FEASIBILITY TEST UPDATES

Resampling/SMOTE

Resampling and SMOTE are possible through the Imbalance Learning (*imblearn*) library (already added on CH3: Libraries). For Oversampling, imblearn's RandomOverSampler was used. For SMOTE, SMOTEN (Synthetic Minority Over-sampling Technique for Nominal) was used (SMOTE but designed to handle categorical data). Do note that resampling (regardless of the technique) may incur overfitting for Benign samples for the case the Oliveira dataset. Hence, there must be proof that can debunk or at least alleviate overfitting concerns which can be done through k-folds testing or model robustness test.

Recent tests on Oliveira suggests that Oversampling is more 'equal' than SMOTE is due to the latter being skewed on certain benign samples. However, in terms of model performance results, the training dataset that had undergone SMOTE outperforms Oversampling.

Oversampling

Top 5 Most Repeated Samples

Sample	Quantity in Resampled	% in Resampled
03384ab6368b68ed16ecb9e6352539af	90	0.22%
0822ec2ba98d291e5bfc836bc3686096	90	0.22%
f78ea80cec007b2c32fb10f9c6c82f39	88	0.21%
075323e77815ee8bcc7854ce23955a15	79	0.19%
79b78bb3d583748040c41ded09555fd3	72	0.17%

Model Performance with Oversampled Dataset

	precision	recall	f1-score	support
0	0.91	0.86	0.88	12827
1	0.87	0.92	0.89	12852
accuracy			0.89	25679
macro avg	0.89	0.89	0.89	25679
weighted avg	0.89	0.89	0.89	25679

SMOTEN

Top 5 Most Repeated Samples

Sample	Quantity in Resampled	% in Resampled
3cedd98ea184c22ee3b024c72a96e075	5965	14.30%
0fbe9eac4ff5af1a392d92881c70c559	3728	8.94%
0b7e7bc7598abe9cfdc594e17e795cf0	1895	4.54%
125d4cdb14dbe86841037e5bbfc6a0bc	895	2.15%
35dd2f5d51ba224735732424f8ab6398	860	2.06%

Model Performance with SMOTEN Dataset

	precision	recall	f1-score	support
0	0.94	0.95	0.94	12827
1	0.95	0.94	0.94	12852
accuracy			0.94	25679
macro avg	0.94	0.94	0.94	25679
weighted avg	0.94	0.94	0.94	25679

Concerns on Model Robustness Test

While it is certain that it will be removed at its current definition (i.e., using different datasets), it is also possible that it will be retained, albeit being redefined in certain aspects. It is possible that part of Oliveira can serve as a stand-in for the external dataset.

Recent tests suggests that it is indeed possible. But do note that Oliveira, at its core, is skewed for Malicious samples, hence it will never score ~90% for on certain metrics (i.e., precision and recall) on individual labels. The two images below show the 90:10 ratio for train (left) and reserve (right) splits respectively.

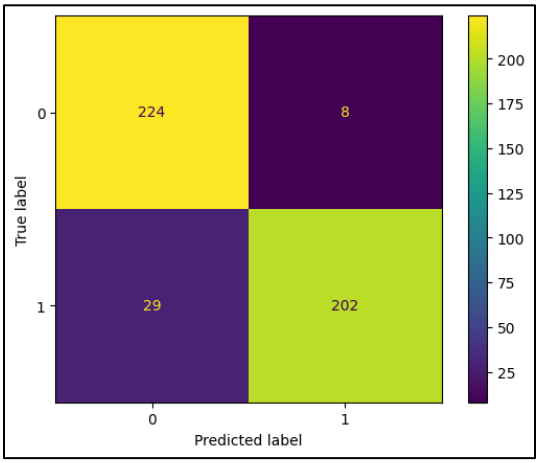
```
train_split shape: (39488, 102)
train_split value_counts:
malware
1    38513
0      975
Name: count, dtype: int64
```

```
reserve_split shape: (4388, 102)
malware
1    4284
0     104
Name: count, dtype: int64
```

LightGBM

LightGBM was tested on MalbehavD-V1 to determine implementation on similar datasets (i.e., one with encoded/numerical data). The dataset (both train and input/test) must be encoded (e.g., LabelEncoded) before use as its implementation of 'support for categorical data' quite misleading as per its documentation. Recent simple tests also suggest that GPUs (tested using AMD GPU on Win11) are supported.

Split Sample Test Results:



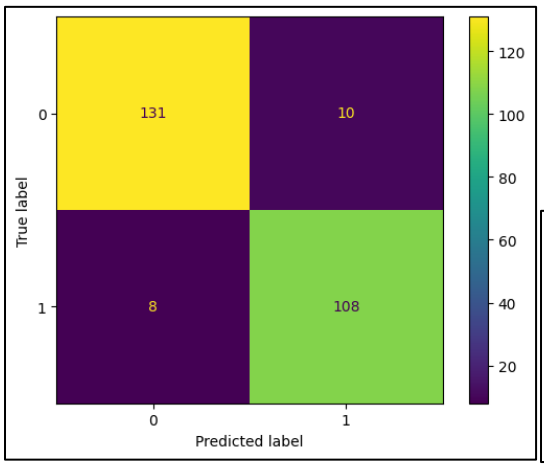
	precision	recall	f1-score	support
0	0.8854	0.9655	0.9237	232
1	0.9619	0.8745	0.9161	231
accuracy			0.9201	463
macro avg	0.9236	0.9200	0.9199	463
weighted avg	0.9236	0.9201	0.9199	463

AUC-ROC Score: 0.92

K-Folds Test Results:

```
[['accuracy', 'f1_score', 'precision', 'recall', 'roc-auc', 'time'],
 [0.9071, 0.907, 0.9401, 0.8718, 0.9075, 25.8268],
 [0.9395, 0.9395, 0.9682, 0.9103, 0.9398, 23.8312],
 [0.9266, 0.9265, 0.963, 0.8889, 0.927, 25.7359],
 [0.9177, 0.9177, 0.9535, 0.8798, 0.9181, 26.5012],
 [0.9156, 0.9155, 0.9535, 0.8761, 0.9161, 26.9401]]
```

Model Robustness Results:



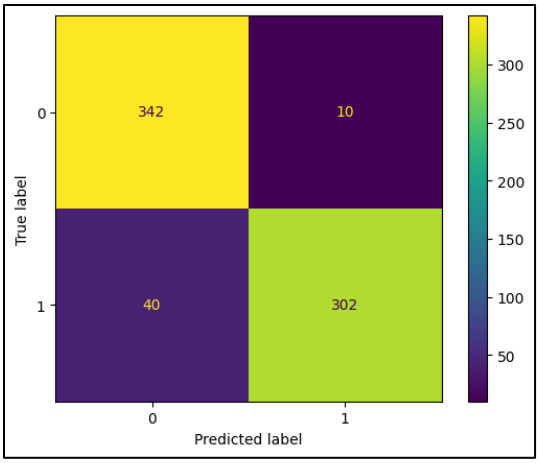
	precision	recall	f1-score	support
0	0.9424	0.9291	0.9357	141
1	0.9153	0.9310	0.9231	116
accuracy			0.9300	257
macro avg	0.9289	0.9301	0.9294	257
weighted avg	0.9302	0.9300	0.9300	257

AUC-ROC Score: 0.9301

CatBoost

CatBoost was tested on MalbehavD-V1 to determine implementation on similar datasets (i.e., one with categorical/string data). Recent simple tests suggests that it is actually capable of handling categorical data, albeit with certain condition(s). Mainly, NaN values shall be converted into a string instead.

Split Sample Test Results:



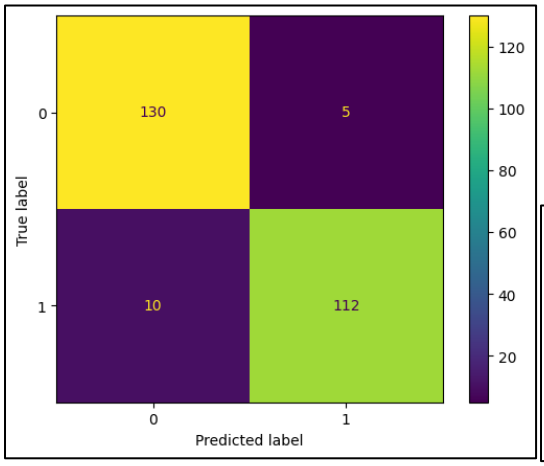
	precision	recall	f1-score	support
0	0.8953	0.9716	0.9319	352
1	0.9679	0.8830	0.9235	342
accuracy			0.9280	694
macro avg	0.9316	0.9273	0.9277	694
weighted avg	0.9311	0.9280	0.9278	694

AUC-ROC Score: 0.9273

K-Folds Test Results:

```
[['accuracy', 'f1_score', 'precision', 'recall', 'roc-auc', 'time'],
 [0.9201, 0.9198, 0.9804, 0.8584, 0.9205, 232.7766],
 [0.9309, 0.9309, 0.9427, 0.9185, 0.931, 224.2083],
 [0.9309, 0.9309, 0.9548, 0.9056, 0.9311, 261.7223],
 [0.9372, 0.9372, 0.9677, 0.9052, 0.9374, 293.2394],
 [0.9156, 0.9155, 0.9447, 0.8836, 0.9157, 238.1865]]
```

Model Robustness Results:



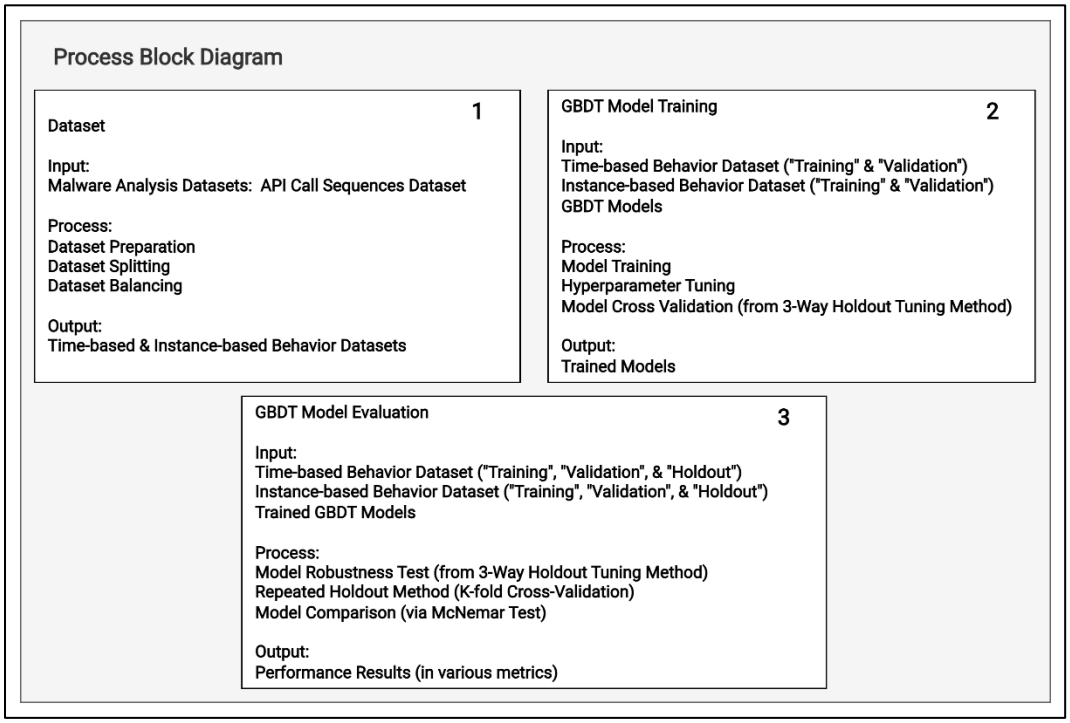
	precision	recall	f1-score	support
0	0.9286	0.9630	0.9455	135
1	0.9573	0.9180	0.9372	122
accuracy			0.9416	257
macro avg	0.9429	0.9405	0.9413	257
weighted avg	0.9422	0.9416	0.9416	257

AUC-ROC Score: 0.9405

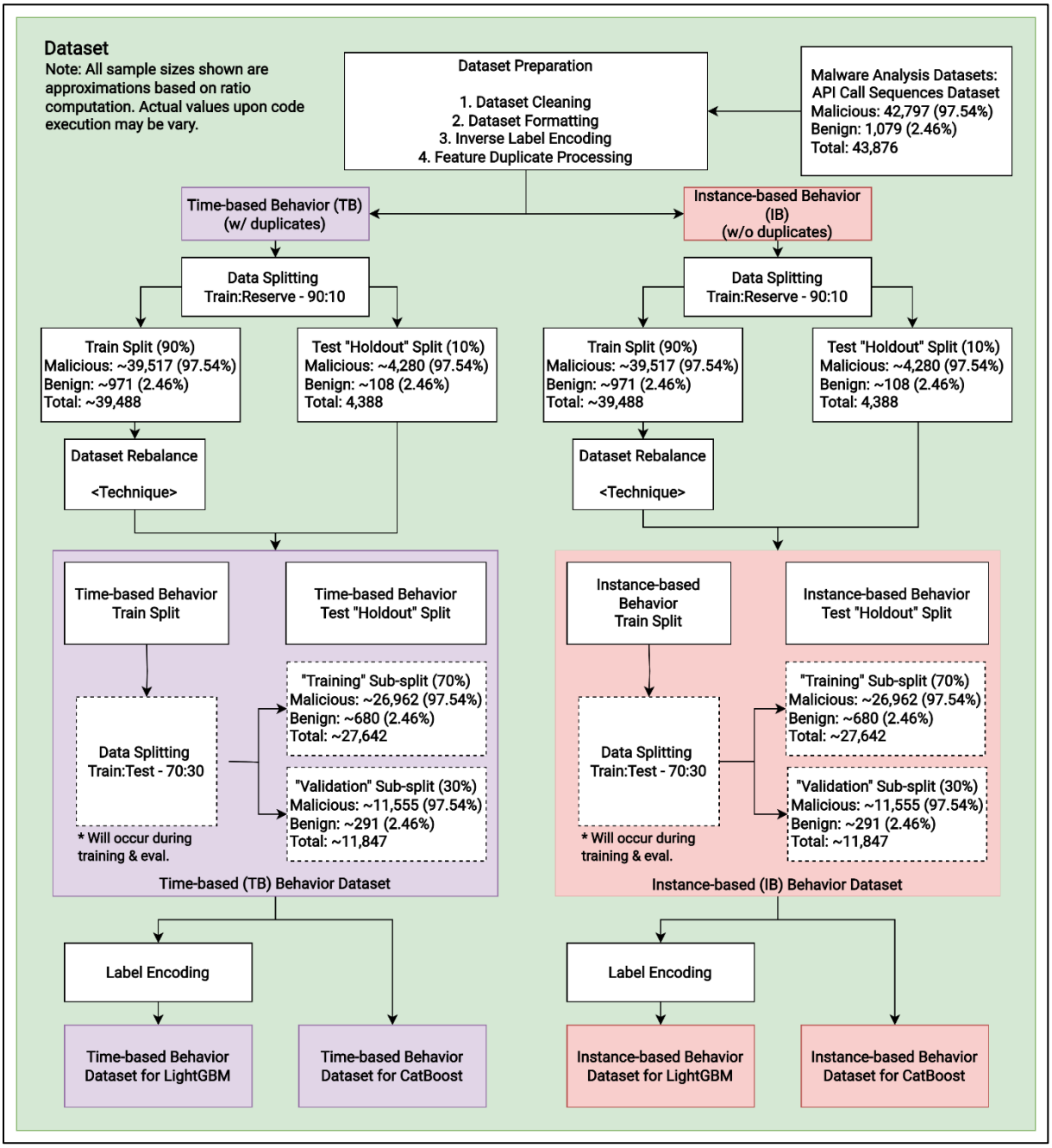
Proposed Process Block Diagram (PBD)

The proposed Process Block Diagrams will be a total of four diagrams which are the Overview, Dataset, GBDT Model Training, and GBDT Model Evaluation. The last 3 diagrams go in line with the study's Gantt chart where each of the major section of the study is divided into 3 sections for each of the terms (i.e., THES1, THES2, THES3). Note that any of the diagrams shown here are prototypes based on the concepts discussed on CH3 which is then arranged according to the study's objectives, needs, & available resources; & is not yet final. It is also apparent that, in the interest of time and available resources, not all concepts discussed in CH3 will be used in CH4.

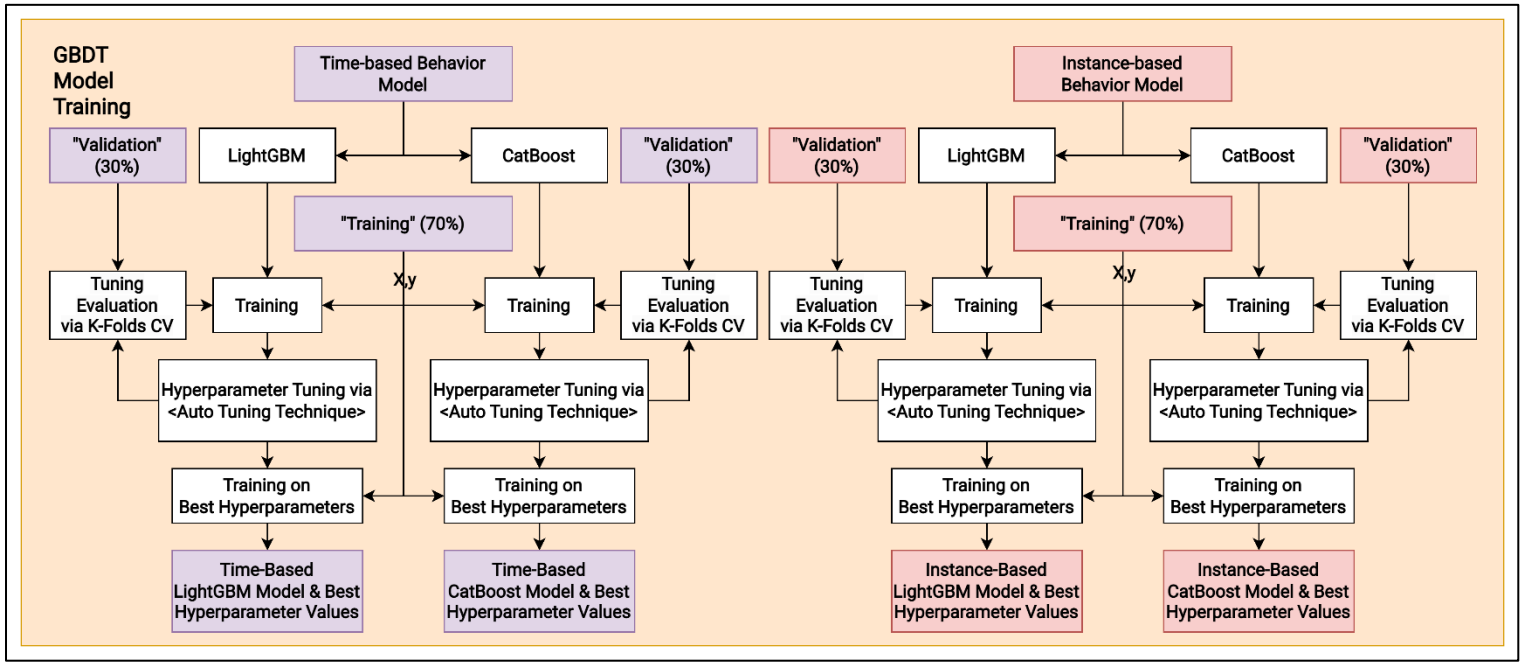
The diagram below shows the Overview PBD where it shows the three major areas of study and the summary of the processes involved in each of the three areas.



The diagram below is about the Dataset. It shows all the processes involved in processing the dataset for it to be useable in the study. The processes shown here are mentioned in CH3. To summarize, the objective of this area of study is to produce the necessary processed datasets which will be used in the next area of the study (i.e., GBDT Model Tuning).



The diagram below is about the GBDT Model Training. It shows all the processes involved in training the model, which is the very essence of the study. The processes shown here are mentioned in CH3. To summarize, the objective of this area of study is to produce the trained models as files to be used on most of the model evaluation (except K-Folds Test) in the next area of the study (i.e., GBDT Model Evaluation).



The diagram below is about the GBDT Model Evaluation. It shows all the processes involved in evaluating the model or the tuning parameters developed in the study. The processes shown here are mentioned in CH3. To summarize, the objective of this area of study is to produce the performance results from various evaluation metrics which can be used to give insights and conclusions as to which model and behavior-type (time-based or instance-based) is better.