MOSQUITOES CLASSIFICATION

Random Forest and Boosting Application

Corrà Sara - Shaboian Goar

OVERVIEW

1. Introduction
    a. Research Problem
    b. Audio Data Analysis
2. Data preparation
    1. Class Imbalance
    2. Data Augmentation
    3. Rebalancing Procedure
3. Ensemble Learning
    1. Random Forest
    2. Boosting
4. Model Interpretation
5. Conclusions

PROBLEM STATEMENT

01 MOSQUITOS: VECTORS OF PATHOGENS

Mosquito-borne diseases are extremely widespread globally, with billions of people at risk of contracting dangerous and life-threatening infections and viruses. Approximately 700,000 people die annually because of mosquito-borne illnesses.

Thus, there is an urgent necessity to promote better methods of disease prevention. One possible ways of development is implementing data science approaches aimed at detecting and recognizing mosquitoes in the surrounding environment, from which point warning or elimination systems may be introduced.

To this end, this analysis is aimed at developing a machine learning approach to classify mosquitoes based on the sounds they emit.

02 RESEARCH TASK

In the analysis, 6 classes of mosquitoes are introduced, each of which is a vector for dangerous illnesses.

For the purpose of disease surveillance and vector control, the stakeholder for this task is a research institute that has received a grant for developing statistical approaches to epidemiological modeling and disease prediction. As such, a comprehensive approach, with a full pipeline from raw dataset to producing final classification results, is required.

The stakeholder is invested in this project because it aligns with their mission to advance scientific understanding and practical solutions for combating vector-borne diseases

03 WINGBEATS DATASET

The analysis is conducted on the Wingbeats dataset, which was collected by Biogents AG, a company specializing in mosquito control research, with the aid of Irideon IoT, a company based in Spain.

The data in the study were collected using an optoelectronic sensor system designed to record the fragile signature of insect wingbeats. The obtained pseudo-acoustic signals were used to retrieve audio format data by using a microprocessor.

Aedes Aegypti

(Yellow Fever Mosquito)

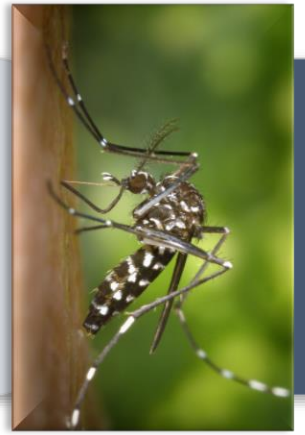
Dengue fever, Zika fever, Yellow Fever viruses, Marburg

- Tropical, subtropical, throughout the world
- Among the most widespread specied

Aedes Albopictus

(Asian Tiger Mosquito)

Yellow fever, Dengue, ZIka



- Native to tropical areas of Southeast Asia
- In the past centuries, has spread to many countries

Anopheles Arabiensis

(African Malaria Mosquito)

Malaria, Plasmodium falciparum

- Zoophilic mosquito
- Endemic to Africa, north into the Sahel and south into steppe of Namibia

Anopheles Gambiae

(African Malaria Mosquito)

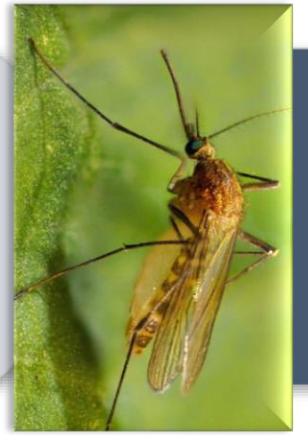
Malaria

- The most efficient malaria vector
- Distributed across sub-Saharan Africa

Culex Pipiens

(Noerthern House Mosquito)

West Nile Virus, Saint Louis encephalitis viruses



- Can be found both in urban and suburban temperature regions
- Prevalent on most continents

Culex quinquefasciatus

(South House Mosquito)

Avian malaria, West Nile, encephalitis viruses, Zika

- Among most abundant peridomestic mosquitos
- Found in tropical and subtropical regions

PROBLEM STATEMENT

01 MOSQUITOS: VECTORS OF PATHOGENS

Mosquito-borne diseases are extremely widespread globally, with billions of people at risk of contracting dangerous and life-threatening infections and viruses. Approximately 700,000 people die annually because of mosquito-borne illnesses.

Thus, there is an urgent necessity to promote better methods of disease prevention. One possible ways of development is implementing data science approaches aimed at detecting and recognizing mosquitoes in the surrounding environment, from which point warning or elimination systems may be introduced.

To this end, this analysis is aimed at developing a machine learning approach to classify mosquitoes based on the sounds they emit.

02 RESEARCH TASK

In the analysis, 6 classes of mosquitoes are introduced, each of which is a vector for dangerous illnesses.

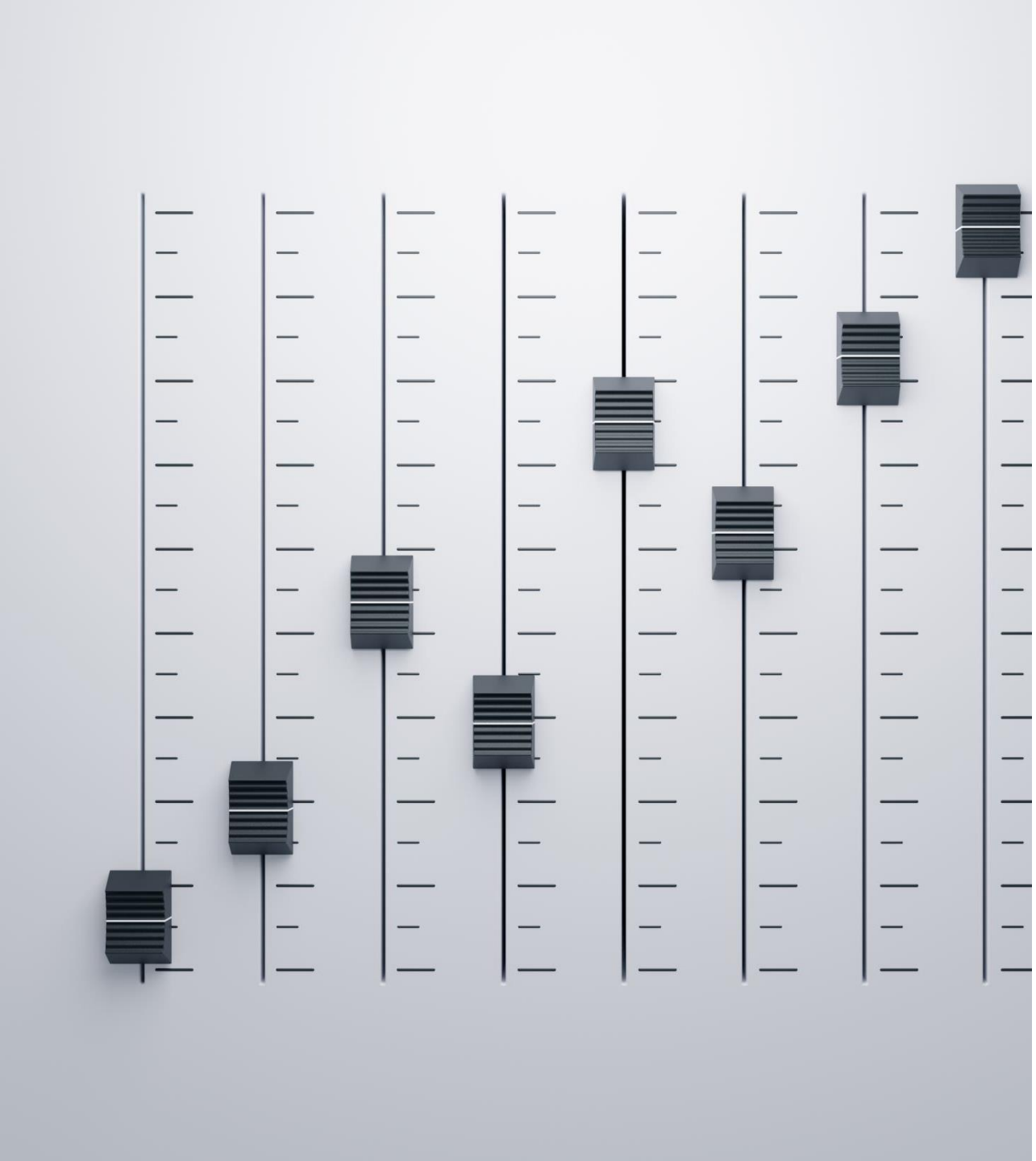
For the purpose of disease surveillance and vector control, the stakeholder for this task is a research institute that has received a grant for developing statistical approaches to epidemiological modeling and disease prediction. As such, a comprehensive approach, with a full pipeline from raw dataset to producing final classification results, is required.

The stakeholder is invested in this project because it aligns with their mission to advance scientific understanding and practical solutions for combating vector-borne diseases

03 WINGBEATS DATASET

The analysis is conducted on the Wingbeats dataset, which was collected by Biogents AG, a company specializing in mosquito control research, with the aid of Irideon IoT, a company based in Spain.

The data in the study were collected using an optoelectronic sensor system designed to record the fragile signature of insect wingbeats. The obtained pseudo-acoustic signals were used to retrieve audio format data by using a microprocessor.



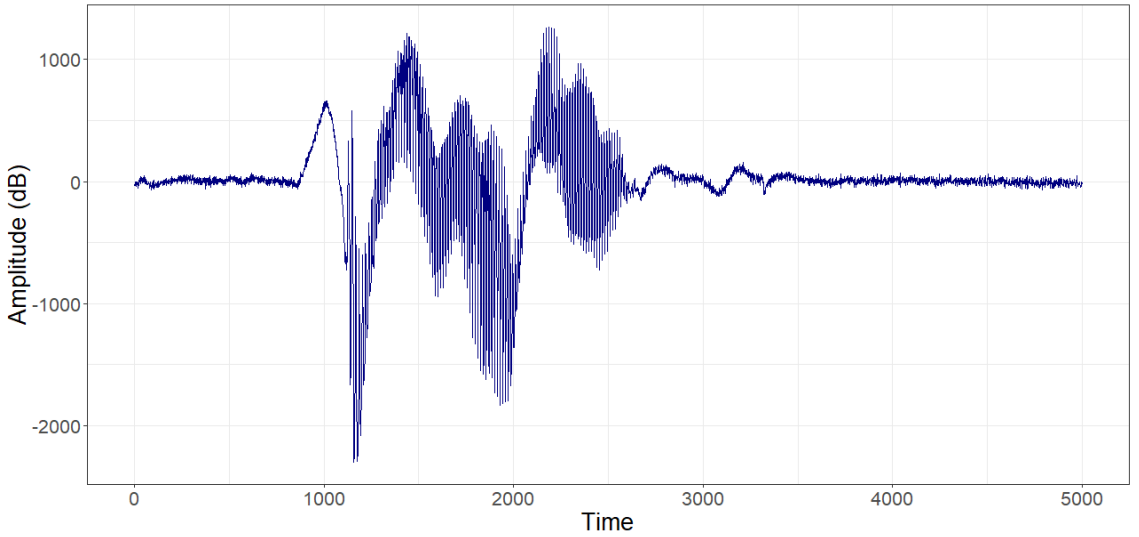
SOUND THEORY

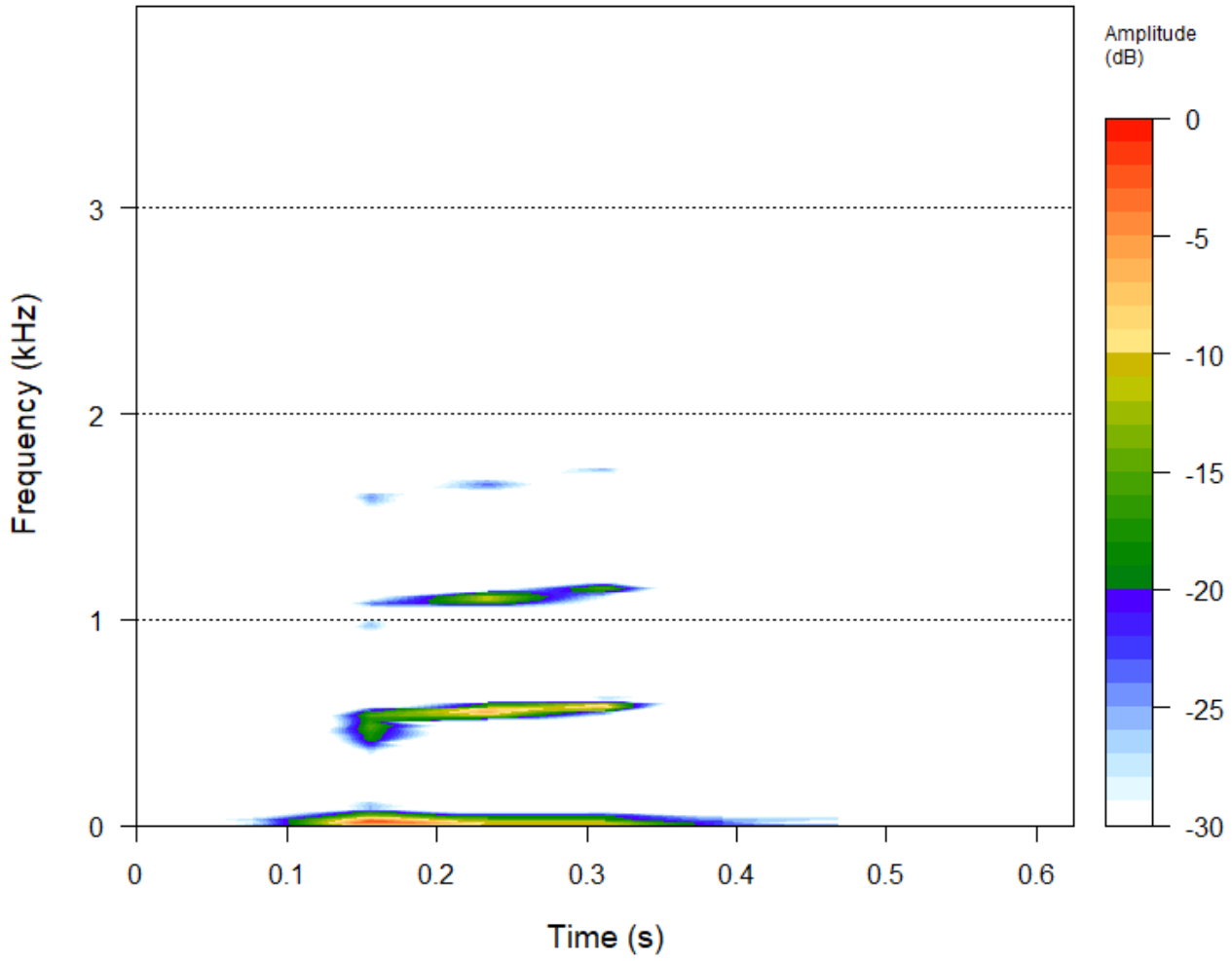
HOW HUMANS PERCEIVE SOUNDS

Audio signals are signals that vibrate in the audible frequency range. When someone talks, it generates air pressure signals; the ear takes in these air pressure differences and communicates with the brain. That's how the brain helps a person recognize that the signal is speech and understand what someone is saying

TIME DOMAIN REPRESENTATION OF MOSQUITO SOUND

```
Wave Object
        Number of Samples:      5000
        Duration (seconds):     0.62
        Samplingrate (Hertz):   8000
        Channels (Mono/Stereo): Mono
        PCM (integer format):   TRUE
        Bit (8/16/24/32/64):    16
```



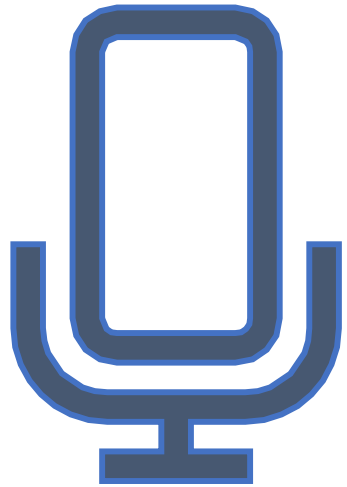


SPECTROGRAM

Mathematically, a spectrum is the Fourier transform of a signal. A Fourier transform converts a time-domain signal to the frequency domain. A spectrogram displays the strength of the signal or loudness (in decibels) over a period of time at different values of frequencies of the wave.

INTRODUCTION TO MFCC COEFFICIENTS

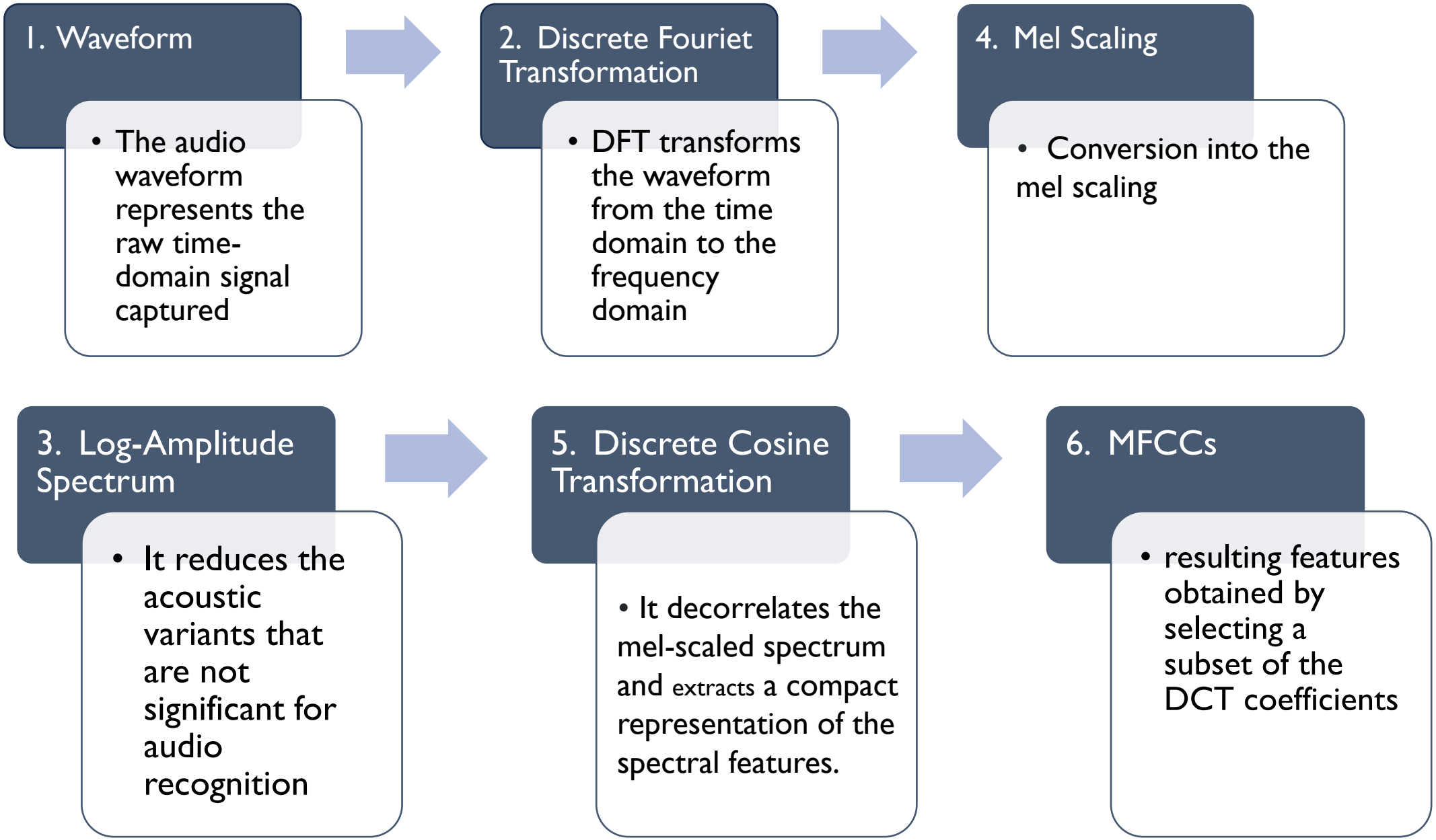
To prepare sound for statistical analysis, it must be converted into numerical data with relevant covariates for algorithm processing. This conversion is achieved through Mel-Frequency Cepstral Coefficients (MFCC), a technique commonly used in speech and audio processing. MFCC extracts key features from sound signals, mimicking the human auditory system's sensitivity to sound frequency and intensity.

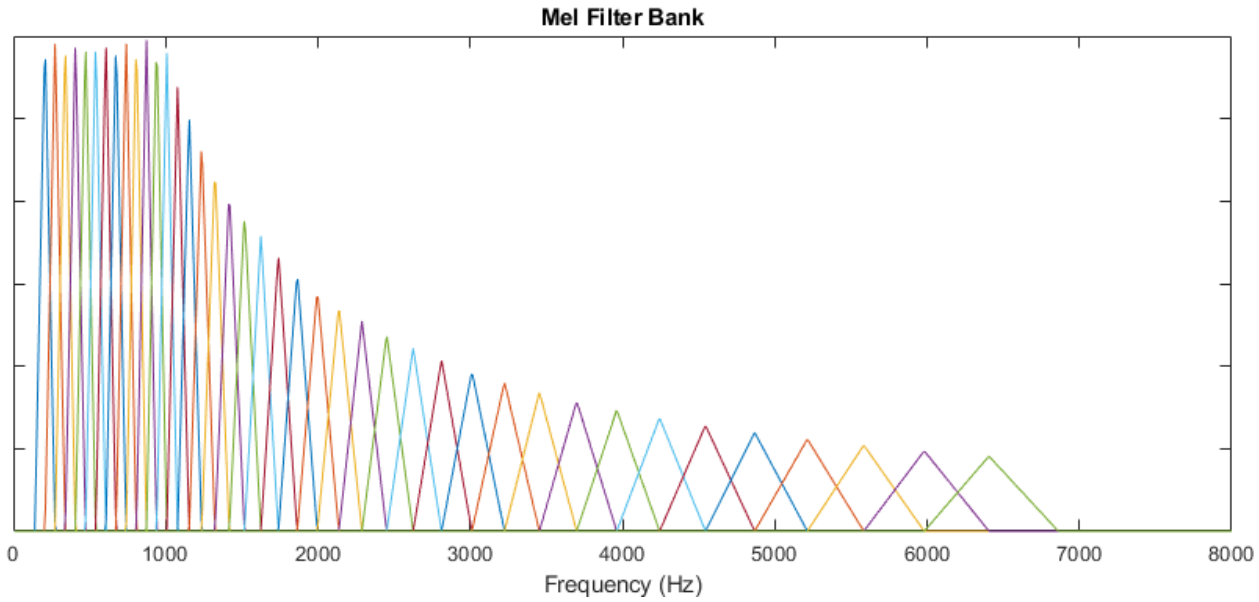


MFCC OBJECTIVES

Make the extracted features independent.

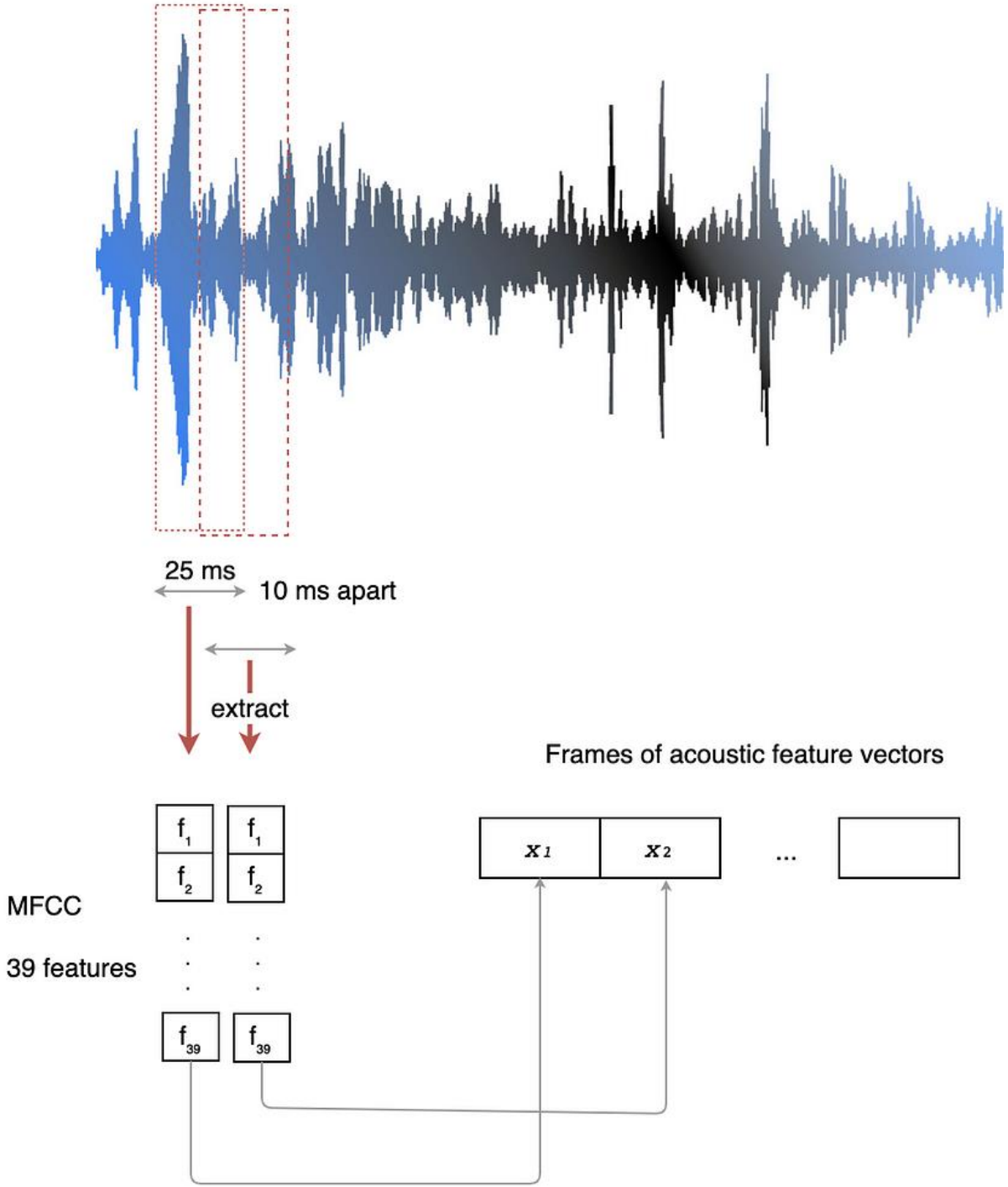
Adjust to how humans perceive loudness and frequency of sound.


1. Waveform
• The audio waveform represents the raw time-domain signal captured
2. Discrete Fouriet Transformation
• DFT transforms the waveform from the time domain to the frequency domain
4. Mel Scaling
• Conversion into the mel scaling
3. Log-Amplitude Spectrum
• It reduces the acoustic variants that are not significant for audio recognition
5. Discrete Cosine Transformation
• It decorrelates the mel-scaled spectrum and extracts a compact representation of the spectral features.
6. MFCCs
• resulting features obtained by selecting a subset of the DCT coefficients



MEL FILTER BANK

- Mel filterbanks mimic human hearing better than linear frequency scales. They use overlapping triangular filters to convert linear frequencies to the mel scale, aligning better with human perception.



FINAL COEFFICIENTS

The function operates by processing each frame of the sound signal, with the frame length specified by the window size. Within each frame, a set of features is extracted and then vectorized, resulting in a single vector representing each sound. These features encompass 12 variables along with 12 delta variables, which encode temporal information

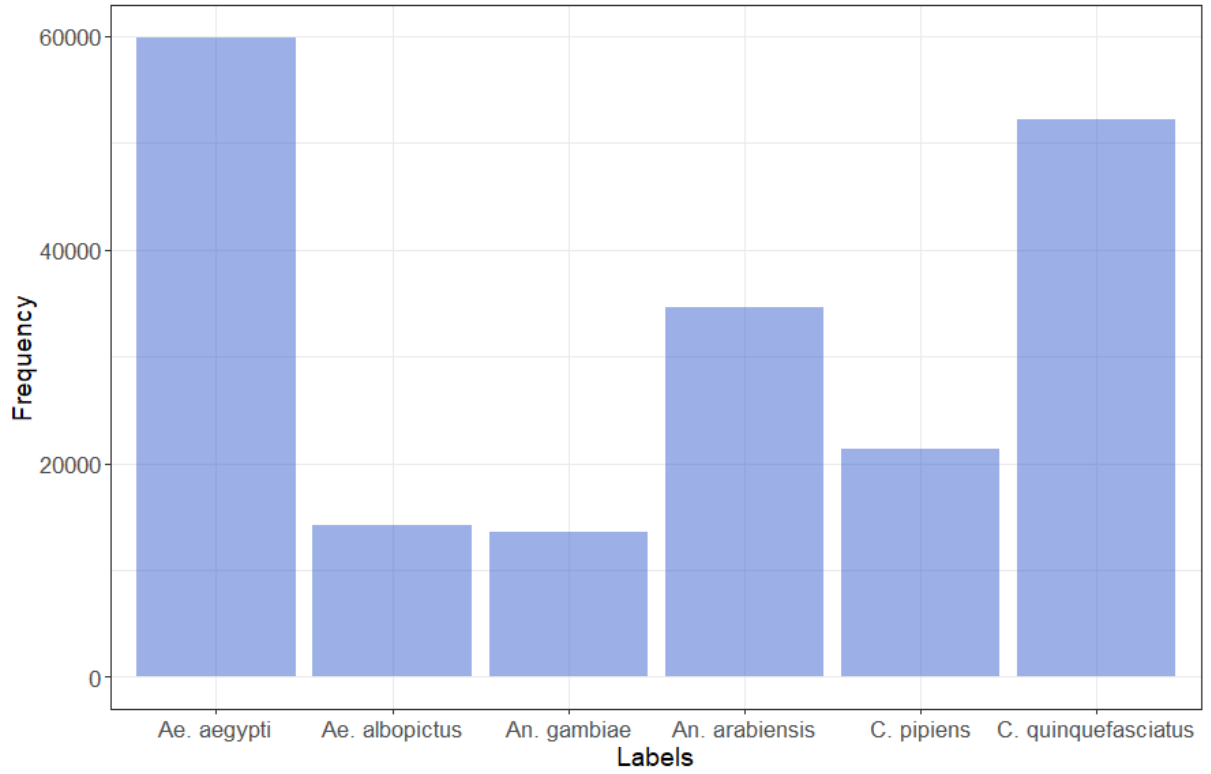
DATASET SPLIT

	TRAIN	VALIDATION	TEST	Total
Absolute frequency	195,698	41,936	41,931	279,565
Relative frequency	70.00 %	15.00 %	15.00 %	100 %

TRAIN DATASET

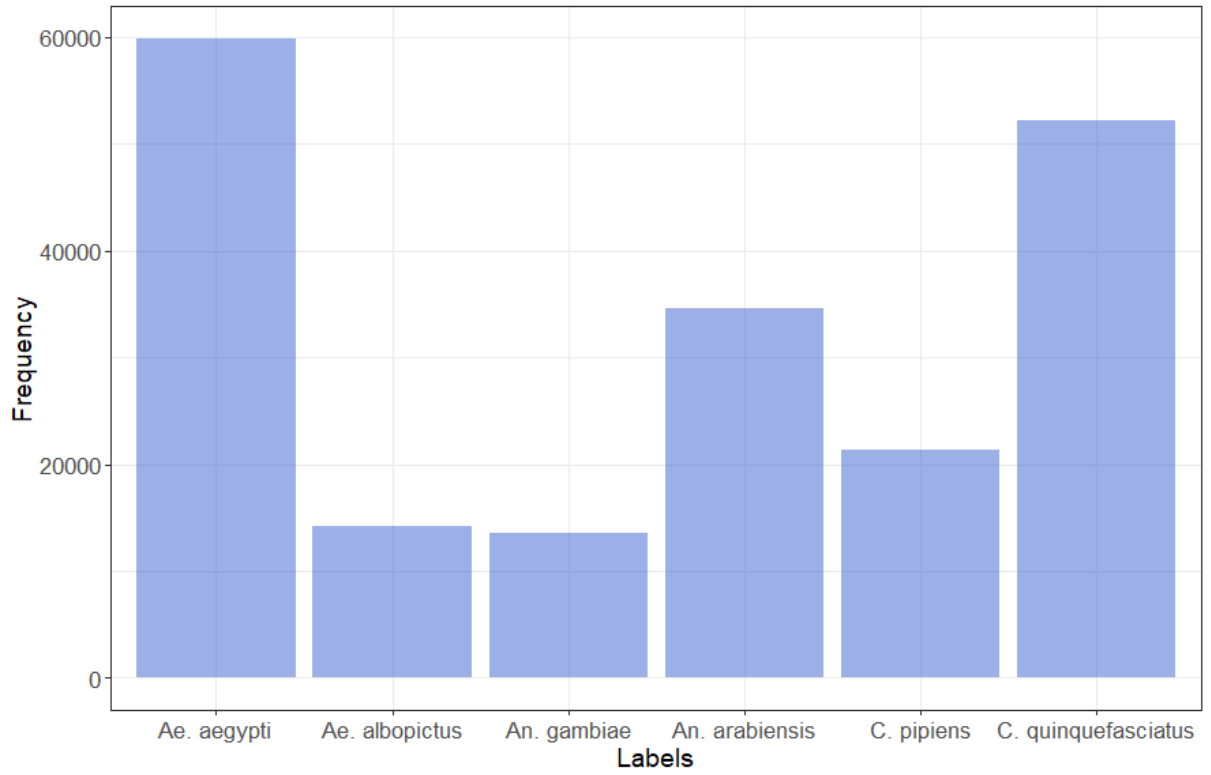
	1	2	3	...	176	label
1	-9.135040	-5.371376	-2.5797885	…	-6.611027	1
2	-7.588031	-6.362967	-1.1263650	…	-10.166367	1
…	…	…	…	…	…	…
195698	-6.611027	-10.166367	-2.8656027	…	465.46366	6

CLASS IMBALANCE



	Ae. aegypti	Ae. albopictus	An. Gambiae	An. arabiensis	C. pipiens	C. quinquefasciatus	Total
Absolute frequency	59'888	14'162	13'508	34'630	21'291	52'219	195698
Relative frequency	30.602 %	7.236 %	6.902 %	17.695 %	10.879 %	26.683 %	100 %

CLASS IMBALANCE



	Ae. aegypti	Ae. albopictus	An. Gambiae	An. arabiensis	C. pipiens	C. quinquefasciatus	Total
Absolute frequency	59'888	14'162	13'508	34'630	21'291	52'219	195698
Relative frequency	30.602 %	7.236 %	6.902 %	17.695 %	10.879 %	26.683 %	100 %

TO ADDRESS THIS CHALLENGE, TWO METHODS ARE PROPOSED:

DATA AUGMENTATION

In order to address the challenge of imbalanced classes within our dataset, we employ data augmentation techniques, focusing specifically on the classes of Ae. albopictus and An. Gambiae, which collectively represent 14.138% of the total data and have fewer observations. To bolster the representation of these classes, we augment the dataset by adding 10,000 observations to each.

This augmentation process involves two distinct steps, which will be outlined in the subsequent flowchart.

ADVANTAGE

The benefit of handling unstructured data is the possibility of augmenting data without introducing a lot of bias, unlike in tabular data.

Choose values for:
- Pitch
- Formant
- Timestretch

Pitch: The perceived frequency of a sound, determining its high or low tone.
Formants: Resonant frequencies in the human voice that shape its timbre and vowel quality.
Timestretch: Altering the duration of a sound without affecting its pitch.

Read wave file and apply the designed transformation to the sound.

With $p = 0.5$

Do nothing

Add noise to the sound

OVERSAMPLING AND UNDERSAMPLING

- Most approaches available for addressing class imbalance are provided for binary classification problems, which cannot be applied directly for multi-class classification.
- An important consideration remains within-class imbalance.
- Random oversampling leads to overfitting and increase in computational burden.
- Random undersampling leads to loss of information, especially if sub-clusters are present in the data.

OVERSAMPLING AND UNDERSAMPLING

- SMOTE oversampling provides impovement by generating synthetic examples.
- Cluster-based undersampling allows to preserve the within-class structure.
- Combining under- and ovesamplng allows to address the drawbacks of both methods in a comprehensive way.
- SMOTE and Clustered Undersampling Technique (SCUT) is proposed to address between-class and within-class imbalance in multi-class problems.

OVERSAMPLING

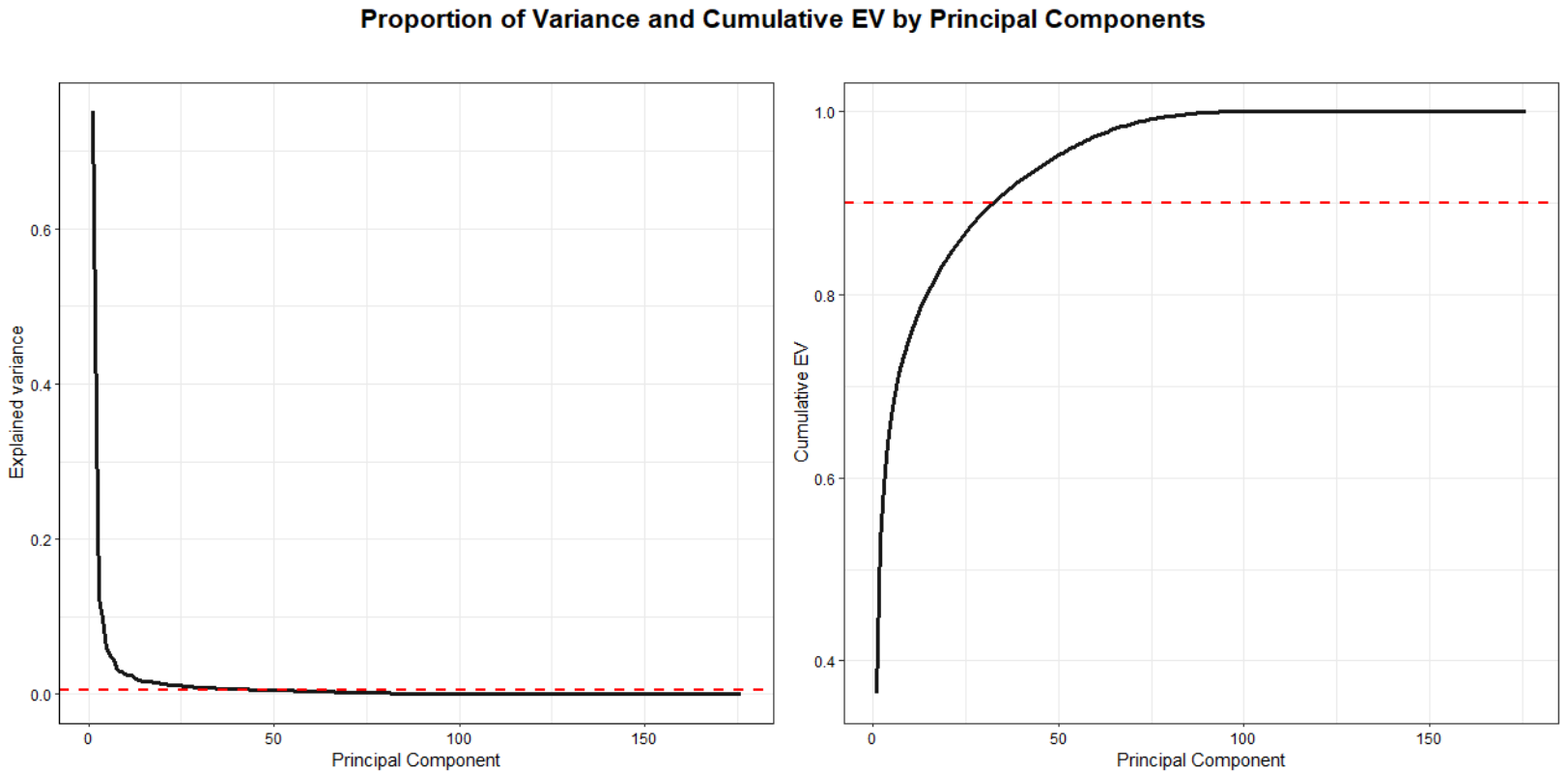
SMOTE algorithm uses KNN-based approach to create synthetic instancs of the underrepresented class, by discounting the distance from a randomly selected observation to the neighbours.

UNDERSAMPLING

Expectation-Maximization algorithm uses probability distributions formed by a mixture of Gaussians. The data points are assigned probabilities of belonging to a particular cluster.

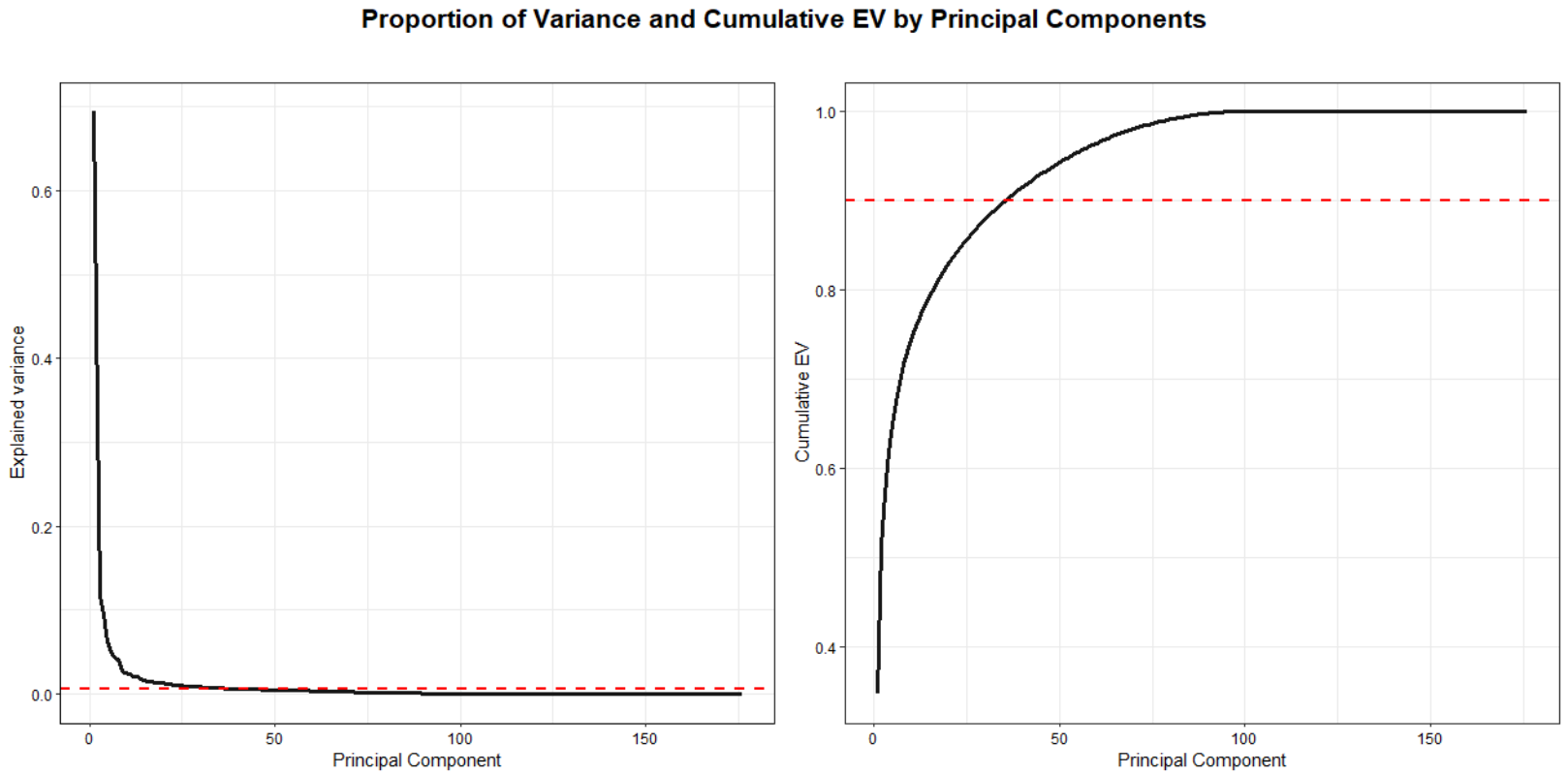
SCUT algorithm	Set m: average number of instances in partitions
	SMOTE oversampling: for classes with $N_j < m$, perform KNN-based procedure
	Undersampling: for classes with $N_j > m$, perform EM clustering-based procedure
	Outcome: balanced dataset

PCA ON REGULAR DATA



	Tot variance explained > 0.9
Number of components	33

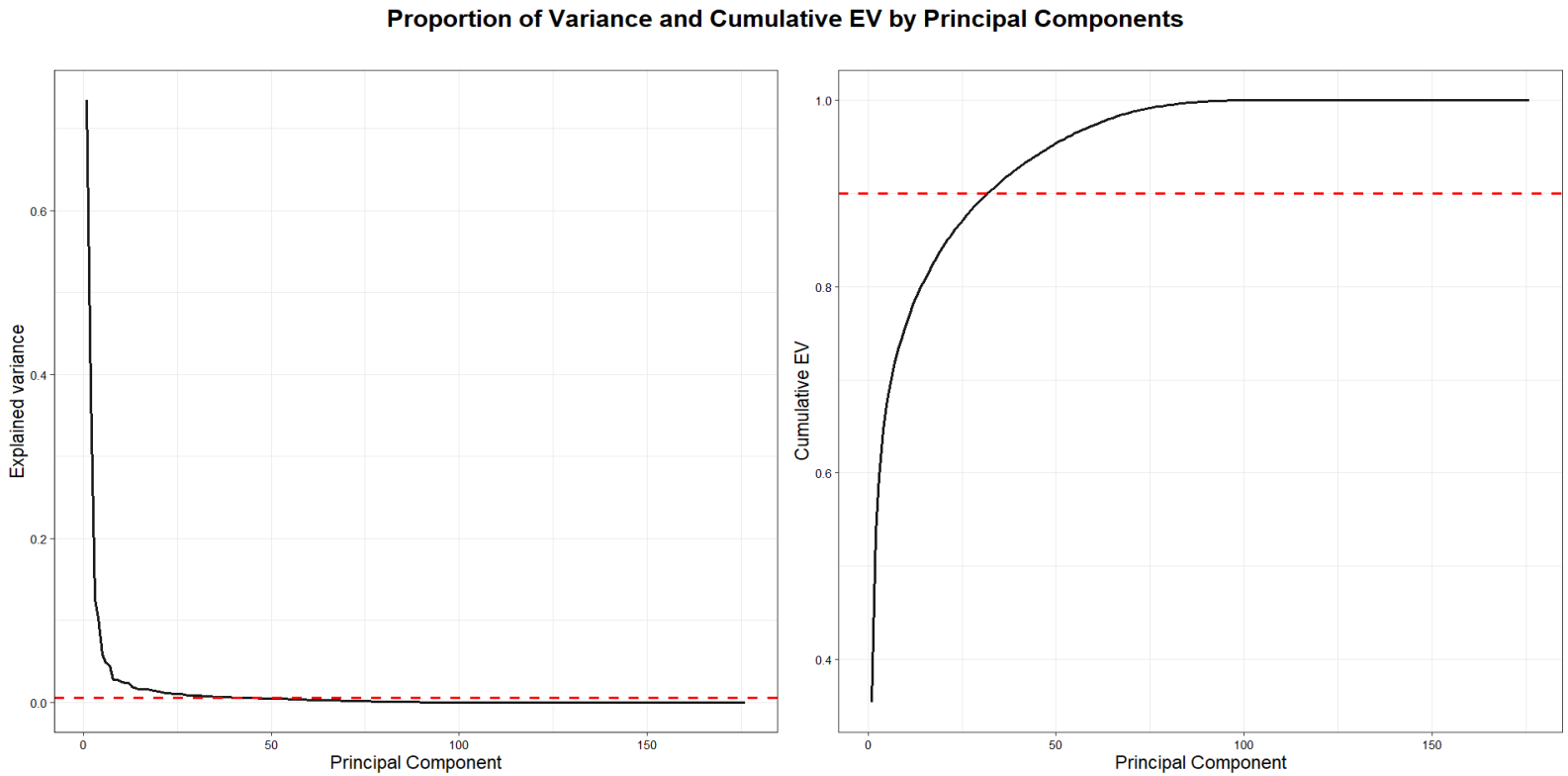
PCA ON AUGMENTED DATA



	Tot variance explained > 0.85
Number of components	30

	Tot variance explained > 0.9
Number of components	43

PCA ON SCUT DATA



	Tot variance explained > 0.9
Number of components	32

METRICS

- MISCLASSIFICATION ERROR RATE
- WEIGHTED ACCURACY
- WEIGHTED F1
- MATTHEWS CORRELATION COEFFICIENT (MCC)

MISCLASSIFICATION ERROR RATE

This metris, denoted as:

$$\frac{FP + FN}{TP + TN + FP + FN}$$

calculates the **proportion of misclassified instances** in a classification model.

It is commonly used to evaluate the overall performance of a classifier. However, it is **not ideal** for imbalanced classes because it tends to be biased towards the majority class.

WEIGHTED AVERAGED ACCURACY

It is calculated by averaging the accuracy of each class, weighted by the class frequency. It's represented as

$$\frac{1}{C} \sum_{i=1}^{C} w_i * Accuracy_i$$

Weighted accuracy provides a **more balanced evaluation** across classes compared to standard accuracy.

WEIGHTED AVERAGED F1

Weighted F1-score is computed by averaging the F1-score of each class, weighted by the class frequency.

$$\frac{1}{C}\sum_{i=1}^{C} w_i * F1_i$$

It provides a balanced evaluation across classes, taking into account both **precision and recall**.

While it gives higher importance to classes with more instances, it may still be influenced by the class distribution, potentially leading to an overestimation of performance on the majority class and an underestimation on the minority class.

MATTHEW CORRELATION COEFFICIENT

It's computed using the formula:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

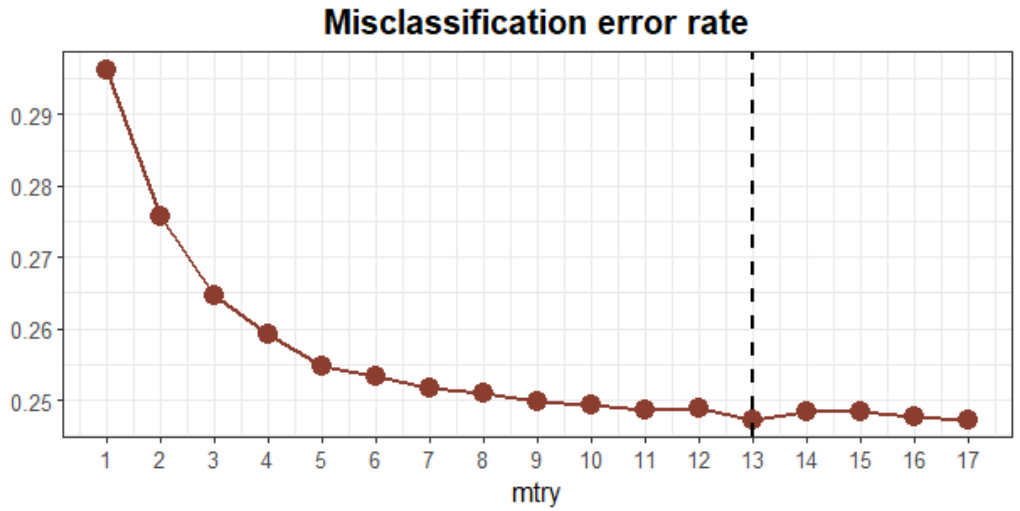
Its scale from -1 to 1 offers a clear interpretation: values close to 1 indicate strong agreement between predictions and observations, while values close to -1 suggest strong disagreement. Furthermore, MCC considers all four elements of the confusion matrix, providing a balanced assessment of model performance even in complex classification scenarios.

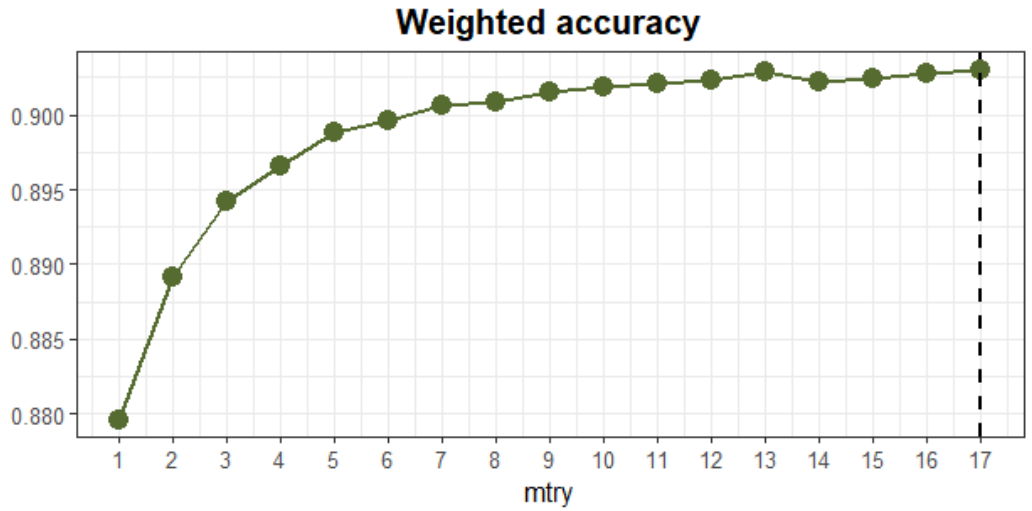
RANDOM FOREST

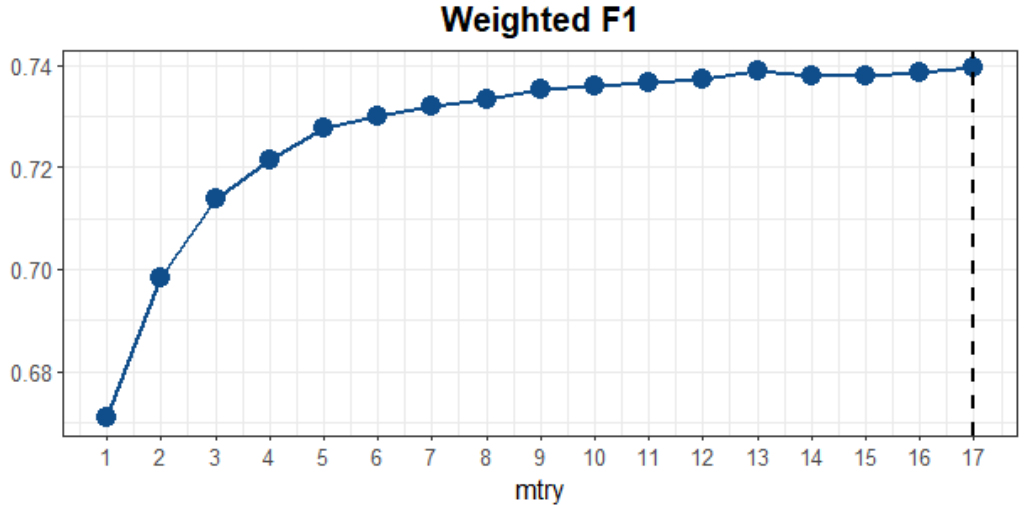
CHOICE NUMBER OF COVARIATES AT EACH SPLIT

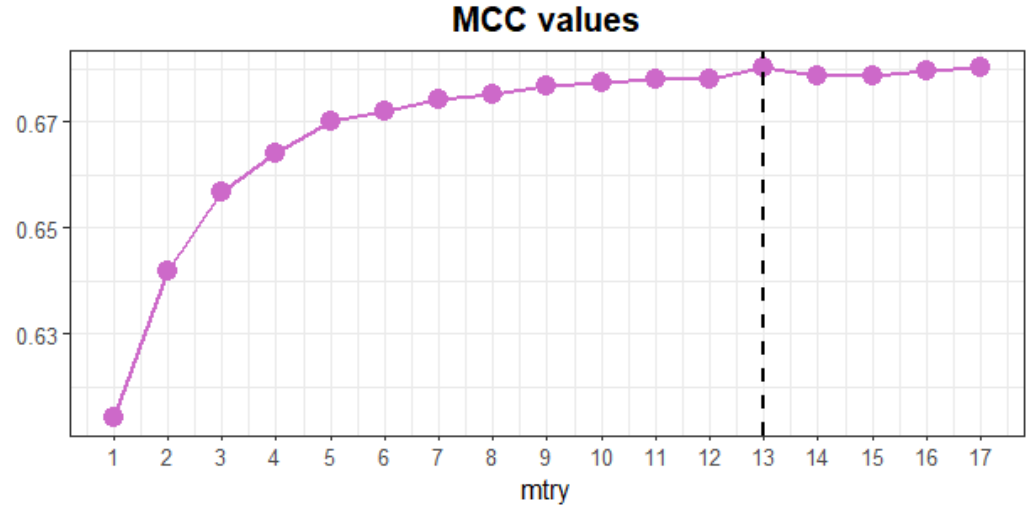
In the context of random forests, the parameter determining the number of covariates considered for each split at every branch is crucial for ensuring the model's effectiveness and generalization capability. By controlling this parameter, we regulate the diversity and complexity of individual trees within the forest. Selecting too few features can lead to overly simplistic trees that fail to capture the complexity of the underlying data, potentially resulting in underfitting. Conversely, employing too many features may lead to overfitting, where the model memorizes noise in the training data rather than learning meaningful patterns. Thus, tuning this parameter is essential to strike a balance between bias and variance, optimizing the model's performance on unseen data.

REGULAR DATASET

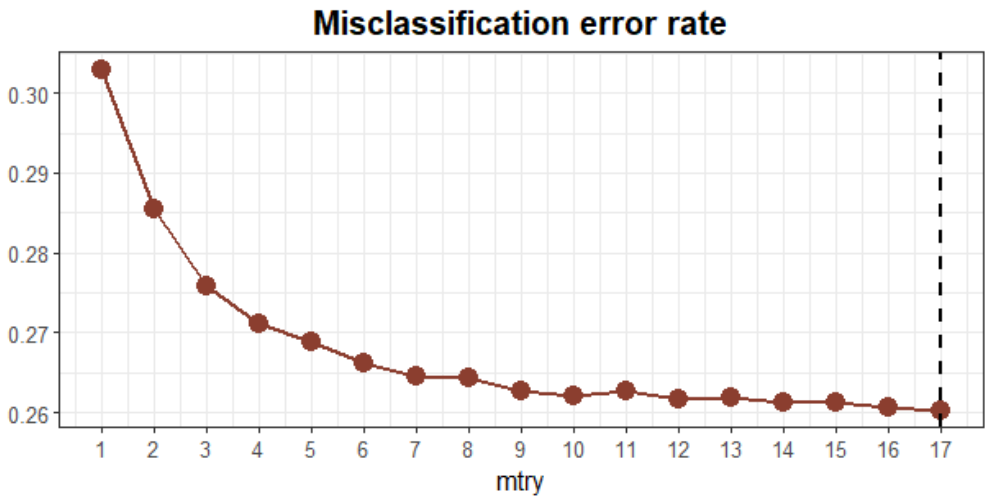


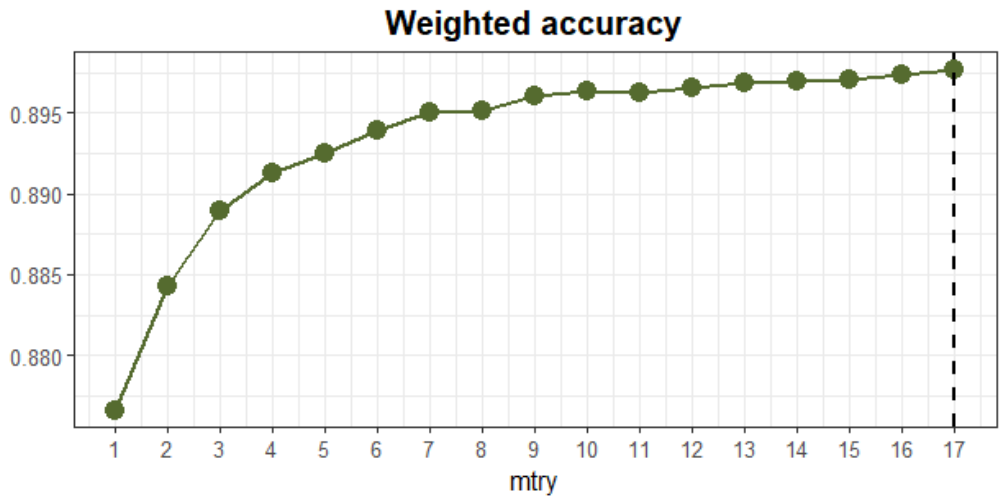


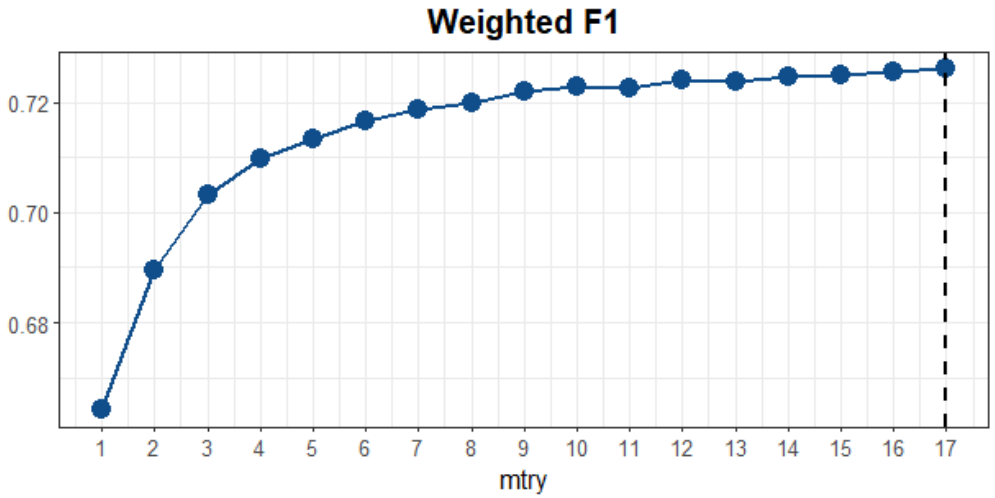


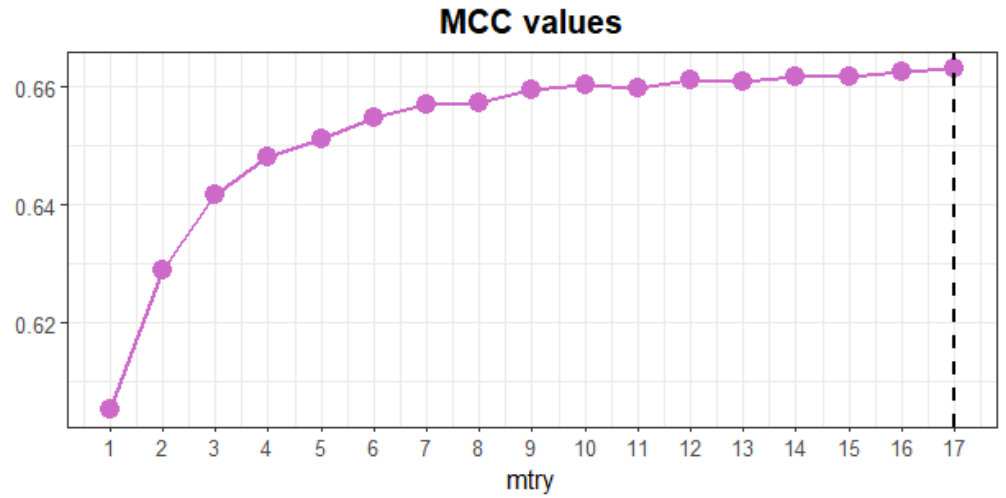


REGULAR DATASET WITH WEIGHT PARAMETER

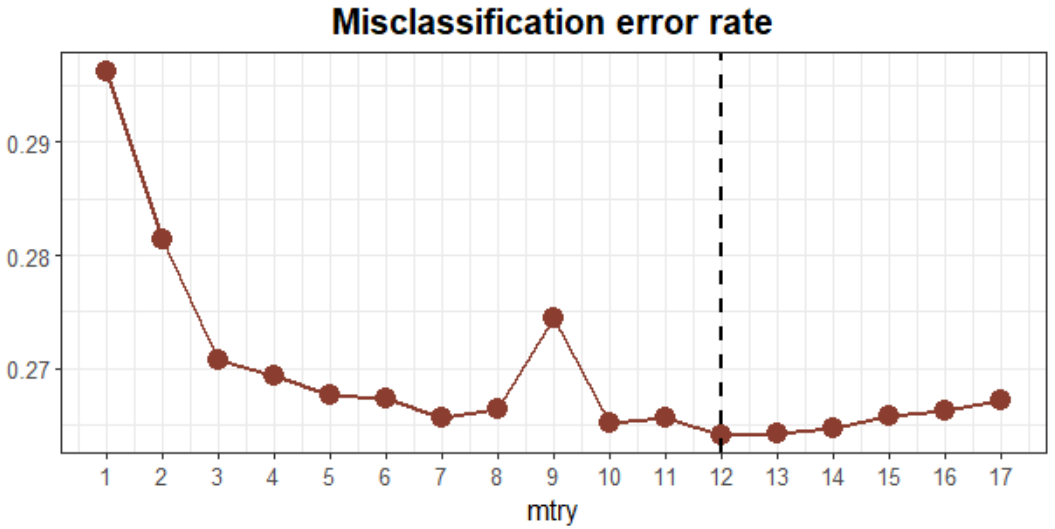


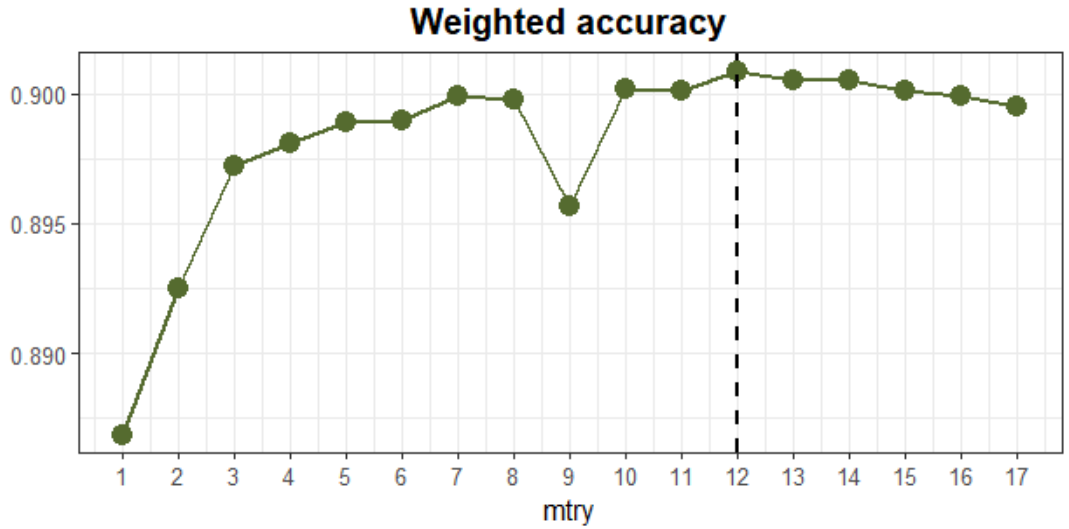


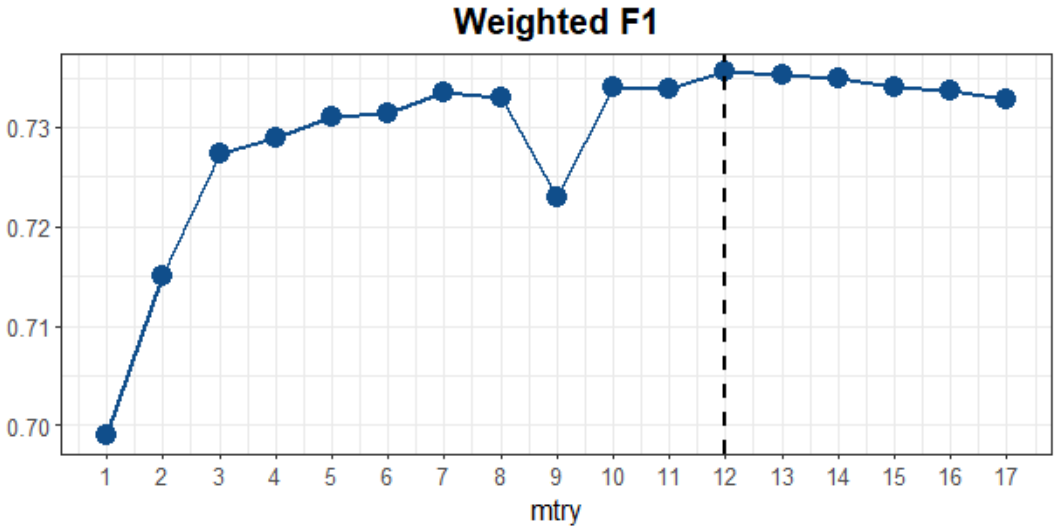


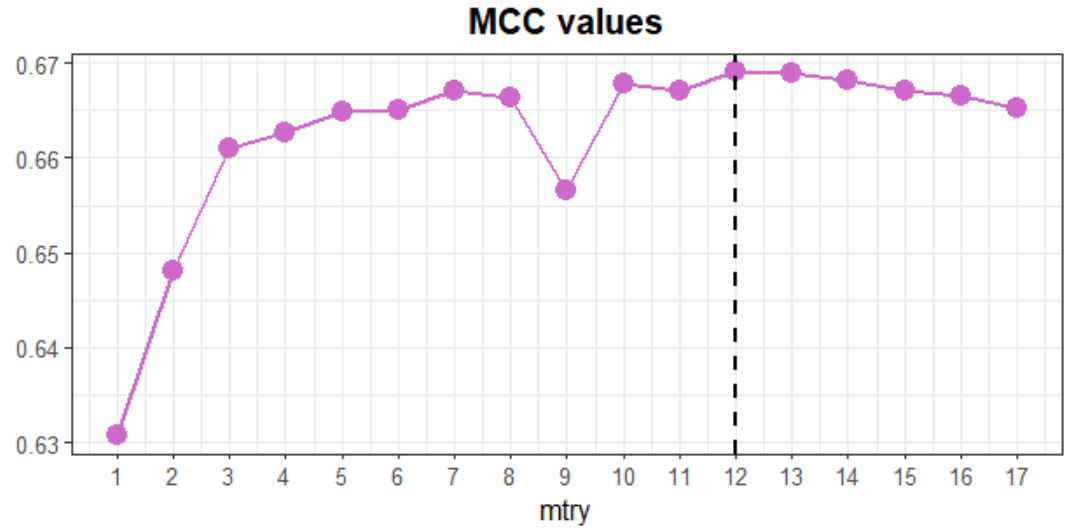


SMOTE AUGMENTED DATASET

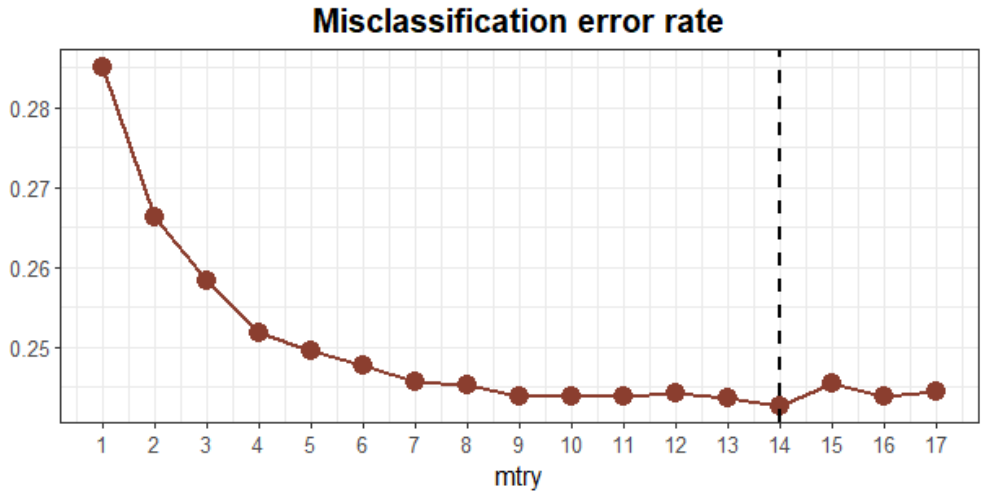


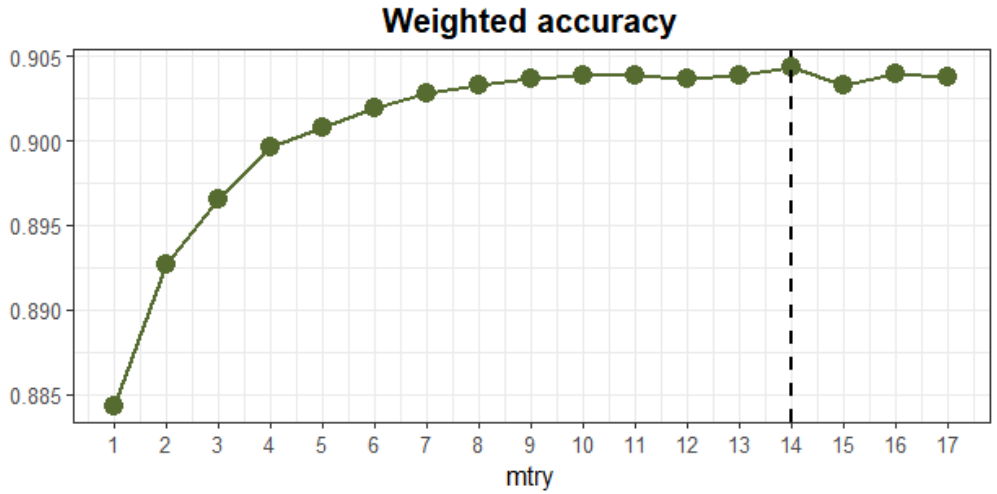


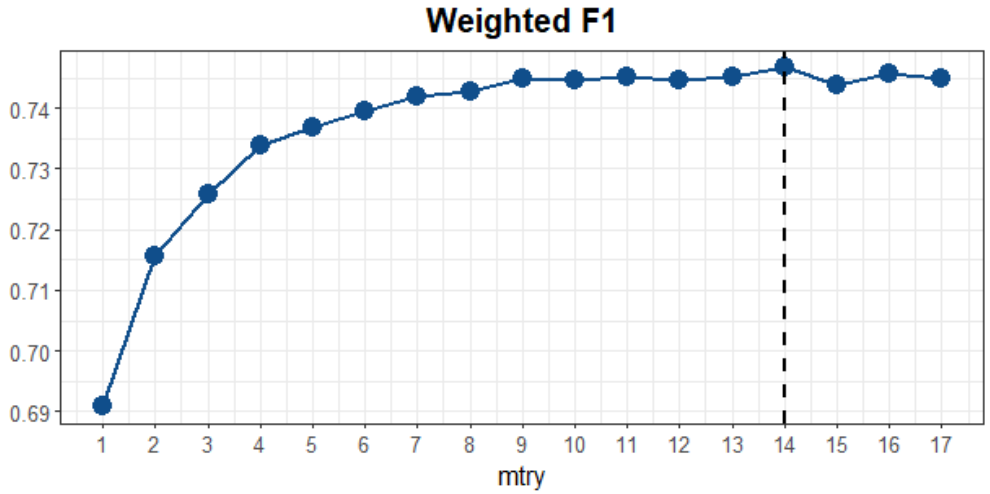


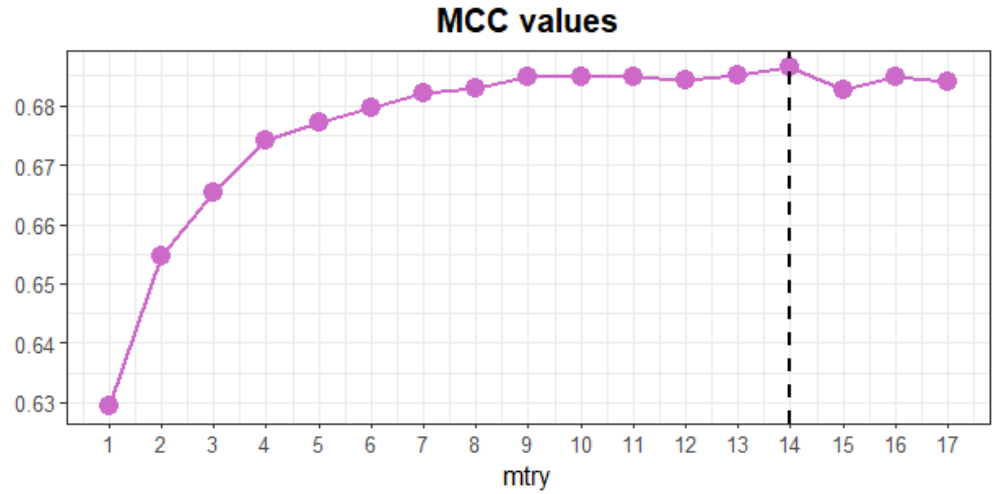


AUGMENTED DATASET



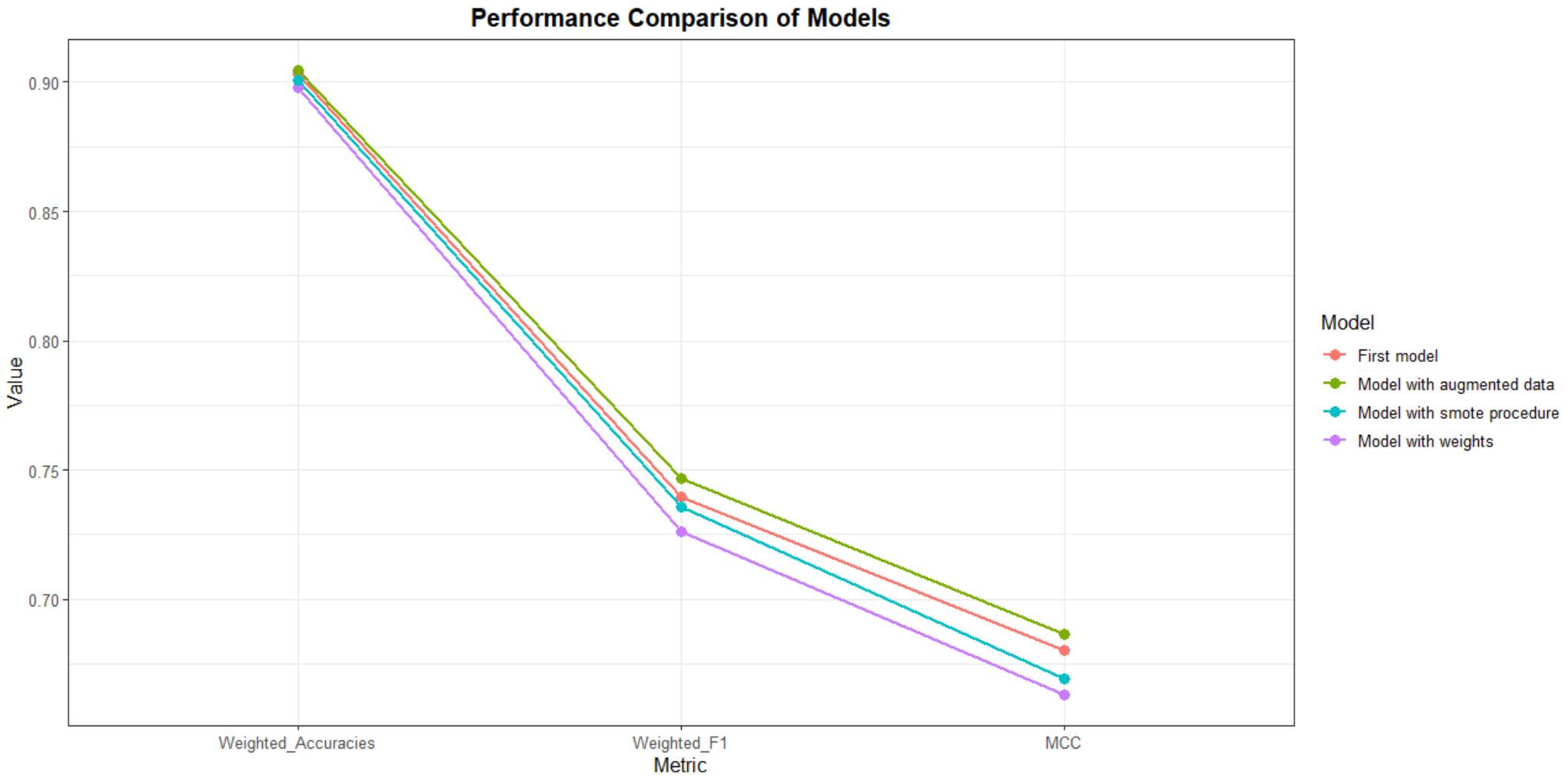






CHOSEN PARAMETER SPECIFICATION: 9

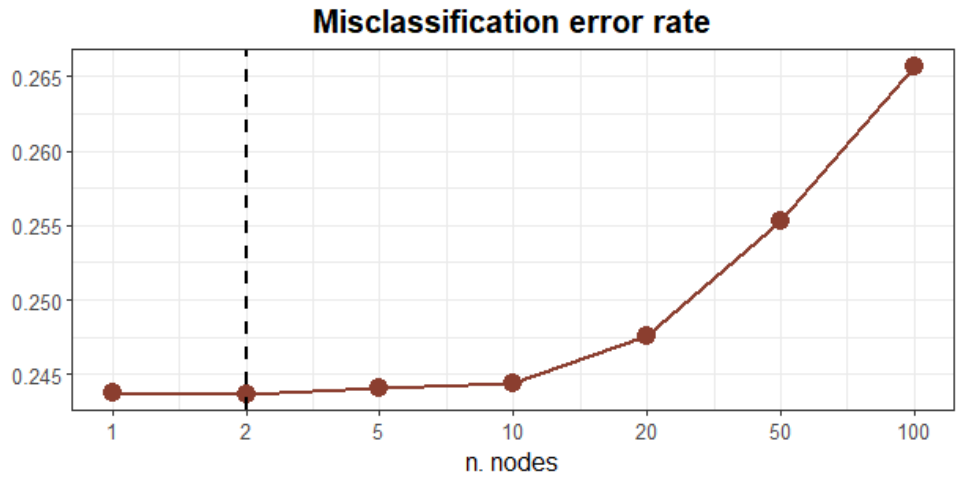
FINAL COMPARISON

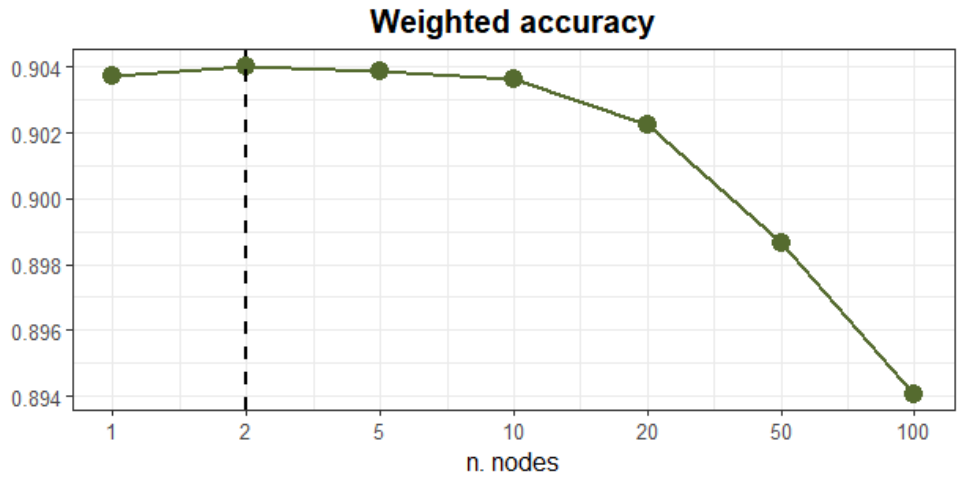


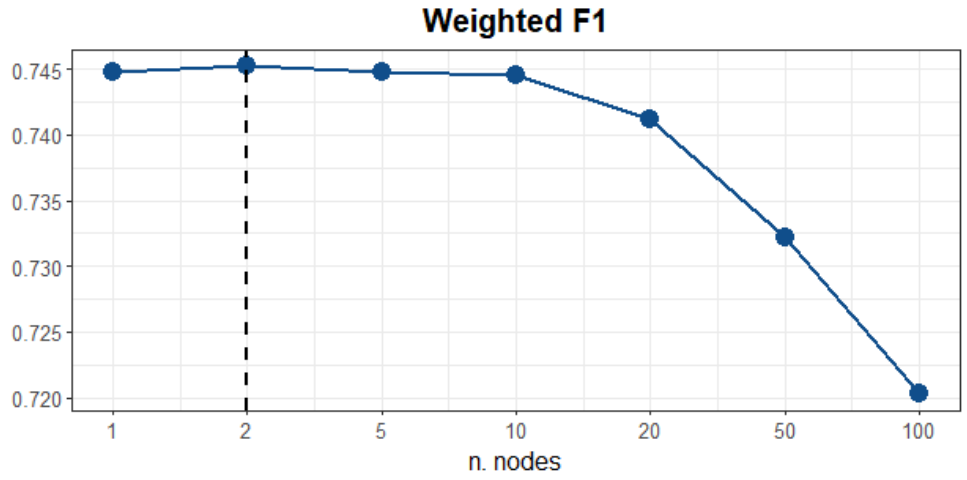
OPTIMIZING NUMBER OF NODES

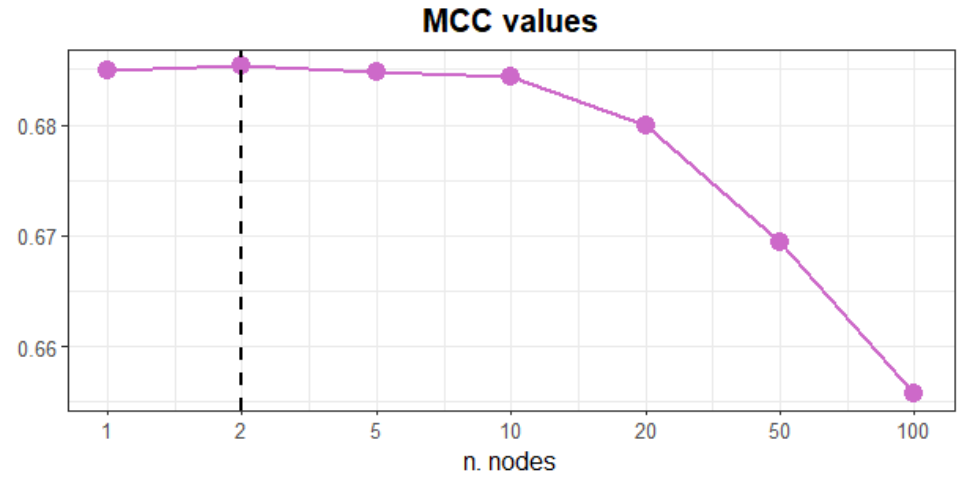
A shallow tree with few nodes may oversimplify the relationships within the data, leading to high bias and poor predictive performance. Conversely, an excessively deep tree with many nodes can capture intricate patterns in the training data, potentially resulting in overfitting and reduced generalization to unseen data. Therefore, tuning the parameters related to the number of nodes is paramount to strike a balance between model complexity and performance.

PARAMETER CHOICE





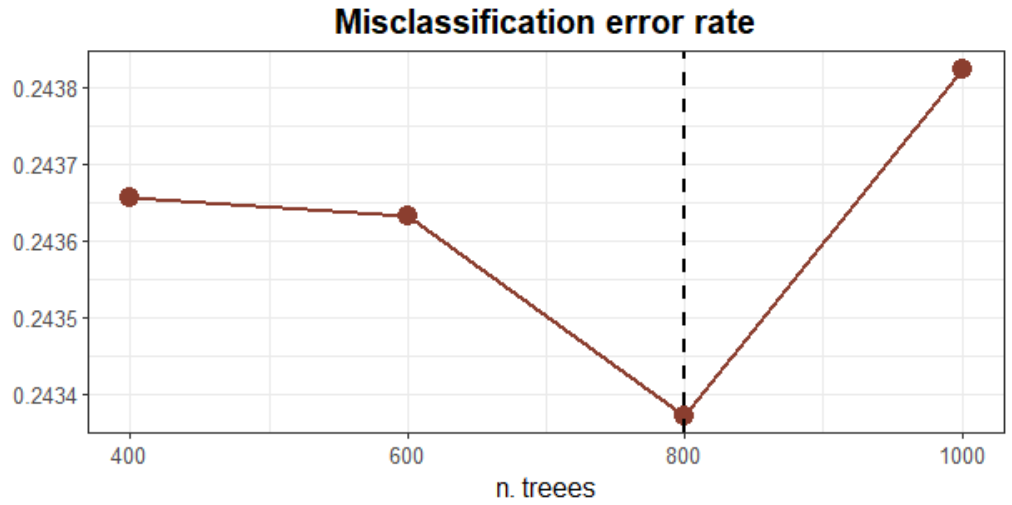


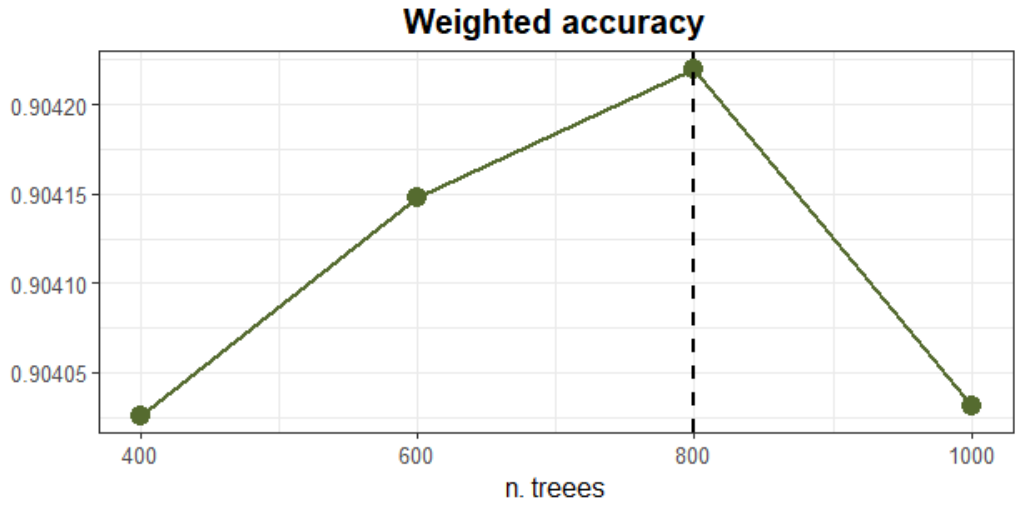


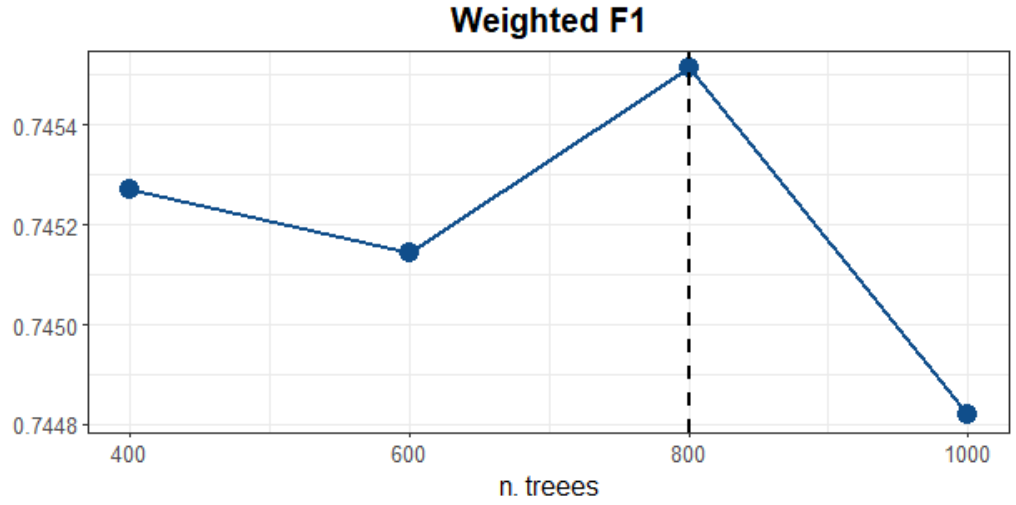
OPTIMIZING NUMBER OF TREES

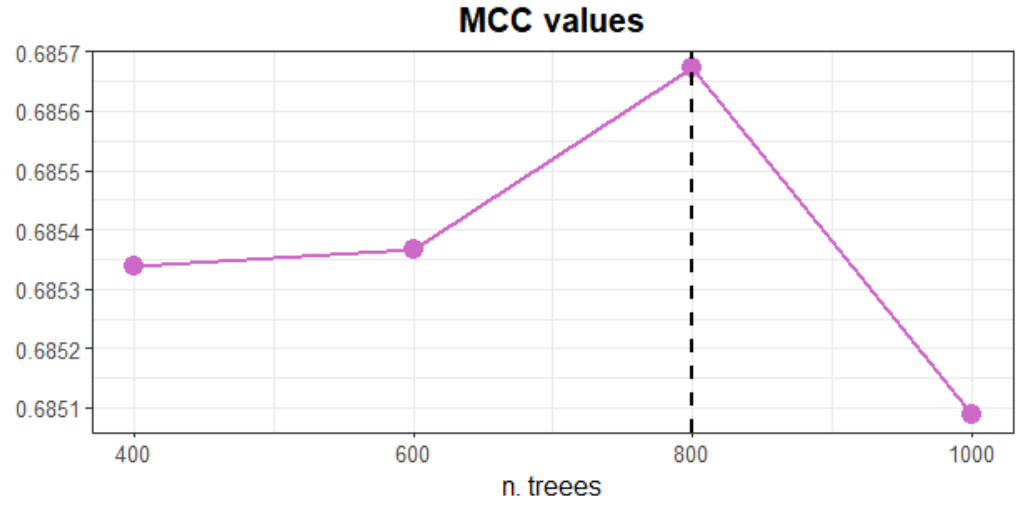
Each tree contributes to the overall prediction, and increasing the number of trees typically improves the model's accuracy and generalization performance. However, adding more trees also comes with computational costs, as training and inference times increase proportionally. Additionally, there's a point of diminishing returns, where the marginal benefit of adding more trees diminishes, and the model's performance plateaus. Hence, optimizing the number of trees is essential.

SELECTING BEST MODEL









BEST MODEL PERFORMANCES

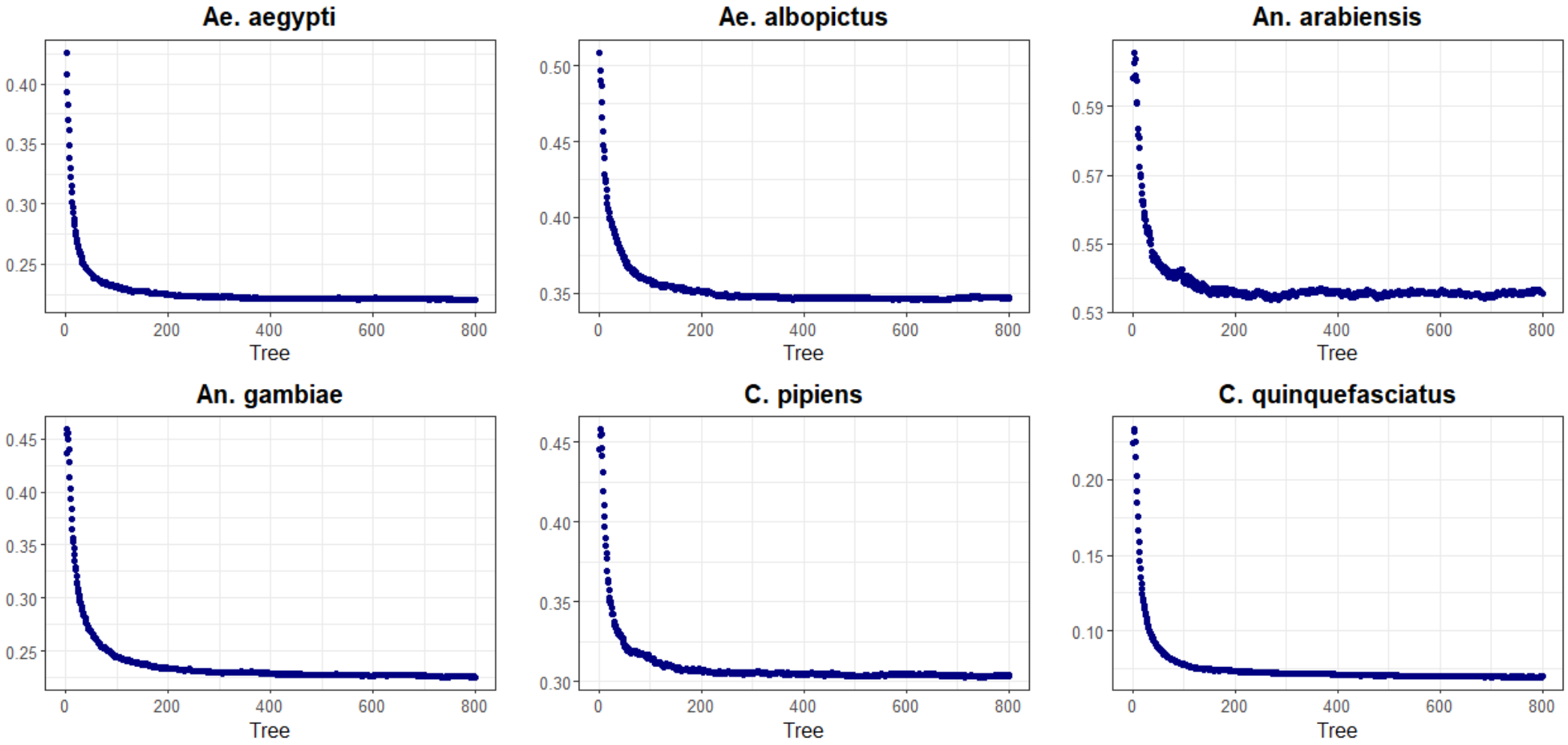
List of contents:

- Confusion matrix
- OOB times distribution
- Treesize distribution
- Error rate for class
- Frequency of predicted classes
- PC importance for class

CONFUSION MATRIX

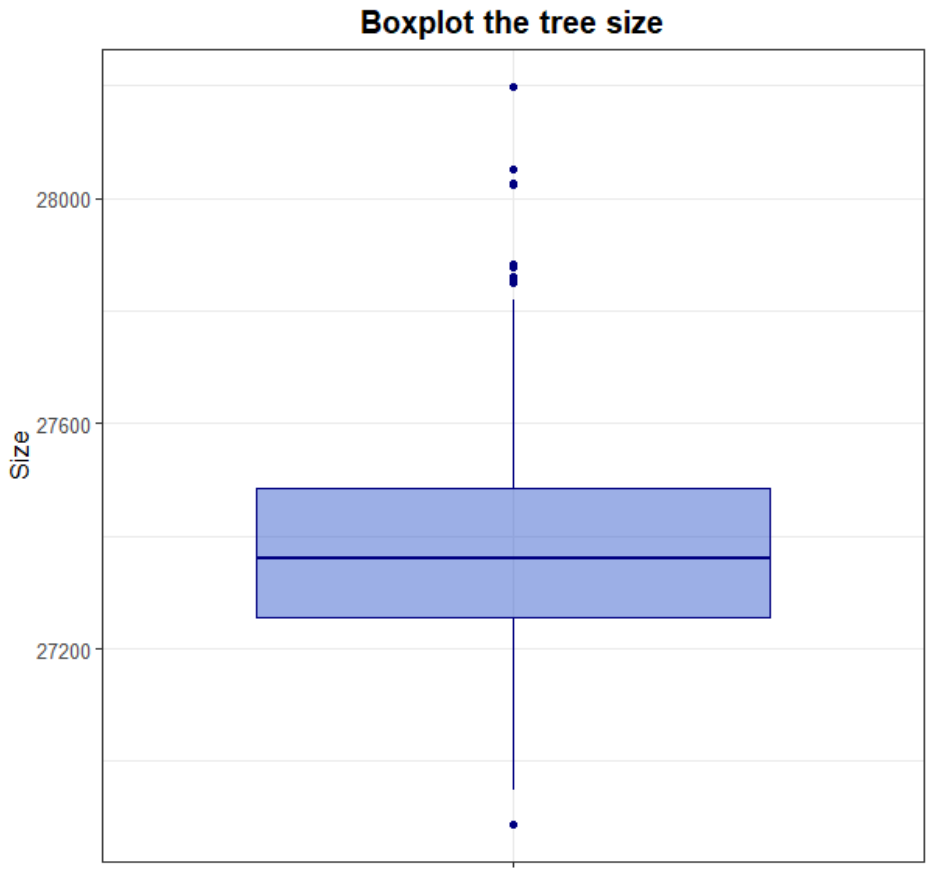
	Yellow Fever	Asian Tiger	Gambiae	Arabiensis	Northern House	Southern House
Yellow Fever	10014	494	476	927	1100	591
Asian Tiger	101	1680	216	352	6	17
Gambiae	161	66	696	250	33	86
Arabiensis	533	671	971	5686	111	28
Northern House	555	93	166	1273	3207	21
Southern House	1469	31	370	79	105	10447
Class error	**0.21966**	**0.44645**	**0.75958**	**0.23379**	**0.29701**	**0.06639**

ERROR RATE PER CLASS

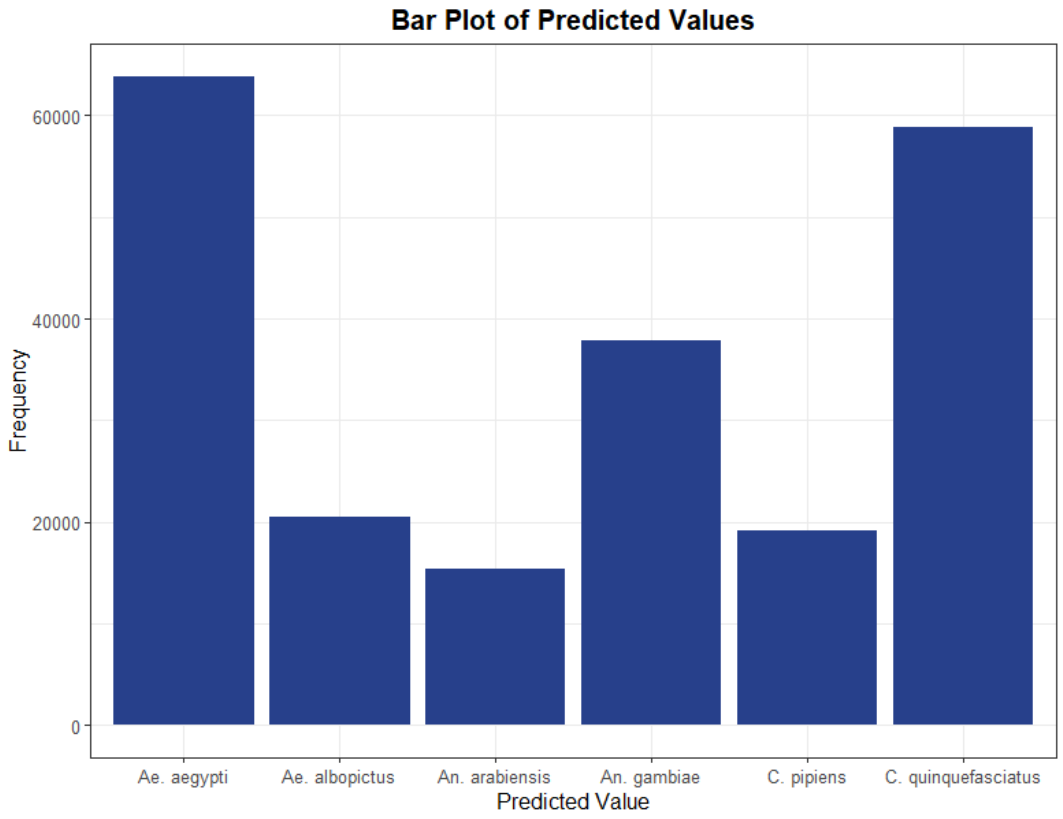


OOB times

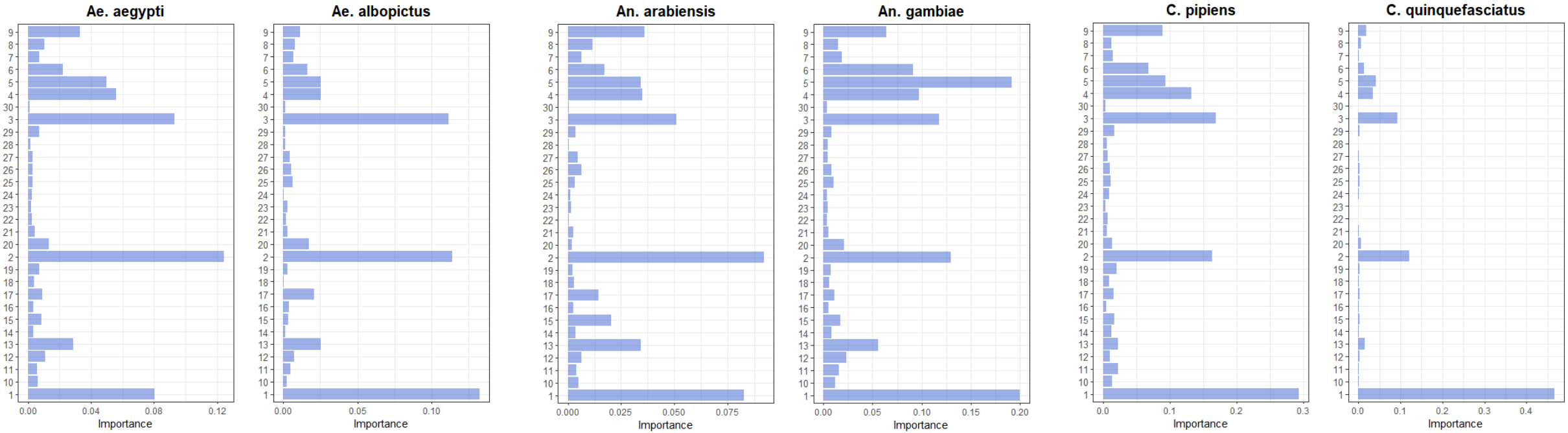
Treesize



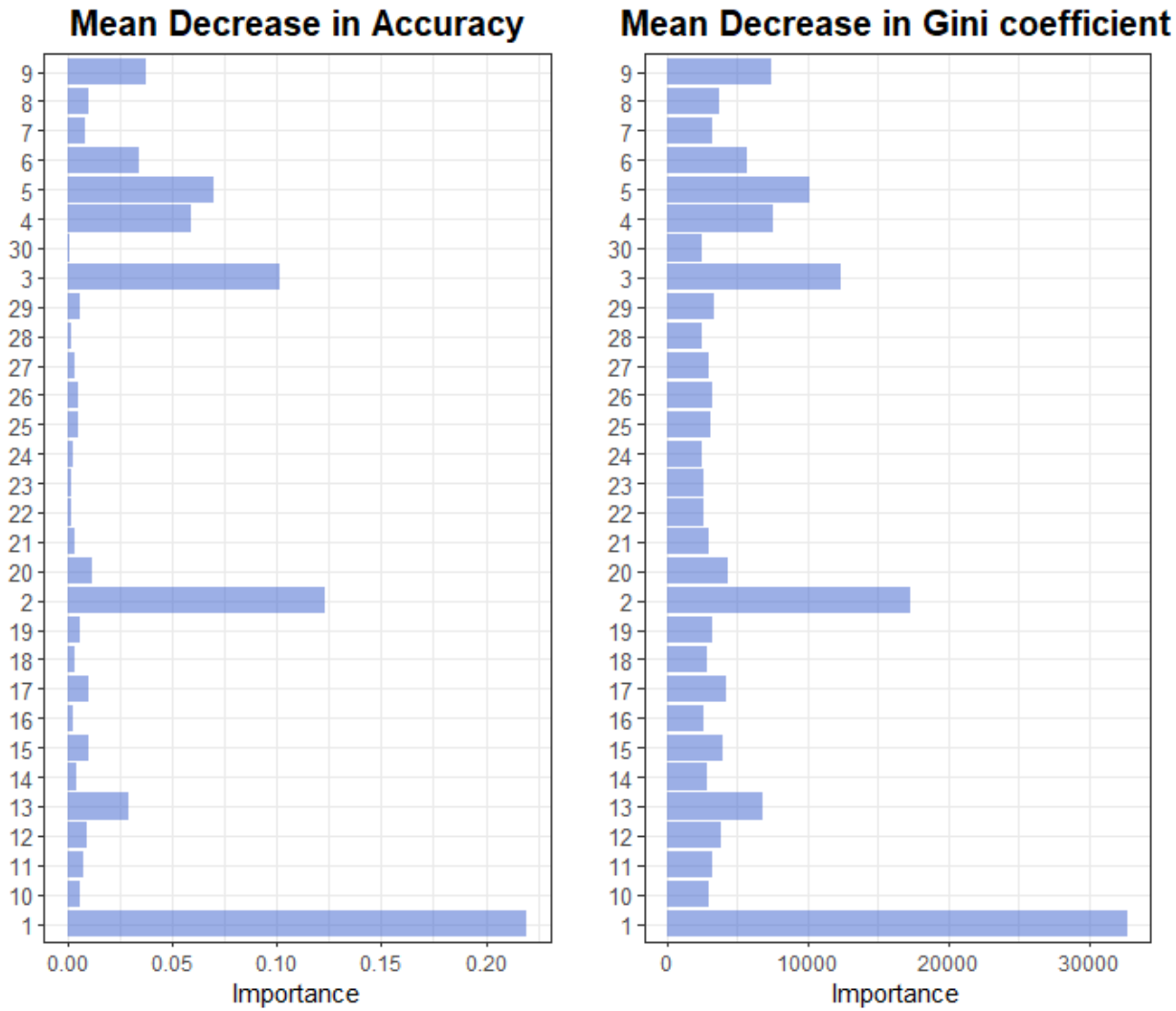
FREQUENCY OF PREDICTED CLASS



PC IMPORTANCE FOR EACH CLASS



MEAN DECREASE IN ACCURACY AND GINI COEFFICIENT BY COMPONENT



FINAL METRICS

MISCLASSIFICATION ERROR RATE	WEIGHTED ACCURACY	WEIGHTED F1	MCC VALUE
0.24337	0.90422	0.74551	0.68567

BOOSTING ENSEMBLE LEARNING

ADVANTAGES OF THE XGBOOST ALGORITHM

Gradient tree boosting

Introducing a differentiable objective function that is minimized at each iteration using gradient descent

Shrinkage

Scaling the newly added weights provides space for future trees to improve the model

Regularisation

L1 and L2 penalization in the optimization task to prevent overfitting

Splitting procedure

Approximate algorithm implemented involving sampling across percentiles of the feature distribution

Sparsity-aware algorithm

Algorithm suitable for sparse datasets, for example, with missing values or frequent zero entries

Computational advantages

Parallel computations using block structures in the system design, along with cache-aware prefetching

PARAMETER TUNING

η Learning rate

The contribution of each tree when we add it to the model.

Values:
{0.001, 0.01, 0.3, 0.5}

γ Minimum loss reduction

Requirement on the mimimum multi-class cross-entropy reduction for further splits

Values:
{1, 3}

d Maximum depth

Controls complexity of the model in terms of the maximal depth allowed for each split

Values:
{2, 5}

B Number of iterations

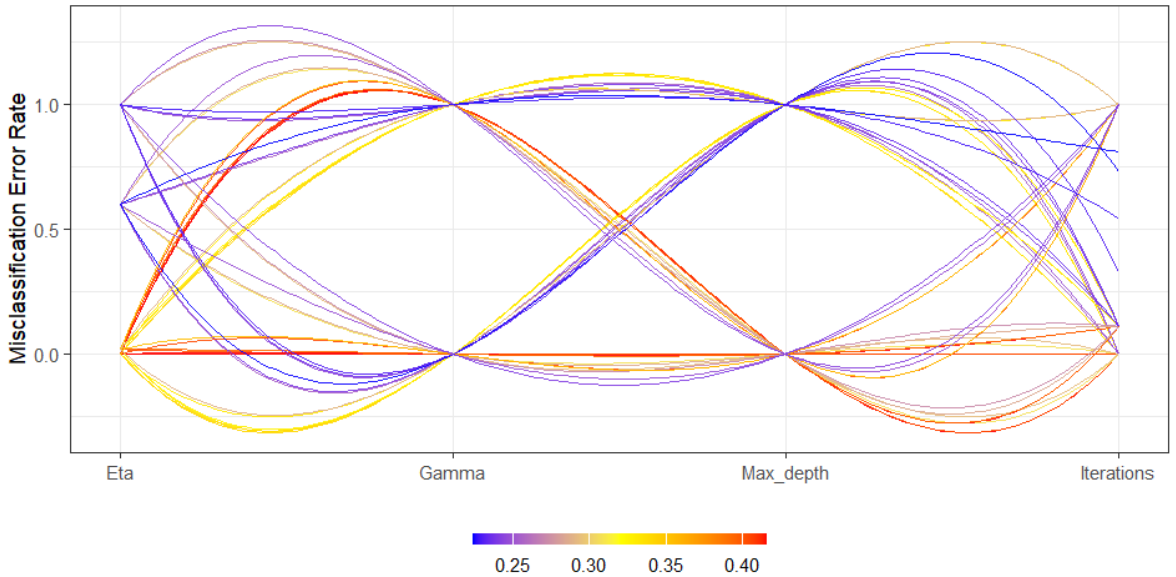
Maximum number of boosting iterations. Early stopping criterion imposed.

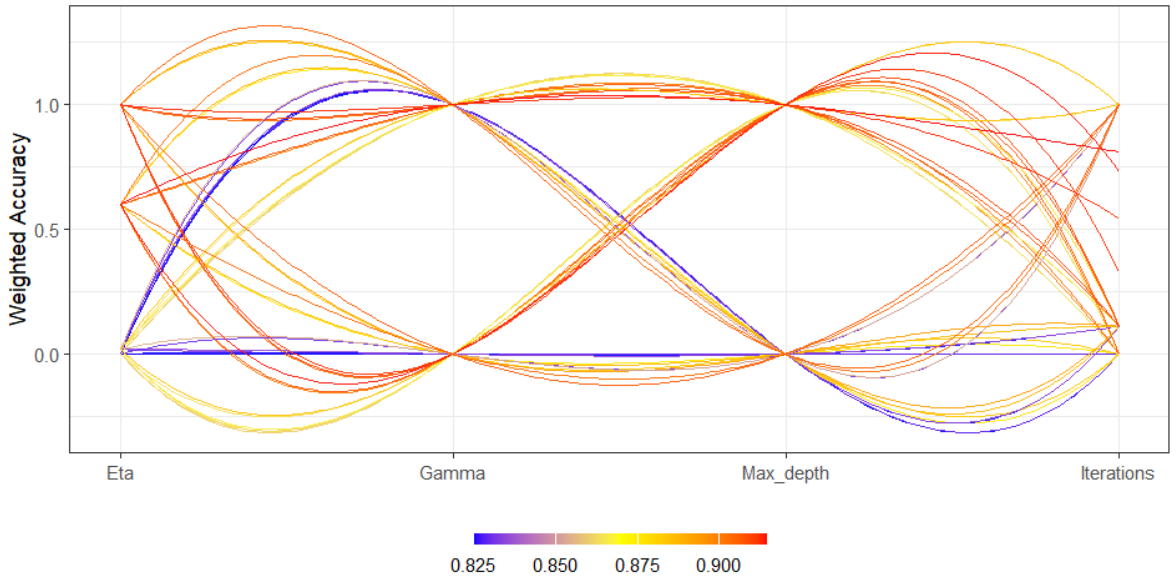
Values:
{50, 100, 500}

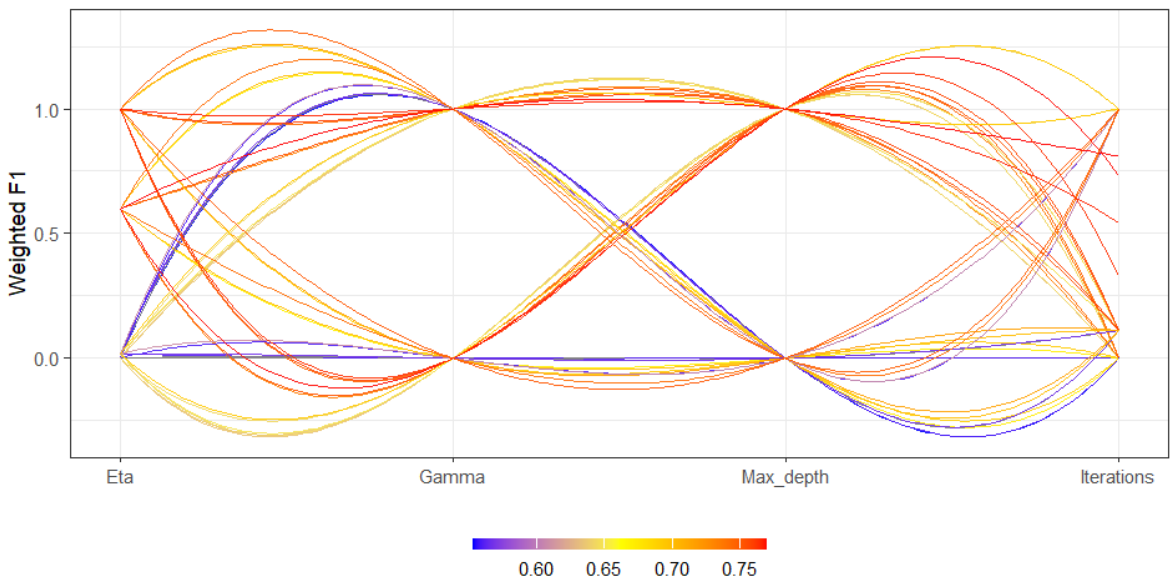
Other parameters

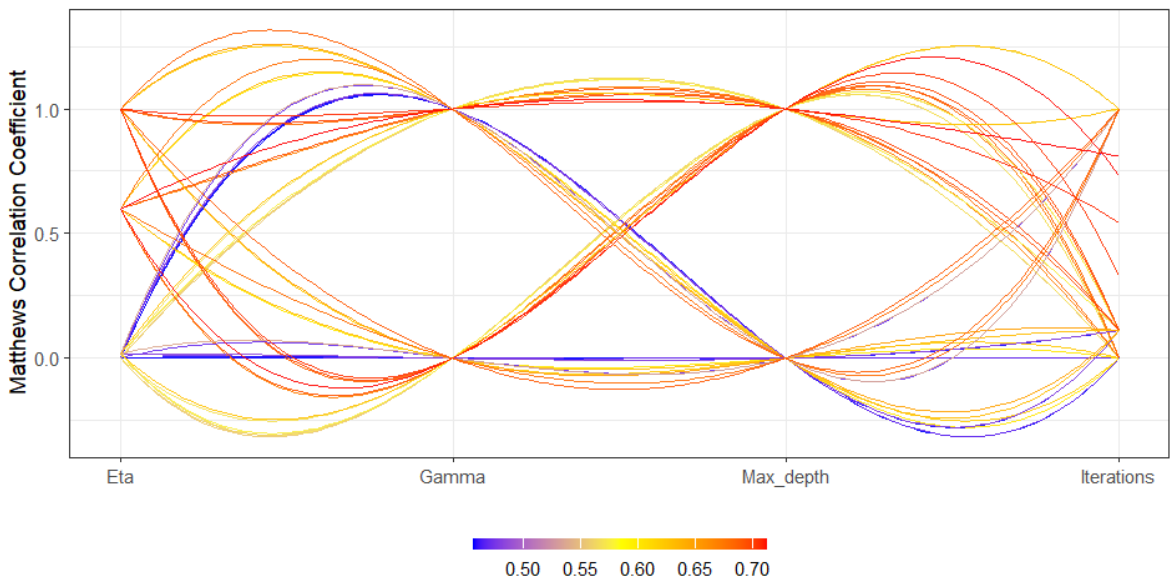
- Booster = 'gbtree'
- Subsample = 0.75
- $\lambda = 1$
- $\alpha = 0$
- objective="multi: softprob"
- nthread = 2

Parallel coordinates plots for XGBOOST, original data



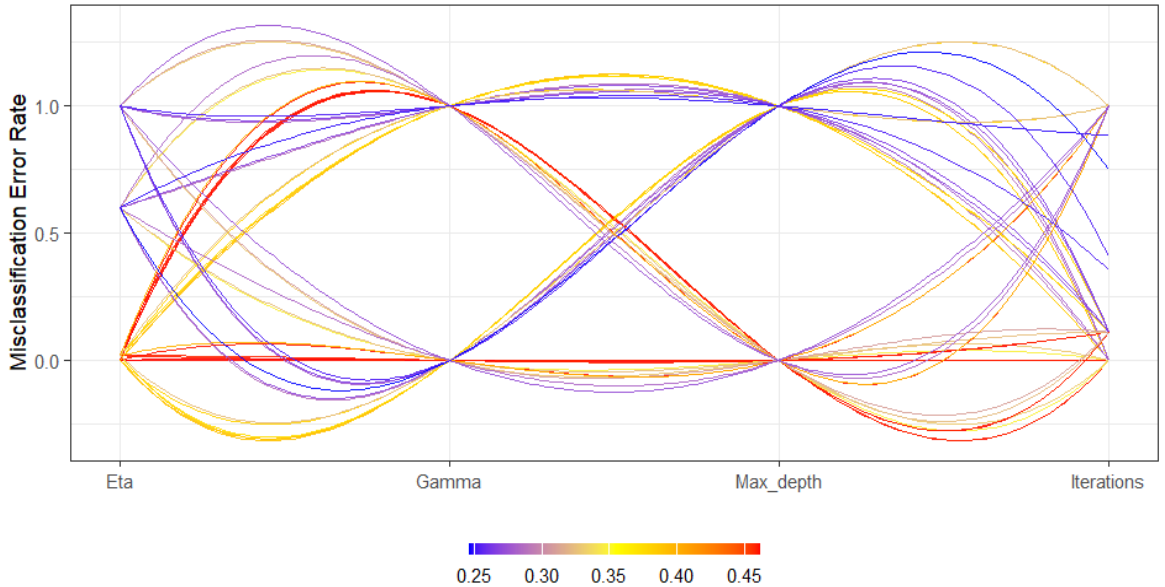


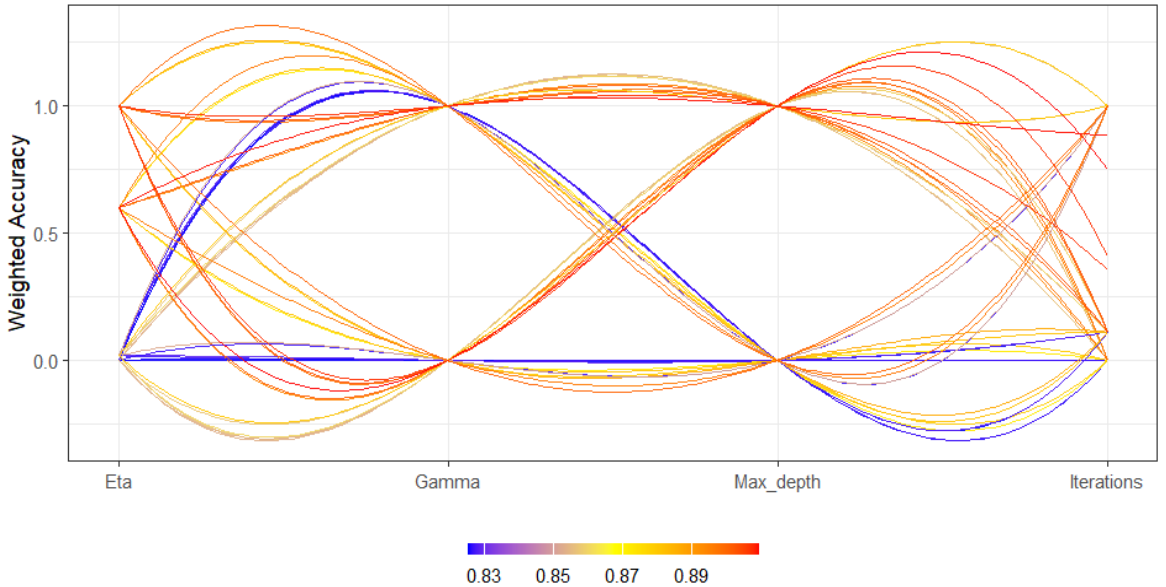


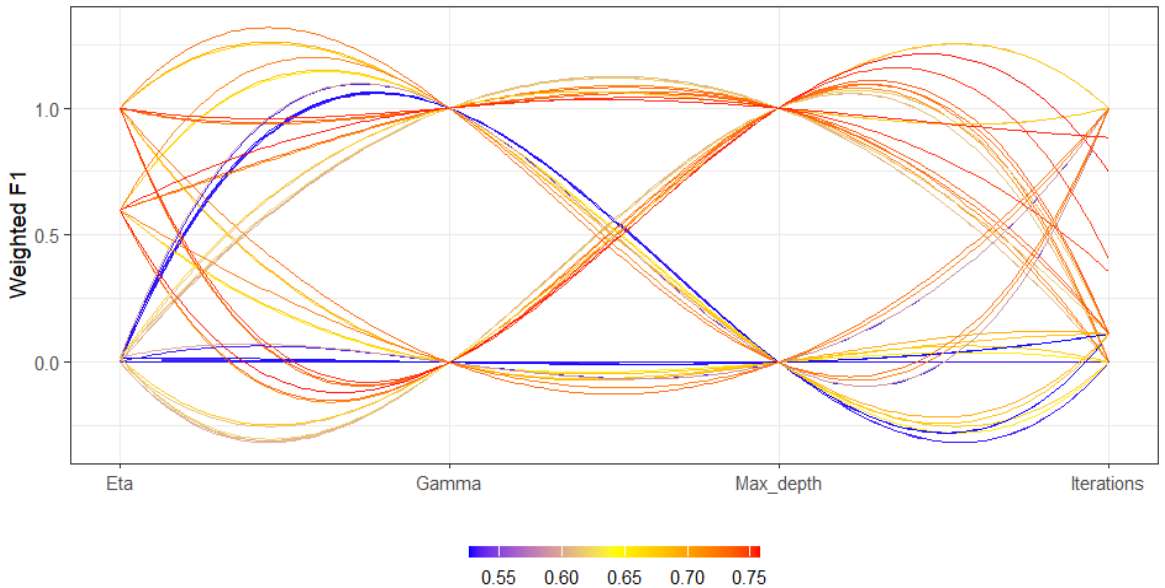


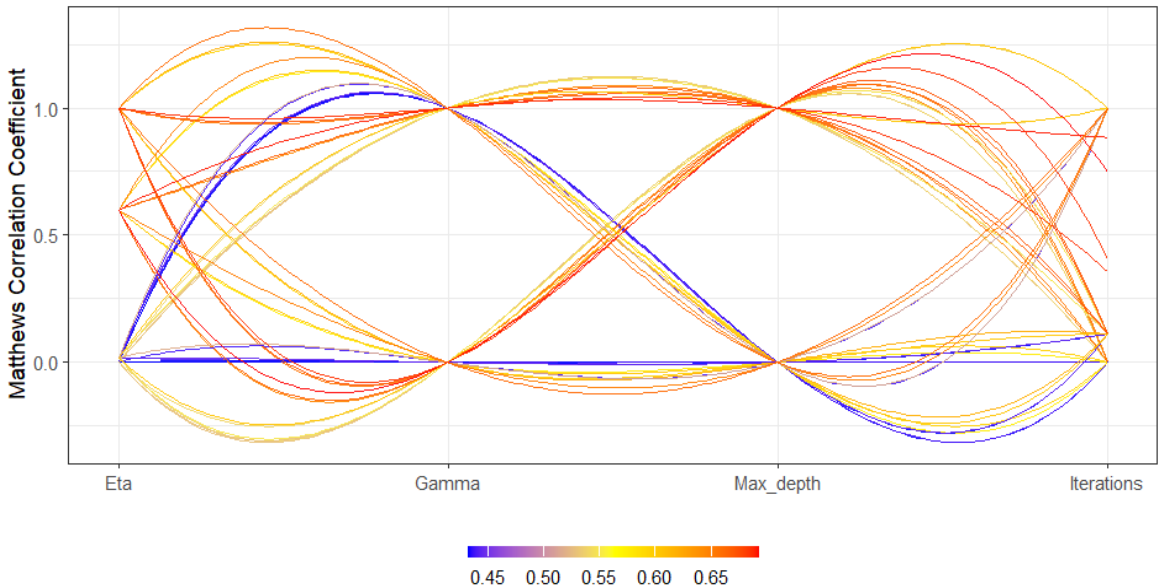
OPTIMAL MODEL SPECIFICATION: $\eta = 0.3$, $\gamma = 3$, $d = 5$, $B = 415$

Parallel coordinates plots for XGBOOST, SCUT



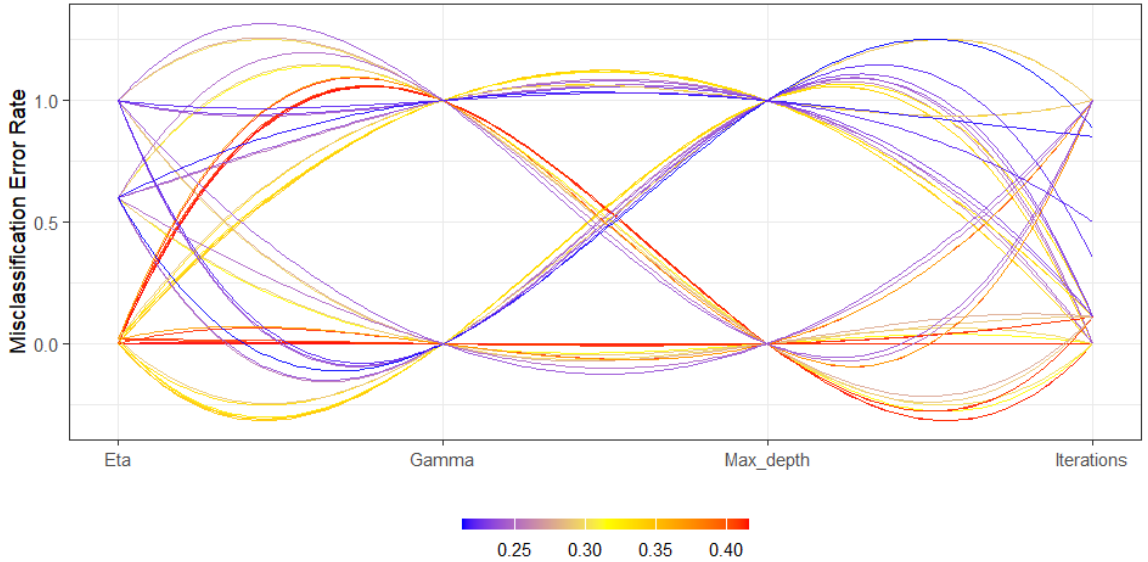


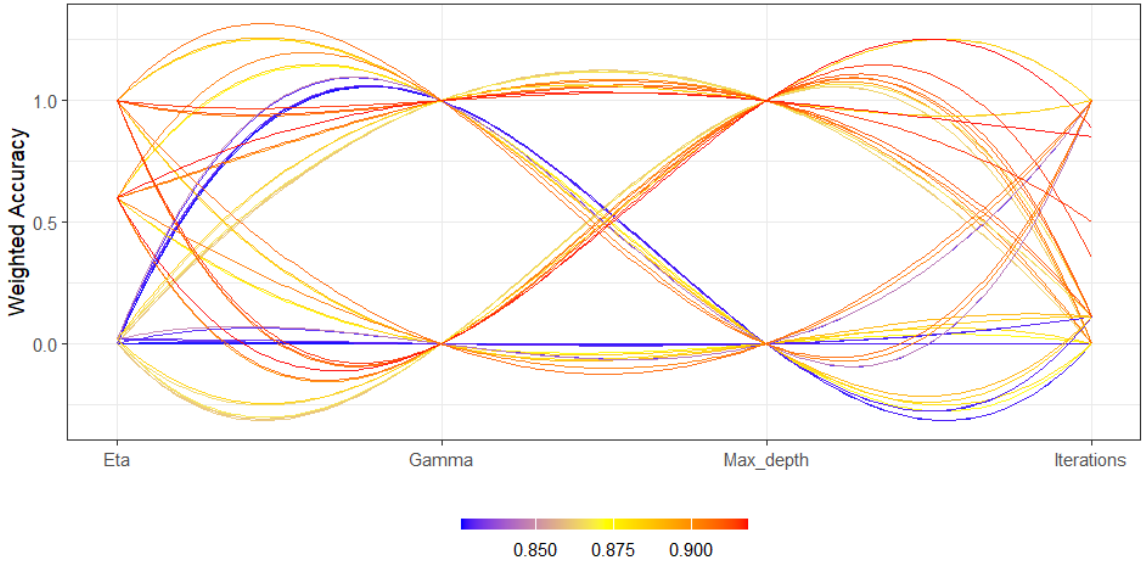


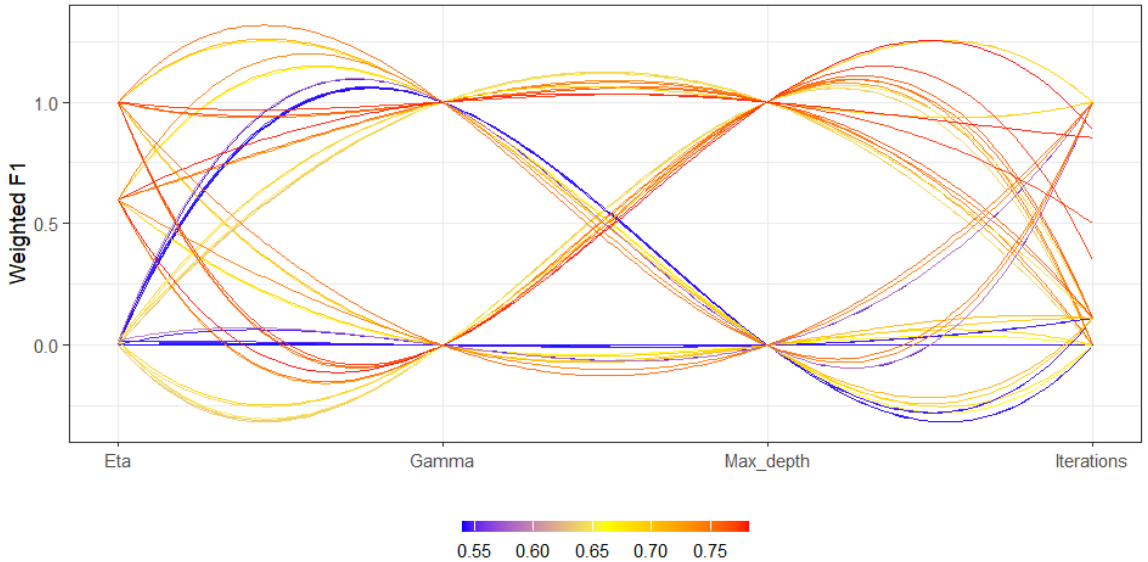


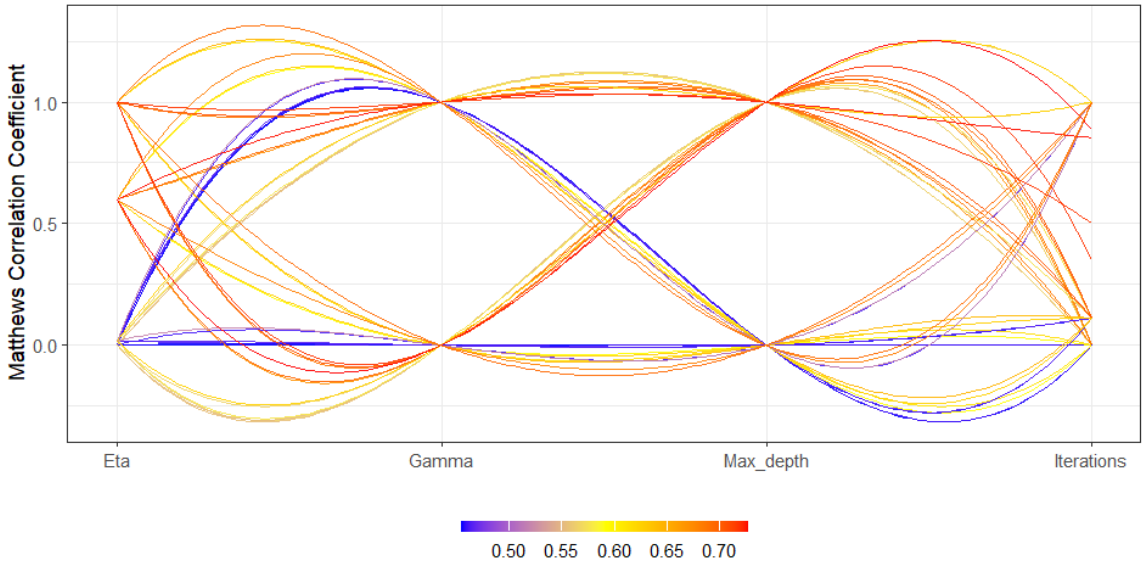
OPTIMAL MODEL SPECIFICATION: $\eta = 0.3$, $\gamma = 1$, $d = 5$, $B = 387$

Parallel coordinates plots for XGBOOST, data augmentation





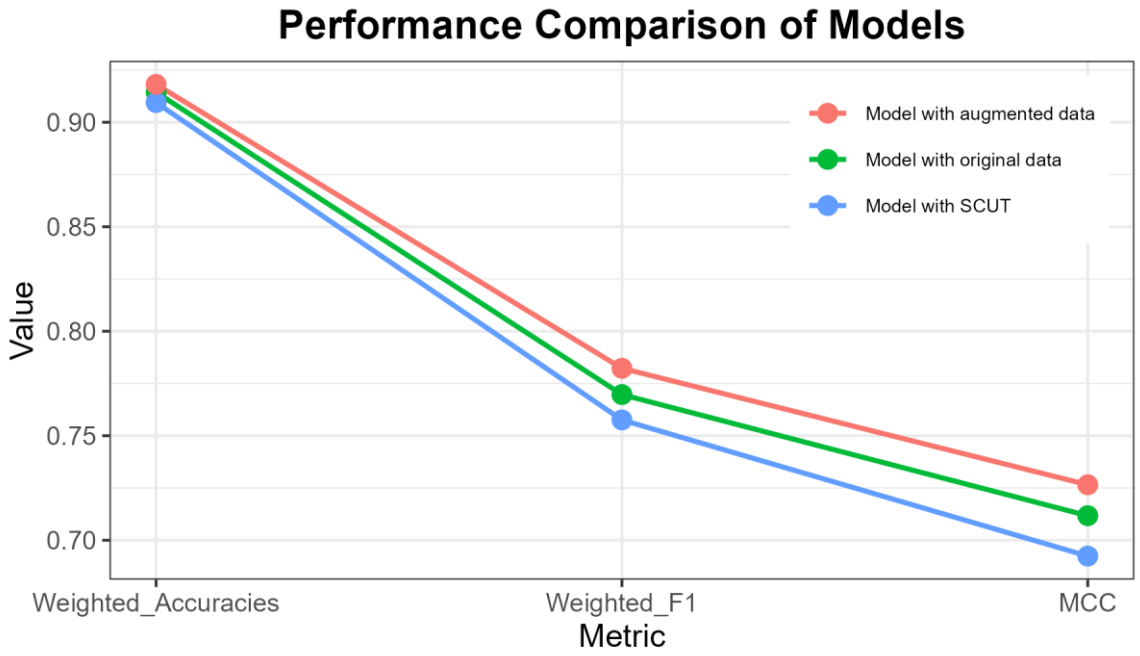


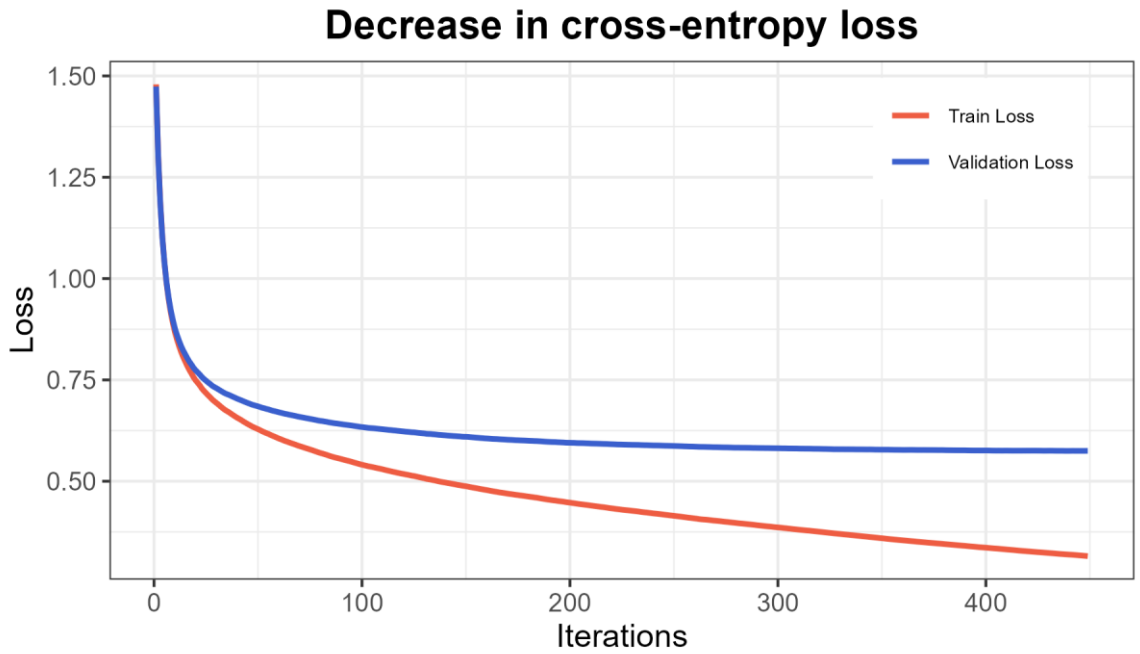


OPTIMAL MODEL SPECIFICATION: $\eta = 0.3$, $\gamma = 3$, $d = 5$, $B = 415$

OPTIMAL BOOSTING MODEL

- The selected model was fitted on augmented dataset with synthetically generated audio data instances
- The learning rate parameter of 0.3 provided best performance, allowing each tree to have quite substatial contribution
- Lower regularisation restriction provided better fit
- Early stopping was triggered, with 415 iterations providing optimal validation error





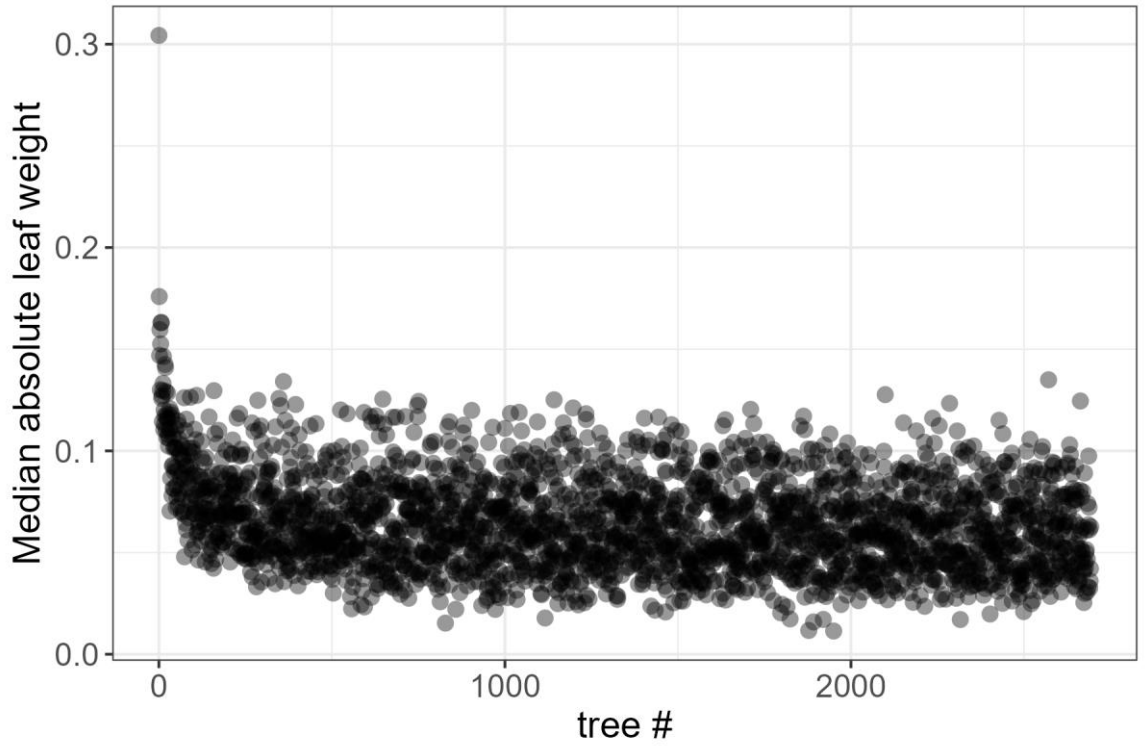
COMPARISON: RANDOM FOREST AND BOOSTING

METRIC	MISCLASSIFICATION ERROR RATE	WEIGHTED ACCURACY	WEIGHTED F1	MATTHEW'S CORRELATION
Random Forest	0.24337	0.90422	0.74551	0.68567
Boosting	0.21266	0.91812	0.78226	0.72654

The boosting ensemble learning model provided a better fit for the training data in terms of the analysed evaluation metrics.

OPTIMAL BOOSTING MODEL

- Final evaluation metrics demonstrate good generalisation capability of the model on the test data with class imbalance



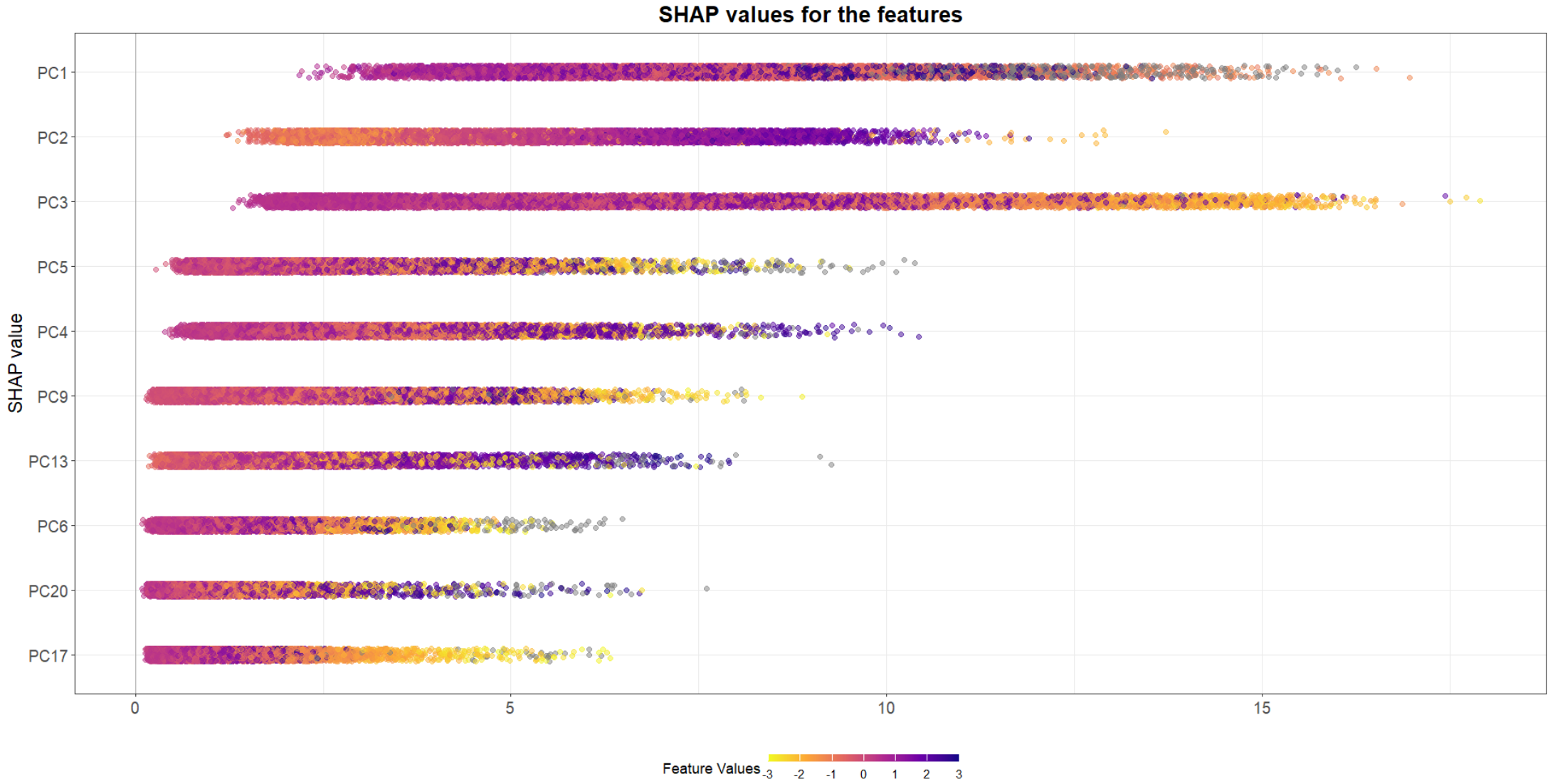
Misclassification Error	Weighted Accuracy	Weighted F1	Matthew's Correlation
0.21039	0.91928	0.78495	0.72957

CONFUSION MATRIX

	Yellow Fever	Asian Tiger	Gambiae	Arabiensis	Northern House	Southern House
Yellow Fever	10278	363	396	616	792	604
Asian Tiger	214	1927	182	285	32	22
Gambiae	243	159	1094	420	60	125
Arabiensis	466	472	796	5932	63	35
Northern House	642	83	109	106	3527	52
Southern House	989	30	317	61	88	10351
Class error	**0.19903**	**0.36486**	**0.62198**	**0.20054**	**0.22687**	**0.07489**

The most common mosquitos are classified with low misclassification error.
The error is higher for smaller classes, especially for Anopheles Gambiae mosquito.

RELATIVE CONTRIBUTION OF THE PRINCIPAL COMPONENTS FOR THE MODEL, WITH CLUSTER STRUCTURE



SHAP values provide an easily interpretable way to observe the contribution of each feature to the predictions made by the model. It leverages additivity of the contributions, treating them independently, as well as local accuracy, by quantifying the differences between observed and fitted values.

It can be observed that the 10 selected principal components contribute positively towards the model predictions. Additionally, for PC3 it higher SHAP values are assocuated with lower scores, while for PC2 the opposite holds.

CONCLUSIONS

- The machine learning model provides a way to classify successfully between 6 classes of mosquitoes based on the sounds they emit.
- Investigate the impact of different feature engineering techniques on model performance.
- Investigate the effect of tuning regularisation λ and α parameters
- Apply methodology to other audio datasets, such as InsectWingbeat and Abuzz, as well as other species of mosquitoes.