of hyperparameters. These algorithms include Logistic Regression (LG, Defazio, Bach and Lacoste-Julien, 2014), Gaussian Naive Bayes (GNB, Zhang, 2004), k-Nearest Neighbours (KNN), Support Vector Machine (SVM, Chang and Lin, 2011), Random Forest (RF, Breiman, 2001) and Gradient Boosted Trees (GB, Friedman, 2001). A total of 4,885 models were trained, with 2,000 GB, 2,000 RF, 462 SVM, 253 LG, 126 LSVM, 44 KNN and 1 GNB. For full details on the hyperparameter chosen of these models and the model configurations, see **Appendix A** and **B,** respectively.

LG fits a sigmoid curve which has horizontal asymptotes at 0 and 1, denoting the negative and positive class, respectively. SVM finds the optimal hyperplane which differentiates two classes. To generate a non-linear decision boundary, a SVM uses a kernel function to transform the data points into a multi-dimensional space. Typically kernels include a polynomial or radial basis function (RBF) which allows a non-linear boundary for more complex problems. Two implementations of SVM were used, one specialised to a linear kernel (LSVM) and one capable of different kernels (SVM). GNB employs Bayes' rule with the assumption the features are independent of the classes

to estimate the probability a sample belongs to each class. LG and SVM fit multiple one-versus-rest models to extend to the multiclass setting, where the maximum value from the decision function determines the assigned label. Likewise, GNB obtains the probability the sample belongs to each class and selects the maximum value. KNN defines a boundary to separate the different classes by looking at a *k* number of neighbouring points, with larger *k* values generating a smoother line. RF and GB are both ensemble methods which train a collection of decision trees in parallel and sequentially, respectively. RF employs bootstrapping of the samples to ensure enough diversity in the trees to generate different errors that cancel out when combining the output of all the trees. GB does not use bootstrapping but iteratively utilises the errors of the previous tree to influence the training of the next tree.

Through a 5-fold cross validation with stratified sampling, each model generates 5 values for each metric to get a more reliable estimation of performance compared to a single test set. Stratified sampling ensures the proportions of the classes are identical between the 5 folds, keeping the class imbalance consistent across the folds. The Pearson correlation between the metrics are presented in **Appendix D**. Additionally, the confusion matrix for each model was obtained to investigate the performance of individual groups to illustrate what subtype the model may struggle to classify. Furthermore, the class-specific metrics are compared to the aggregated metrics (see **Figure 3** and **Appendix E**).

Given some models have more hyperparameters to tune, this generates more unique combinations. To prevent the bias towards the more numerous models, only the top 10 values for each metric and for each type of model were used. Finally, the “best” set of hyperparameters for each model are selected by the maximum value achieved for each single metric. Then the TPR (or Recall) for each class was compared between

these selected models to illustrate the potential down-stream consequences of using that metric in model selection (see **Figure 4** and **Appendix F**).

The training of the models used Python (version 3.6.8) with imported modules such as NumPy (version 1.19.5), Pandas (version 1.1.5) and Sci-Kit Learn (version 0.24.2). Graphical representations and data parsing was completed through R (version 4.3.2) and the TidyVerse ecosystem (version 2.0.0). (See the “Code” Folder in Appendices and the README for full details)

Results

Evaluation of a Dummy Classifier

To illustrate the behaviour of the metrics, the Dummy (Naive) model was configured to always predict the most frequent class (PHT). Whilst the absolute values are not impressive, there are substantial differences between the metrics. From **Figure 1**, is it clear that accuracy was the most (over-)confident metric compared to Balanced accuracy, MCC and Cohen's Kappa. The difference behind Accuracy and Balanced Accuracy is also apparent, as Accuracy returns the proportion of the largest class (0.334) while Balanced Accuracy returns one divided by the number of classes (0.25). With more classes or a greater imbalance, the difference would become even more clear. As expected, the metrics that used Micro averaging were identical to Accuracy meaning the diagnostic information about the difference between Recall and Precision is unobtainable. Whereas, Macro averaging was more conservative and indicated that the model had better Recall than Precision. MCC and Cohen's Kappa correctly evaluated the usefulness of this model reflected by their value of zero.

Evaluation of Genuine Classifiers

To reflect a more realistic approach to model selection, the top ten models for each metric were extracted. **Figure 2** shows the median scores of different metrics across the averaging methods for each of the models. Overall, there was agreement between the different metrics across all the averaging methods that clearly indicated Gradient Boosted Trees (GB) as the best performing model. Random Forests (RF), Logistic Regression (LR) and Support Vector Machines (SVM) generally performed very similarly to each other as the next best set of models. Similar to the previous figure, Accuracy appeared the most optimistic with MCC and Coehn's Kappa gave markedly lower values. Moreover, MCC and Cohen's Kappa are very similar in both their values

and ranking of the models. However, only Balanced Accuracy and the Macro averaged metrics do not put RF as the second best model, likely suggesting RF is struggling with a class that is being overlooked by the class-agnostic and Micro averaged metrics.

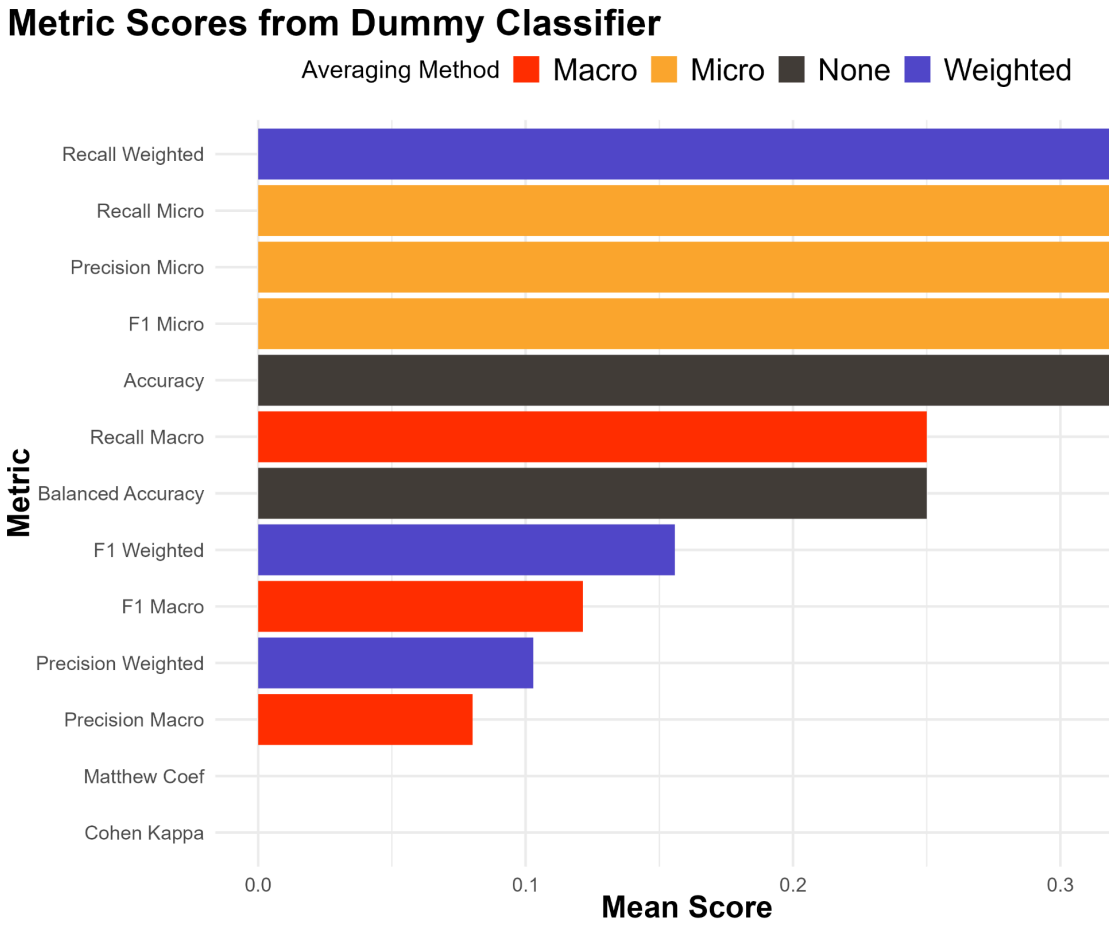

Figure 1: The performance of a Dummy (Naive) classifier, which always predicts the most common class, across a range of metrics, coloured by their averaging method where grey refers to no averaging method (None), red is Macro, yellow is Micro and blue is Weighted. Accuracy obtained the largest values and Micro averaging got identical results to Accuracy, with their value of 0.334 being the prevalence of the largest group (PHT) in the data. MCC and Cohen's Kappa both returned zero, correctly detecting the usefulness of this model. Macro and Weighted averaging were more strict compared to Micro averaging. Micro Averaging has also reduced Recall and Precision (and thus F1) to identical values while Macro Recall is more than twice the value of Macro Precision.

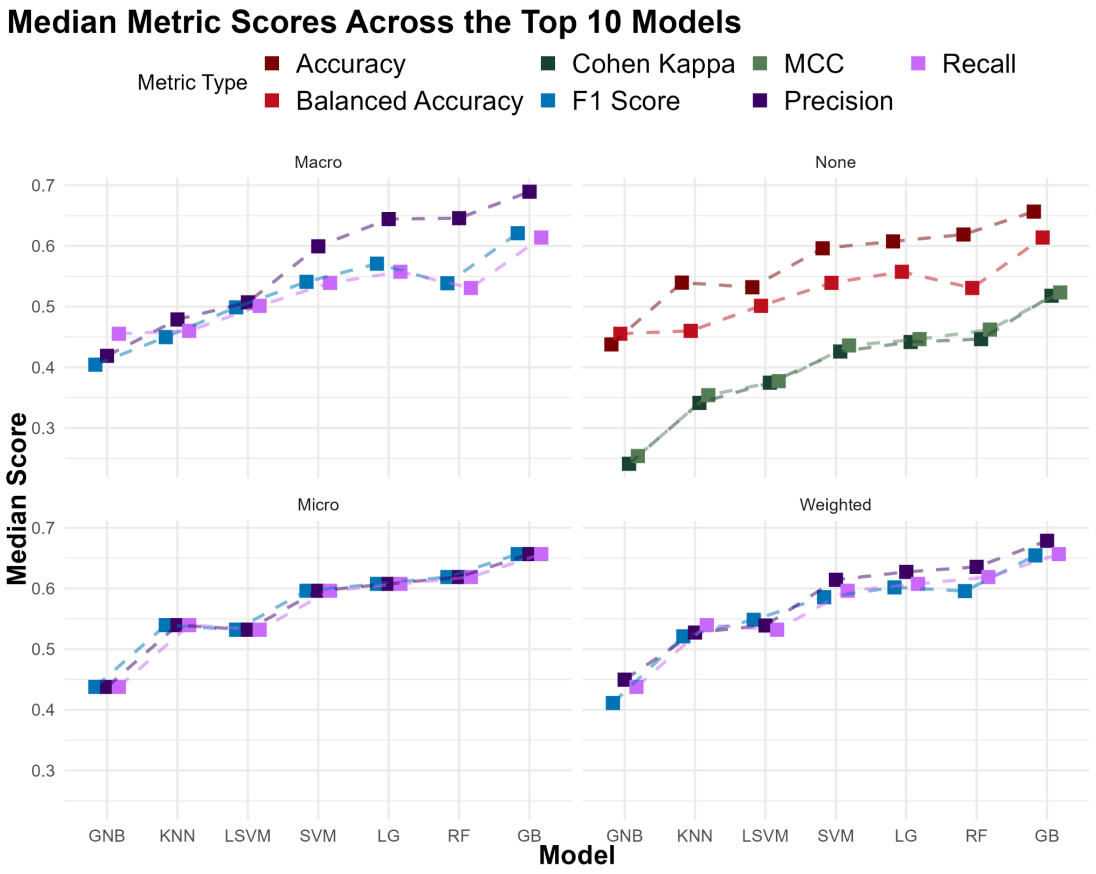

Figure 2: Having selected the top ten models for each metric in each model type, the median scores are represented with squares and their maximum and minimum values with lines. Coloured by their averaging-method: Micro-, Macro- and Weighted-averaged metrics and None, the metrics which do not require aggregation. Overall, there was similarity in their ranking of the best model, but differed in their range of values. With the reduced variation of Micro-averaging, there were disagreements between which model is the second best when compared to Macro-averaging. Concretely, Micro-averaging favoured RF whilst Macro-averaging favoured LR. (Models: **GNB** - Gaussian Naive Bayes, **KNN** - K-Nearest Neighbours, **LSVM** - Linear (Kernel) Support Vector Machine, **SVM** - Support Vector Machine, **LG** - Logistic Regression, **RF** - Random Forest, **GB** - Gradient Boosted Trees and the Metric **MCC** - Matthews Correlation Coefficient)

Evaluation of Averaging Methods

Macro or Micro averaged Recall, Precision and F1-Score are compared to the class-specific metrics in **Figure 3** for the top ten classifiers for each model. Overall, Micro averaging was slightly closer to the larger classes and was more insensitive to poor results for the less frequent classes. Macro averaging was more balanced by decreasing when CS or PPGL were markedly weaker. Concretely, Micro Recall favoured RF over SVM and LG even when RF had far lower Recall for CS and PPGL whereas Macro Recall ranks SVM and LG above RF. Furthermore, Macro Precision is larger than Micro Precision for LG because the reduced performance of the second largest class is more influential to Micro Precision. As mentioned previously, all three Micro averaged metrics have identical values, even though it appears most models perform much better in Precision than Recall.

Evaluation of Metrics to Define “Best” Model

Typically the maximum value of a metric is used to determine the best set of hyperparameters but this depends on the metric used. **Figure 4** shows the model’s True Positive Rate (Recall) of the individual classes with the x-axis indicating the metric used to obtain the optimal values for the hyperparameters (See **Appendix G** for full overview). For simplicity and due to the similarity in performance, the graph focuses on four types of models (GB, RF, SVM and LG) which all displayed substantial changes for CS, notably in Balanced Accuracy, Macro-Recall and Macro-F1 Score. Whilst Micro- and Weighted-averaged metrics are not shown, they selected identical models to Accuracy, which overlooked for a poorer performance in CS. However, whilst MCC and Cohen’s Kappa are significantly more strict, they appeared to mirror these metrics too, likely stemming from their class-agnostic evaluation of the models, which will naturally favour the larger classes. The variation in performance from the choice of metric appears inversely proportional to frequency of the label as the greatest change

in TPR was seen in CS, being the smallest group, and there was minimal change in the two most frequent labels.

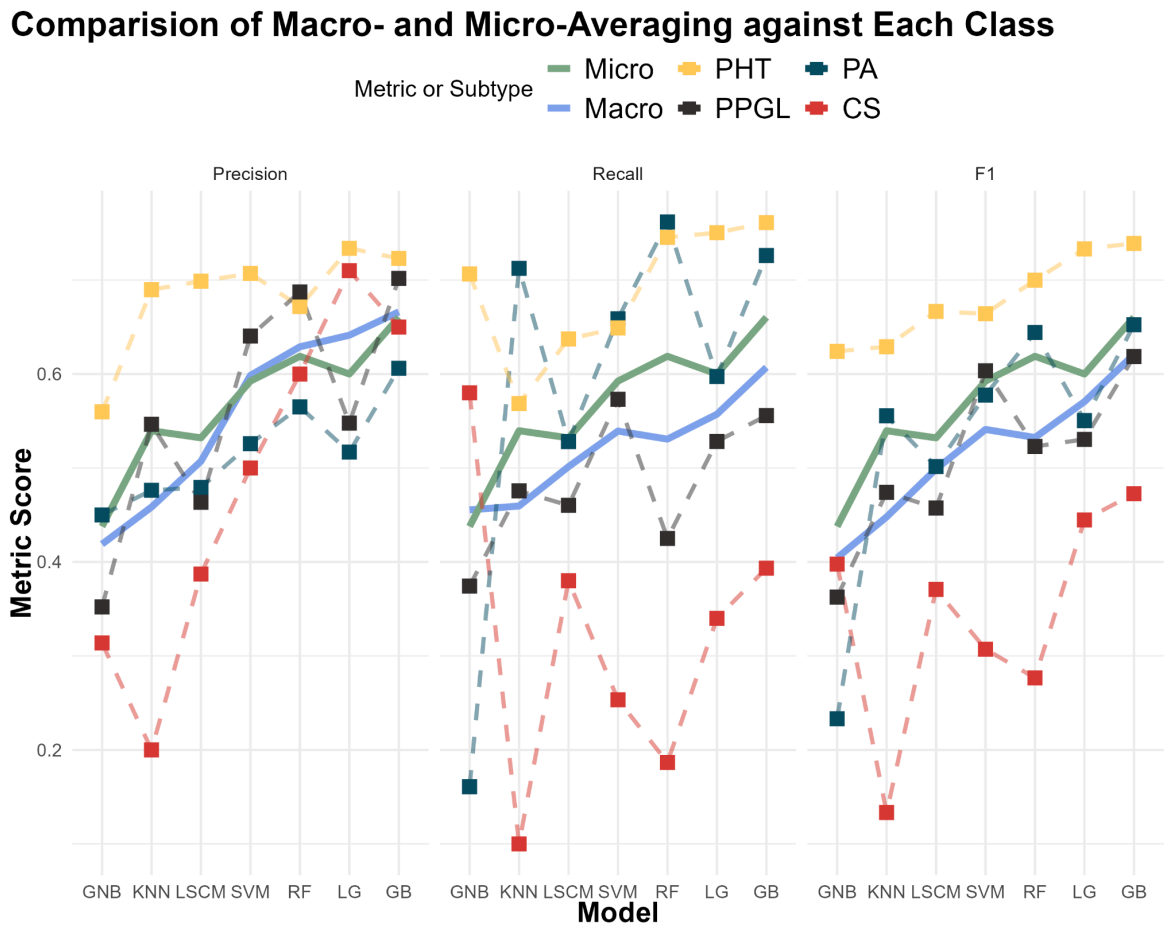

Figure 3: Median Macro- and Micro-averaged Precision, Recall and F1-Score for the top ten classifiers for each model are compared to class-specific metrics. Micro-averaged metrics were less responsive to weak results in the minority classes (CS and PPGL) compared to Macro-averaged metrics. Micro-Recall appeared unphased by the poor Recall for CS in K-Nearest Neighbours and Random Forest. Most models struggled more with Recall than Precision, but Micro-averaging obscured this pattern as all three Micro-averaged metrics have identical values. (Models: **GNB** - Gaussian Naive Bayes, **KNN** - K-Nearest Neighbours, **LSVM** - Linear (Kernel) Support Vector Machine, **SVM** - Support Vector Machine, **LG** - Logistic Regression, **RF** - Random Forest, **GB** - Gradient Boosted Trees and the Subtypes **PHT** - Primary Hypertension, **PA** - Primary Aldosteronism, **PPGL** - Pheochromocytoma/catecholamine-producing Paraganglioma, **CS** - Cushing Syndrome)

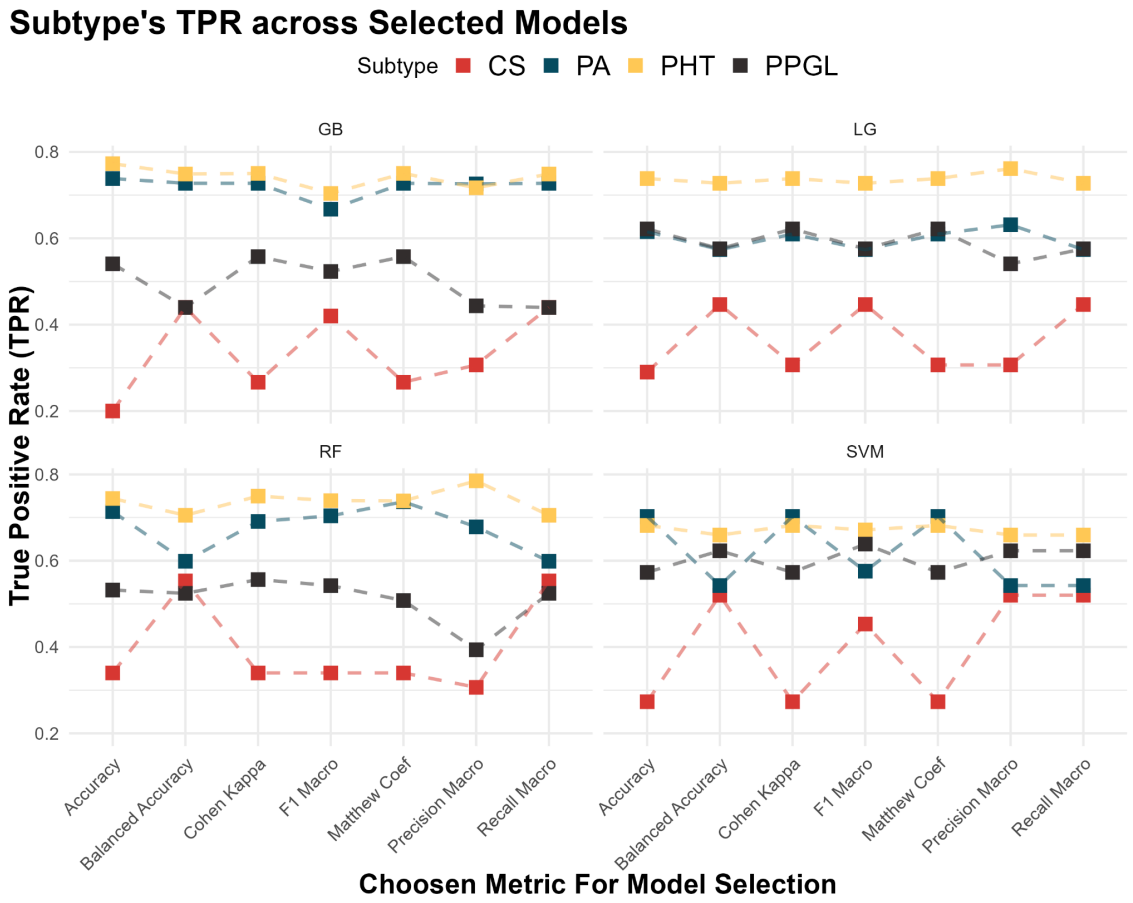

Figure 4: The True Positive Rate (TPR) of the four subtypes are influenced by the metric selected to define the best set of hyperparameters, across the four types of models. For example, F1-Macro is the model with the highest F1-macro rather than the metric itself. This aims to show the consequences of choosing a metric for model selection. The variation in the TPR is most pronounced in the minority classes, while the more frequent classes (PHT and PA) are relatively unaffected by the choice of metric. Macro-averaged metrics differed from the rest, with better performance in the minority classes at the subtle decrement of the larger classes. Micro- and Weighted-averaged metrics were removed given they select identical models to Accuracy. (Models: **GB** - Gradient Boosted Trees, **LG** - Logistic Regression, **RF** - Random Forest and **SVM** - Support Vector Machine, Subtypes: **PHT** - Primary Hypertension, **PA** - Primary Aldosteronism, **PPGL** - Pheochromocytoma/ catecholamine-producing Paraganglioma, **CS** - Cushing Syndrome)

Discussion

The absolute metric values are not the priority of this report, as the focus is on the differences between metrics and their ranking of models. Accuracy and Micro averaged metrics obtained identical values and selected the same model, indicating Micro average offered no unique information about model performance. Especially in the context of model evaluation, having Micro Precision and Micro Recall return equal scores reduced the diagnostic information available. Furthermore, Micro averaging was insensitive to poor performances in the minority class which will have severe practical limitations if used in clinical settings. Macro averaging does appear influenced by the minority class, providing valuable information if the model is struggling to differentiate the less frequent labels. Although it assigns each class the same importance regardless of size, it does not reward a model that is overlooking the minority class(es). This is likely desirable in a medical setting when the performance of every disease subtype is critical, especially when these subtypes have diverse pathophysiologies and treatment plans. Weighted averaging may seem like a convenient median between Macro and Micro, but it selected the same models as Micro-averaging and Accuracy, as the larger classes have proportionally larger weights in the formula. Weighted averaging also diminished and underestimated the difference between Precision and Recall, which is preserved with Macro averaging.

Collectively, the class-agnostic metrics selected the same models but rewarded them with different scores. Accuracy was the most optimistic while MCC and Cohen's Kappa had much lower values. Balanced Accuracy appeared more robust to Accuracy, given the identical results with Macro-Recall. Although they may not provide class-specific information, the notably gap between Accuracy and MCC or Cohen's Kappa could indicate that MCC and Cohen's Kappa acknowledge the limited predictive capabilities of the smaller classes. This would be consistent with Chicco and Jurman (2020), which

illustrated how MCC and Accuracy become increasingly discordant with larger imbalances as MCC is not skewed by the asymmetry in class frequency. The similarity of MCC and Cohen's Kappa also mirrors the results in Wardhani *et al.* (2019), indicating they have overlapping characteristics and only using one may be sufficient.

Nevertheless, the metrics used have important differences between them but they all ranked Gradient Boosted Trees as the best model which was unexpected as the microRNA data is from Reel and colleagues (2022) which had Random Forest as their best classifier. Whereas, Caruana and Niculescu-Mizil (2006) observed Gradient Boosted Trees marginally outperformed Random Forest in most datasets. One surprising element is the similarity in performance for Logistic Regression, Support Vector Machines and Random Forest, which had varying performances in Caruana's and Niculescu-Mizil's comprehensive comparison of supervised models. This potentially could suggest they were equally inhibited by the lack of data, as the capabilities of an ensemble method would likely become more clear with more samples.

With the choice of metric disproportionality affecting how the minority classes are handled, the appropriate metric depends upon the significance of the minority classes. A pragmatic choice of metric would likely be Balanced Accuracy or Macro-averaged metrics, to identify poor predictive results for the smaller classes. Alternatively, visualising the performance on a class-specific level would also prevent a class-agnostic metric to hide the performance on the minority labels. However, MCC has been shown to be robust for imbalanced datasets in biomedical applications and bioinformatics (Boughorbel, Jarray and El-Anbari, 2017; Brown, 2018; Chicco and Jurman, 2020; Chicco, Tötsch and Jurman, 2021). Realistically, a model would require to perform very well for each disease subtype to be considered useful and safe. Whilst it may be convenient to aggregate the model's capabilities into a single value, explicitly

detailing the metrics for each class would improve transparency of the practical limitations of the model. Additionally, this may allow a clearer and more robust comparison between models by judging their capability of predicting each disease rather than reporting the general performance.

Given the models were trained on one relatively small dataset, general metric recommendations for every medical or biological problem cannot be made. One notable characteristic of the dataset used is the label frequency does not reflect the underlying prevalence of each subtype. Metrics like Precision are directly influenced by the prevalence which renders their interpretation less reliable when the genuine prevalences are different from the available dataset(s) and especially if they are unknown. A more comprehensive assessment would increase the number of datasets used to ensure the behaviour of the metrics were consistent across a range of similar problems. Caruana and Niculescu-Mizil (2006) compared several supervised algorithms across 11 large datasets and observed no single model was the best for every problem. Furthermore, the relationship between sample size and imbalance of the dataset could be explored, to determine if the performance in the minority class is symptomatic of limited data or model bias.

Another improvement could be to implement more models to help generate model recommendations. Ensemble methods are widely considered more resistant to imbalanced data at the cost of computation, thus the inclusion of other common models like AdaBoosted or Extreme Gradient (XG) Boosted trees might be valuable to explore (Dietterich, 2000; Friedman, 2001; Galar *et al.*, 2012; Chen and Guestrin, 2016; Wardhani *et al.*, 2019). Additionally, introducing a small artificial neural network (ANN) like a Multi-Layer Perceptron (MLP) could offer some insight if the metric recommendations generalise to the increasingly popular deep learning approaches (Tang, Deng and Huang, 2016). With the increase in the number of models or the

number of datasets, a consideration of memory uses and time required to train the models might become necessary, especially as ANN requires a large quantity of data and computation. Alternatively, in the context of hyperparameter tuning, hyperparameter optimisation may be more replicable and objective as it minimises the underlying loss function of a model instead of evaluating on a performance metric (Bischl *et al.*, 2023).

To explore the available metrics further, a more fine-grained assessment of a classifier's performance could be obtained with Log Loss, sometimes called cross-entropy loss, which differentiates between confident and uncertain predictions (Mao, Mohri and Zhong, 2023). Whilst this metric is common for evaluating ANN in multiclass problems, some methods do not natively calculate a probability for the label, like a Support Vector Machine which motivated the probabilistic variation termed the Relevance Vector Machine (Tipping, 1999; Mao, Mohri and Zhong, 2023). With the growing investment into machine learning and deep learning, metrics are being invented or repurposed to improve model selection. Powers introduced Informedness and Markedness to counter the population prevalence and class frequency bias present in Precision and Recall (Powers, 2020). These metrics aim to estimate the probabilities of a model identifying the label and of a label being identified by a model, by also considering the underlying chance of a random prediction (Powers, 2020). Additionally, the metric G-mean, the geometric mean of the Recall for each group, is often used for imbalanced datasets (Wardhani *et al.*, 2019; Ri and Kim, 2020). As each score is multiplied by each other, all Recall values must be relatively good. Stemming from Information Theory, Confusion Entropy was recently introduced to have better discriminative power than Accuracy and comparable or improved robustness relative to MCC, especially in imbalanced multiclass datasets (Jurman, Riccadonna and Furlanello, 2012; Delgado and Núñez-González, 2019).

In conclusion, the choice of metric influences model selection and most notably the down-stream performance of the minority classes. Therefore, the choice of metric is dependent on the importance and relevance of the less frequent labels. However, the overall agreement and similarity between the metrics should not be overlooked. Class-agnostic metrics provide a convenient score to gauge the overall performance of the classifier but they carry the risk of allowing the performance of the less frequent labels being ignored. Macro averaging offers a pragmatic choice to highlight potential weakness in the smaller classes. Nevertheless, presenting and visualising the class-specific metrics may be the most transparent and robust method to communicate the strengths and weaknesses of a classifier. With the growing number of metrics and models, especially ensemble and deep learning methods, the choice of metric and models must be made with a consideration of the underlying biases and tradeoffs.

References

Ahsan, M.M., Luna, S.A. and Siddique, Z. (2022) ‘Machine-Learning-Based Disease Diagnosis: A Comprehensive Review’, *Healthcare*, 10(3), p. 541. Available at: https://doi.org/10.3390/healthcare10030541.

Araf, I., Idri, A. and Chairi, I. (2024) ‘Cost-sensitive learning for imbalanced medical data: a review’, *Artificial Intelligence Review*, 57(4), p. 80. Available at: https://doi.org/10.1007/s10462-023-10652-8.

Azari, H. *et al.* (2023) ‘Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer’, *Scientific Reports*, 13(1), p. 6147. Available at: https://doi.org/10.1038/s41598-023-32332-x.

Baghdadi, N.A. *et al.* (2023) ‘Advanced machine learning techniques for cardiovascular disease early detection and diagnosis’, *Journal of Big Data*, 10(1), p. 144. Available at: https://doi.org/10.1186/s40537-023-00817-1.

Bandyopadhyay, S. *et al.* (2015) ‘MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets’, *Scientific Reports*, 5, p. 8004. Available at: https://doi.org/10.1038/srep08004.

Bischl, B. *et al.* (2023) ‘Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges’, *WIREs Data Mining and Knowledge Discovery*, 13(2), p. e1484. Available at: https://doi.org/10.1002/widm.1484.

Blagus, R. and Lusa, L. (2010) ‘Class prediction for high-dimensional class-imbalanced data’, *BMC Bioinformatics*, 11(1), p. 523. Available at: https://doi.org/10.1186/1471-2105-11-523.

Boughorbel, S., Jarray, F. and El-Anbari, M. (2017) ‘Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric’, *PLOS ONE*, 12(6), p. e0177678. Available at: https://doi.org/10.1371/journal.pone.0177678.

Breiman, L. (2001) ‘Random Forests’, *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Brown, J.B. (2018) ‘Classifiers and their Metrics Quantified’, *Molecular Informatics*, 37(1–2), p. 1700127. Available at: https://doi.org/10.1002/minf.201700127.

Caruana, R. and Niculescu-Mizil, A. (2006) ‘An empirical comparison of supervised learning algorithms’, in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. *the 23rd international conference*, Pittsburgh, Pennsylvania: ACM Press, pp. 161–168. Available at: https://doi.org/10.1145/1143844.1143865.

Cawley, G.C. and Talbot, N.L.C. (2010) ‘On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation’.

Chang, C.-C. and Lin, C.-J. (2011) ‘LIBSVM: A library for support vector machines’, *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp. 1–27. Available at: https://doi.org/10.1145/1961189.1961199.

Chen, T. and Guestrin, C. (2016) ‘XGBoost: A Scalable Tree Boosting System’, in

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery (KDD '16), pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), p. 6. Available at: https://doi.org/10.1186/s12864-019-6413-7.

Chicco, D., Tötsch, N. and Jurman, G. (2021) 'The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation', *BioData Mining*, 14(1), p. 13. Available at: https://doi.org/10.1186/s13040-021-00244-z.

Defazio, A., Bach, F. and Lacoste-Julien, S. (2014) 'SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives'. arXiv. Available at: https://doi.org/10.48550/arXiv.1407.0202.

Delgado, R. and Núñez-González, J.D. (2019) 'Enhancing Confusion Entropy (CEN) for binary and multiclass classification', *PLOS ONE*, 14(1), p. e0210264. Available at: https://doi.org/10.1371/journal.pone.0210264.

Demšar, J. (2006) 'Statistical Comparisons of Classifiers over Multiple Data Sets', *The Journal of Machine learning research* [Preprint].

Dietterich, T.G. (2000) 'Ensemble Methods in Machine Learning', in *Multiple Classifier Systems*. Berlin, Heidelberg: Springer, pp. 1–15. Available at: https://doi.org/10.1007/3-540-45014-9_1.

Ding, J., Li, X. and Hu, H. (2016) 'TarPmiR: a new approach for microRNA target site prediction', *Bioinformatics*, 32(18), pp. 2768–2775. Available at: https://doi.org/10.1093/bioinformatics/btw318.

Farhadpour, S., Warner, T.A. and Maxwell, A.E. (2024) 'Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices', *Remote Sensing*, 16(3), p. 533. Available at: https://doi.org/10.3390/rs16030533.

Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics*, 29(5), pp. 1189–1232. Available at: https://doi.org/10.1214/aos/1013203451.

Galar, M. *et al.* (2012) 'A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), pp. 463–484. Available at: https://doi.org/10.1109/TSMCC.2011.2161285.

Gaudreault, J.-G., Branco, P. and Gama, J. (2021) 'An Analysis of Performance Metrics for Imbalanced Classification', in C. Soares and L. Torgo (eds) *Discovery Science*. Cham: Springer International Publishing, pp. 67–77. Available at: https://doi.org/10.1007/978-3-030-88942-5_6.

GBD 2015 Risk Factors Collaborators (2016) 'Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global

Burden of Disease Study 2015', *Lancet (London, England)*, 388(10053), pp. 1659–1724. Available at: https://doi.org/10.1016/S0140-6736(16)31679-8.

Grandini, M., Bagli, E. and Visani, G. (2020) 'Metrics for Multi-Class Classification: an Overview'. arXiv. Available at: http://arxiv.org/abs/2008.05756

Hanniford, D. *et al.* (2015) 'A miRNA-based signature detected in primary melanoma tissue predicts development of brain metastasis', *Clinical cancer research : an official journal of the American Association for Cancer Research*, 21(21), pp. 4903–4912. Available at: https://doi.org/10.1158/1078-0432.CCR-14-2566.

Jurman, G., Riccadonna, S. and Furlanello, C. (2012) 'A Comparison of MCC and CEN Error Measures in Multi-Class Prediction', *PLOS ONE*, 7(8), p. e41882. Available at: https://doi.org/10.1371/journal.pone.0041882.

Koch, C., Papadopoulou-Marketou, N. and Chrousos, G.P. (2020) 'Overview of Endocrine Hypertension', in *Endotext [Internet]*. MDText.com, Inc. Available at: https://www.ncbi.nlm.nih.gov/sites/books/NBK278980/

Li, W.Y. *et al.* (2014) 'Circulating microRNAs as potential non-invasive biomarkers for the early detection of hypertension-related stroke', *Journal of Human Hypertension*, 28(5), pp. 288–291. Available at: https://doi.org/10.1038/jhh.2013.94.

Livak, K.J. and Schmittgen, T.D. (2001) 'Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method', *Methods*, 25(4), pp. 402–408. Available at: https://doi.org/10.1006/meth.2001.1262.

Mao, A., Mohri, M. and Zhong, Y. (2023) 'Cross-Entropy Loss Functions: Theoretical Analysis and Applications'. arXiv. Available at: http://arxiv.org/abs/2304.07288

Mills, K.T., Stefanescu, A. and He, J. (2020) 'The global epidemiology of hypertension', *Nature reviews. Nephrology*, 16(4), p. 223. Available at: https://doi.org/10.1038/s41581-019-0244-2.

Oparil, S. *et al.* (2018) 'Hypertension', *Nature reviews. Disease primers*, 4, p. 18014. Available at: https://doi.org/10.1038/nrdp.2018.14.

Parveen, A. *et al.* (2019) 'Applications of Machine Learning in miRNA Discovery and Target Prediction', *Current Genomics*, 20(8), pp. 537–544. Available at: https://doi.org/10.2174/1389202921666200106111813.

Powers, D.M.W. (2020) 'Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2010.16061.

Rainio, O., Teuho, J. and Klén, R. (2024) 'Evaluation metrics and statistical tests for machine learning', *Scientific Reports*, 14(1), p. 6086. Available at: https://doi.org/10.1038/s41598-024-56706-x.

Reel, P.S. *et al.* (2022) 'Machine learning for classification of hypertension subtypes using multi-omics: A multi-centre, retrospective, data-driven study', *eBioMedicine*, 84. Available at: https://doi.org/10.1016/j.ebiom.2022.104276.

Ri, J. and Kim, H. (2020) 'G-mean based extreme learning machine for imbalance learning', *Digital Signal Processing*, 98, p. 102637. Available at:

https://doi.org/10.1016/j.dsp.2019.102637.

Robertson, S. *et al.* (2017) ‘Regulation of Corticosteroidogenic Genes by MicroRNAs’, *International Journal of Endocrinology*, 2017, p. 2021903. Available at: https://doi.org/10.1155/2017/2021903.

Samadishadlou, M. *et al.* (2023) ‘Unlocking the potential of microRNAs: machine learning identifies key biomarkers for myocardial infarction diagnosis’, *Cardiovascular Diabetology*, 22(1), p. 247. Available at: https://doi.org/10.1186/s12933-023-01957-7.

Schmittgen, T.D. and Livak, K.J. (2008) ‘Analyzing real-time PCR data by the comparative CT method’, *Nature Protocols*, 3(6), pp. 1101–1108. Available at: https://doi.org/10.1038/nprot.2008.73.

Shi, L. *et al.* (2015) ‘Mechanisms and therapeutic potential of microRNAs in hypertension’, *Drug discovery today*, 20(10), pp. 1188–1204. Available at: https://doi.org/10.1016/j.drudis.2015.05.007.

Song, M.A. *et al.* (2015) ‘Differential expression of microRNAs in ischemic heart disease’, *Drug discovery today*, 20(2), pp. 223–235. Available at: https://doi.org/10.1016/j.drudis.2014.10.004.

Staessen, J.A. *et al.* (2003) ‘Essential hypertension’, *The Lancet*, 361(9369), pp. 1629–1641. Available at: https://doi.org/10.1016/S0140-6736(03)13302-8.

Stehman, S.V. and Foody, G.M. (2019) ‘Key issues in rigorous accuracy assessment of land cover products’, *Remote Sensing of Environment*, 231, p. 111199. Available at: https://doi.org/10.1016/j.rse.2019.05.018.

Sudano, I., Beuschlein, F. and Lüscher, T.F. (2018) ‘Secondary causes of hypertension’, in A.J. Camm et al. (eds) *The ESC Textbook of Cardiovascular Medicine*. Oxford University Press, p. 0. Available at: https://doi.org/10.1093/med/9780198784906.003.0566.

Sudano, I., Suter, P. and Beuschlein, F. (2023) ‘Secondary hypertension as a cause of treatment resistance’, *Blood Pressure*, 32(1), p. 2224898. Available at: https://doi.org/10.1080/08037051.2023.2224898.

Tai, M.C. *et al.* (2016) ‘Blood-borne miRNA profile-based diagnostic classifier for lung adenocarcinoma’, *Scientific Reports*, 6(1), p. 31389. Available at: https://doi.org/10.1038/srep31389.

Tang, J., Deng, C. and Huang, G.-B. (2016) ‘Extreme Learning Machine for Multilayer Perceptron’, *IEEE Transactions on Neural Networks and Learning Systems*, 27(4), pp. 809–821. Available at: https://doi.org/10.1109/TNNLS.2015.2424995.

Tipping, M. (1999) ‘The Relevance Vector Machine’, in *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://proceedings.neurips.cc/paper_files/paper/1999/hash/f3144cefe89a60d6a1afaf7859c5076b-Abstract.html .

Topol, E.J. (2019) ‘High-performance medicine: the convergence of human and artificial intelligence’, *Nature Medicine*, 25(1), pp. 44–56. Available at: https://doi.org/10.1038/s41591-018-0300-7.

Wardhani, N.W.S. *et al.* (2019) ‘Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data’, in *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 14–18. Available at: https://doi.org/10.1109/IC3INA48034.2019.8949568.

Weber, M. *et al.* (2010) ‘MiR-21 Is Induced in Endothelial Cells by Shear Stress and Modulates Apoptosis and eNOS Activity’, *Biochemical and biophysical research communications*, 393(4), pp. 643–648. Available at: https://doi.org/10.1016/j.bbrc.2010.02.045.

Williams, B. *et al.* (2018) ‘2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH)’, *European Heart Journal*, 39(33), pp. 3021–3104. Available at: https://doi.org/10.1093/eurheartj/ehy339.

Yu, K.-H., Beam, A.L. and Kohane, I.S. (2018) ‘Artificial intelligence in healthcare’, *Nature Biomedical Engineering*, 2(10), pp. 719–731. Available at: https://doi.org/10.1038/s41551-018-0305-z.

Zhang, H. (2004) ‘The Optimality of Naive Bayes’.

Zhang, H. *et al.* (2018) ‘Endothelial dysfunction in diabetes and hypertension: Role of microRNAs and long non-coding RNAs’, *Life Sciences*, 213, pp. 258–268. Available at: https://doi.org/10.1016/j.lfs.2018.10.028.