Uma Análise Compreensiva e Exaustiva de Métodos de Seleção de Características para Detecção de Malwares Android

UFAM

unipampa

Universidade Federal do Pampa

Vanderson Rocha, Diego Kreutz, Hendrio Bragança, Joner Assolin, Nicolas Pinto e Eduardo Feitosa

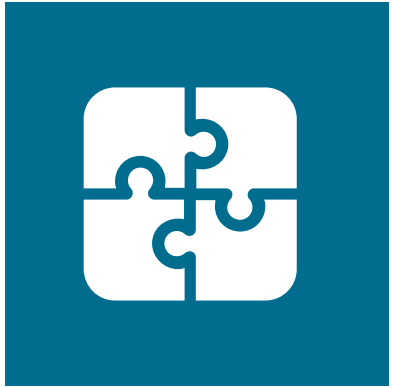
Universidade Federal do Amazonas (UFAM)
Universidade Federal do Pampa (UNIPAMPA)

Desafios

Adaptação
contínua
dos malwares

Desafios

Adaptação
contínua
dos malwares



Seleção
dinâmica
de características

Desafios

Adaptação
contínua
dos malwares

Seleção
dinâmica
de características

Volume
e diversidade
de dados

Desafios

Adaptação contínua dos malwares

Seleção dinâmica de características

Volume e diversidade de dados

Impacto na eficácia dos modelos

Limitações Atuais

Utilização de um único conjunto de dados;

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Limitações Atuais

Dificuldade na reprodutibilidade e verificação independente;

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Limitações Atuais

Risco de overfitting aos dados específicos;

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Limitações Atuais

Comparação entre (poucos) métodos com datasets distintos.

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Contribuições

- Arcabouço de software (esforço de 3 anos);
- Extensiva análise dos métodos seleção;
- Insights contra-intuitivos.

Datasets

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

Datasets

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

Datasets

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

Desempenho dos Métodos: F1 e Recall

Método	F1		Método	Recall	
	Média	Desvio			Desvio
LASSO	0,9071	0,0717	LASSO		0,0586
RFE	0,9030	0,0768	RFE	0,9034	0,0705
SigAPI	0,9008	0,0659	SigAPI	0,9037	0,0573
PCC	0,8997	0,0735	MAD	0,9004	0,0616
...					
SigPID	0,6861	0,3139	PCA	0,6708	0,3294
PCA	0,6650	0,2962	SigPID	0,6525	0,3377
JOWMDroid	0,6361	0,3340	JOWMDroid	0,6486	0,3649
ReliefF	0,6352	0,2780	ReliefF	0,6253	0,3148

Métodos específicos

Desempenho dos Métodos: F1 e Recall

Método	F1		Método	Recall	
	Média	Desvio		Média	Desvio
LASSO	0,9071	0,0717	LASSO	0,9086	0,0586
RFE	0,9030	0,0768	RFE	0,9034	0,0705
SigAPI	0,9008	0,0659	SigAPI	0,9037	0,0573
PCC	0,8997	0,0735	MAD	0,9004	0,0616
...					
SigPID	0,6861	0,3139	PCA	0,6708	0,3294
PCA	0,6650	0,2962	SigPID	0,6525	0,3377
JOWMDroid	0,6361	0,3340	JOWMDroid	0,6486	0,3649
ReliefF	0,6352	0,2780	ReliefF	0,6253	0,3148

Desempenho dos Métodos: F1 e Recall

Método	F1		Método	Recall	
	Média	Desvio		Média	Desvio
LASSO	0,9071	0,0717	LASSO	0,9086	0,0586
RFE	0,9030	0,0768	RFE	0,9034	0,0705
SigAPI	0,9008	0,0659	SigAPI	0,9037	0,0573
PCC	0,8997	0,0735	MAD	0,9004	0,0616
...					
SigPID	0,6861	0,3139	PCA	0,6708	0,3294
PCA	0,6650	0,2962	SigPID	0,6525	0,3377
JOWMDroid	0,6361	0,3340	JOWMDroid	0,6486	0,3649
ReliefF	0,6352	0,2780	ReliefF	0,6253	0,3148

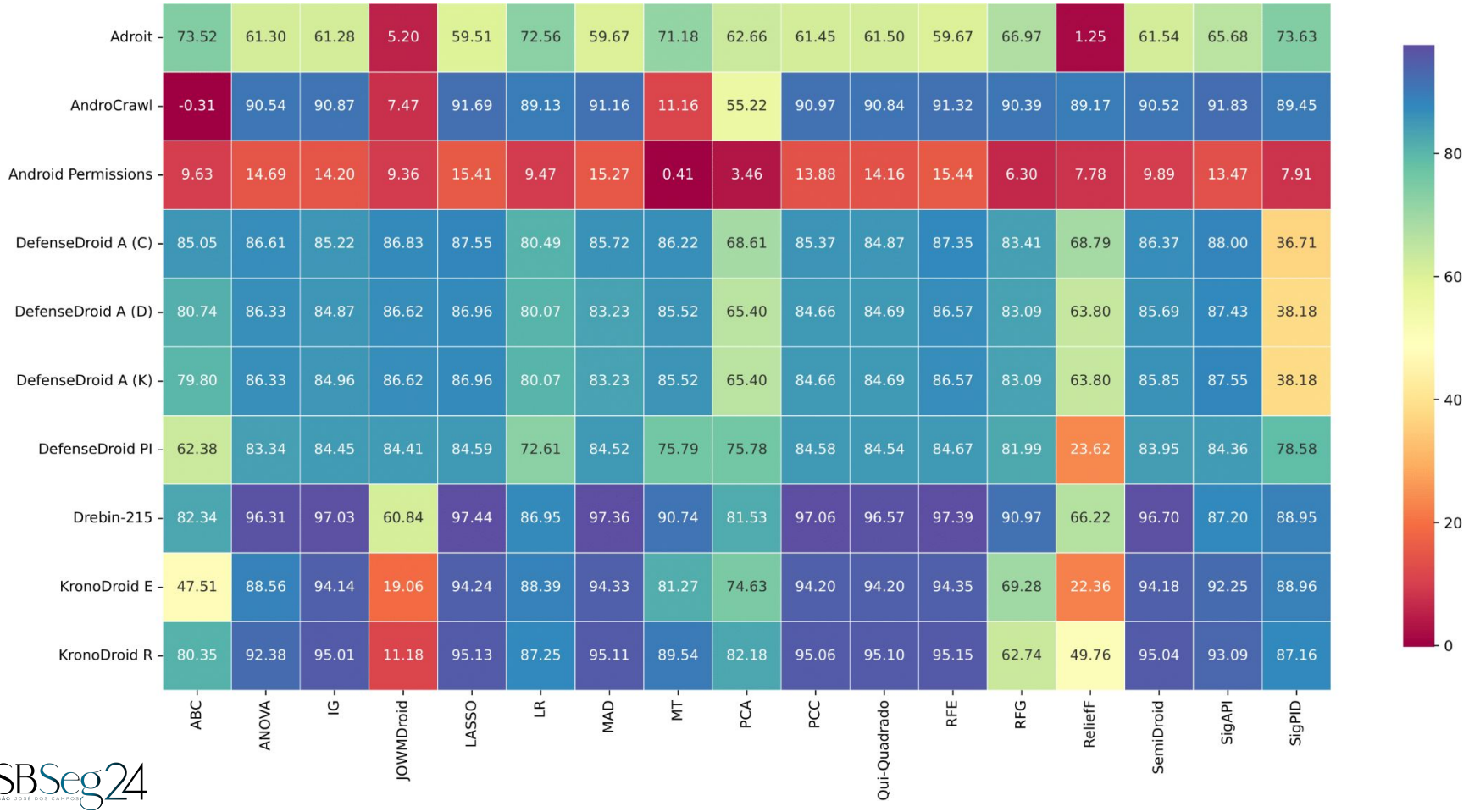
Desempenho dos Métodos: F1 e Recall

Método	F1		Método	Recall	
	Média	Desvio		Média	Desvio
LASSO	0,9071	0,0717	LASSO	0,9086	0,0586
RFE	0,9030	0,0768	RFE	0,9034	0,0705
SigAPI	0,9008	0,0659	SigAPI	0,9037	0,0573
PCC	0,8997	0,0735	MAD	0,9004	0,0616
...					
SigPID	0,6861	0,3139	PCA	0,6708	0,3294
PCA	0,6650	0,2962	SigPID	0,6525	0,3377
JOWMDroid	0,6361	0,3340	JOWMDroid	0,6486	0,3649
ReliefF	0,6352	0,2780	ReliefF	0,6253	0,3148

Desempenho dos Métodos: F1 e Recall

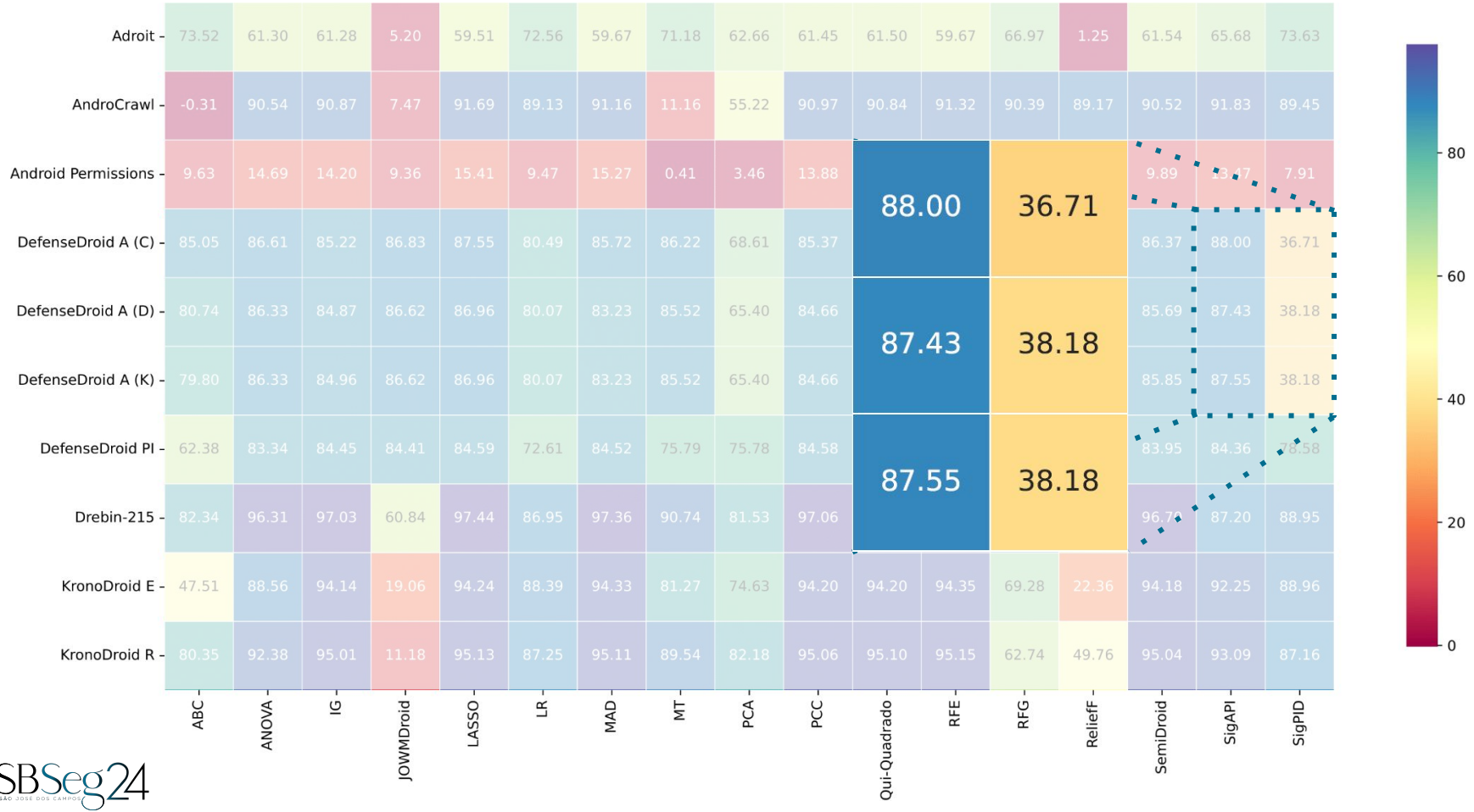
Método	F1		Método	Recall	
	Média	Desvio		Média	Desvio
LASSO	0,9071	0,0717	LASSO	0,9086	0,0586
RFE	0,9030	0,0768	RFE	0,9034	0,0705
SigAPI	0,9008	0,0659	SigAPI	0,9037	0,0573
PCC	0,8997	0,0735	MAD	0,9004	0,0616
...					
SigPID	0,6861	0,3139	PCA	0,6708	0,3294
PCA	0,6650	0,2962	SigPID	0,6525	0,3377
JOWMDroid	0,6361	0,3340	JOWMDroid	0,6486	0,3649
ReliefF	0,6352	0,2780	ReliefF	0,6253	0,3148

Desempenho dos Métodos: MCC

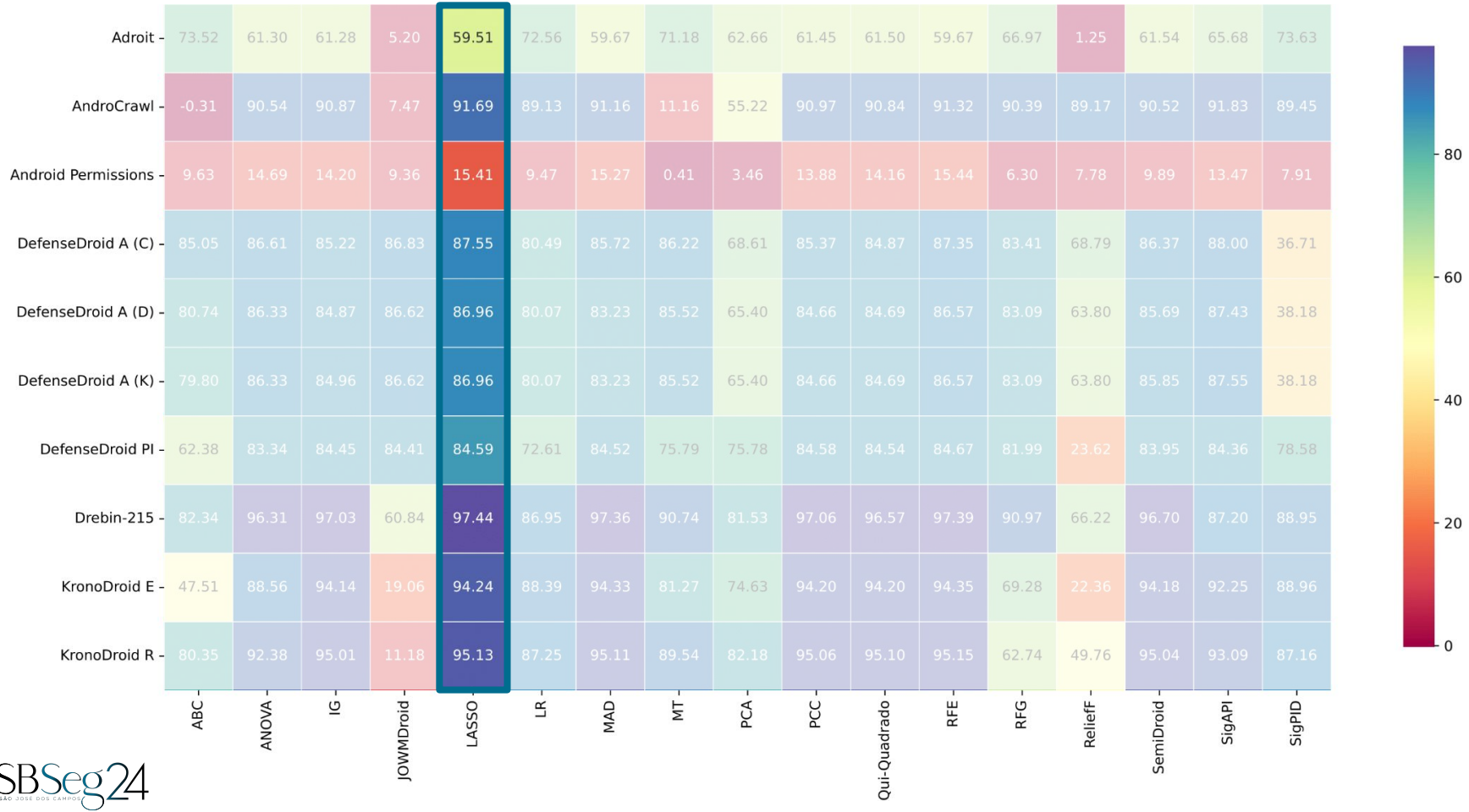


Desempenho dos Métodos: MCC

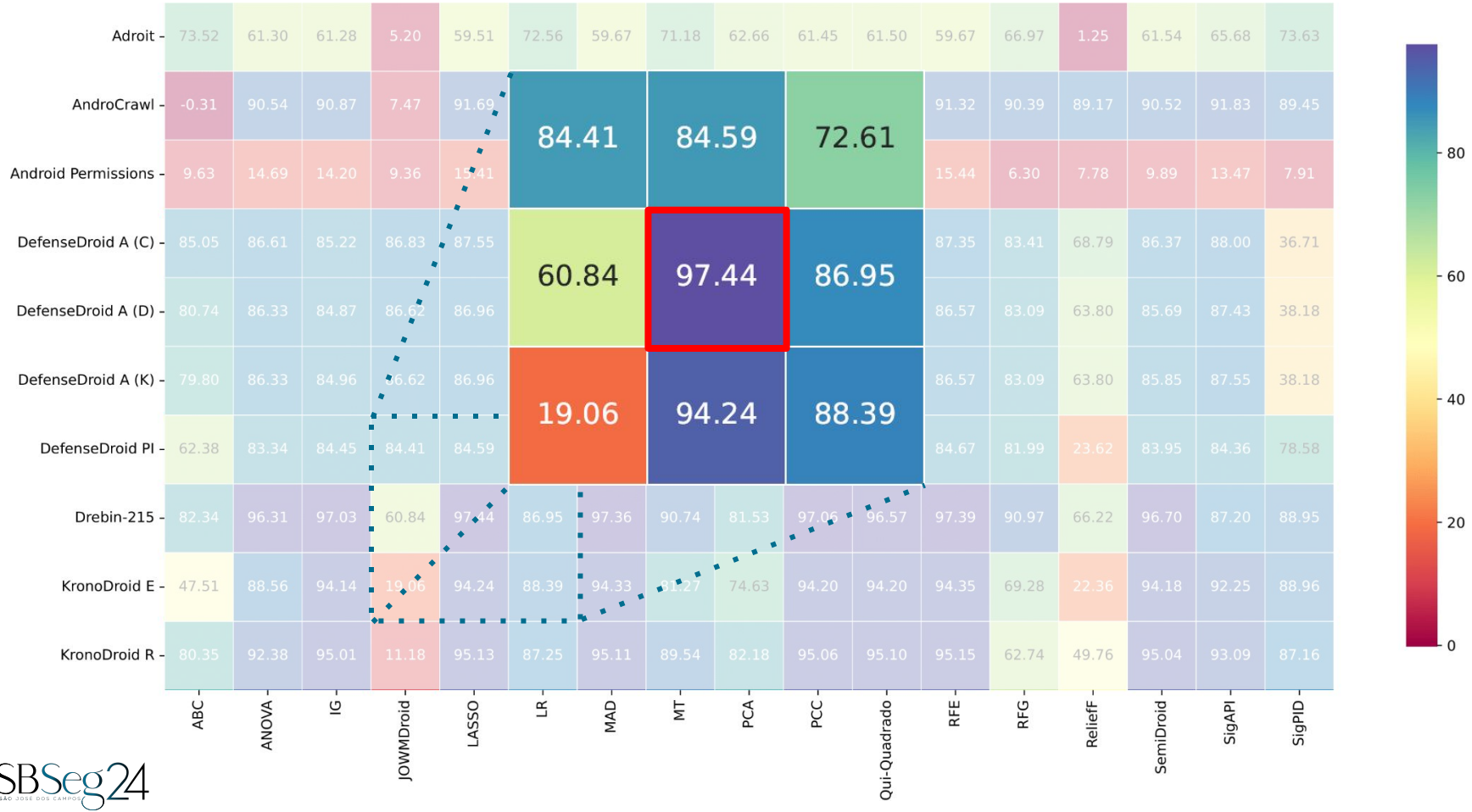
Desempenho dos Métodos: MCC



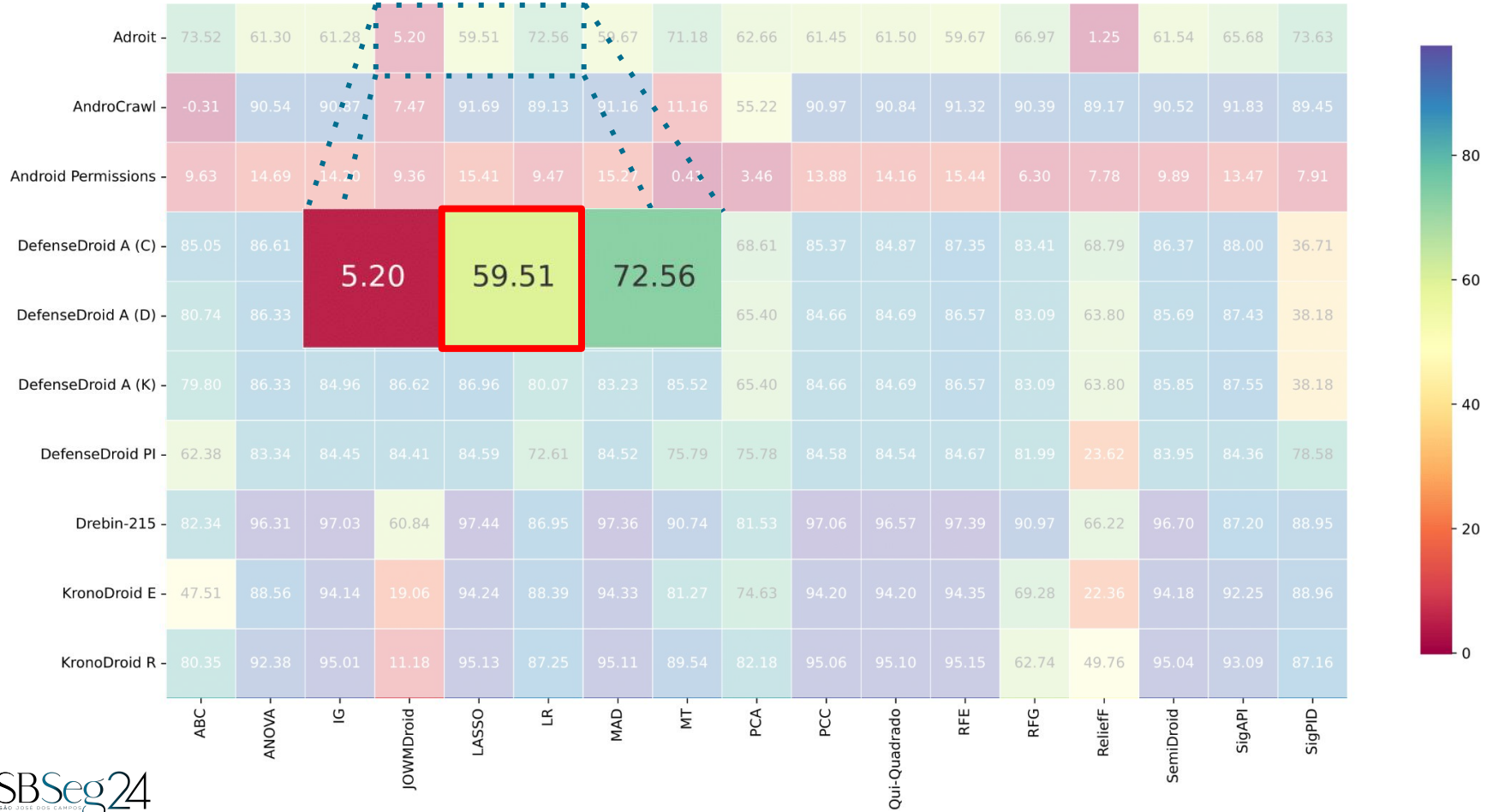
Desempenho dos Métodos: MCC



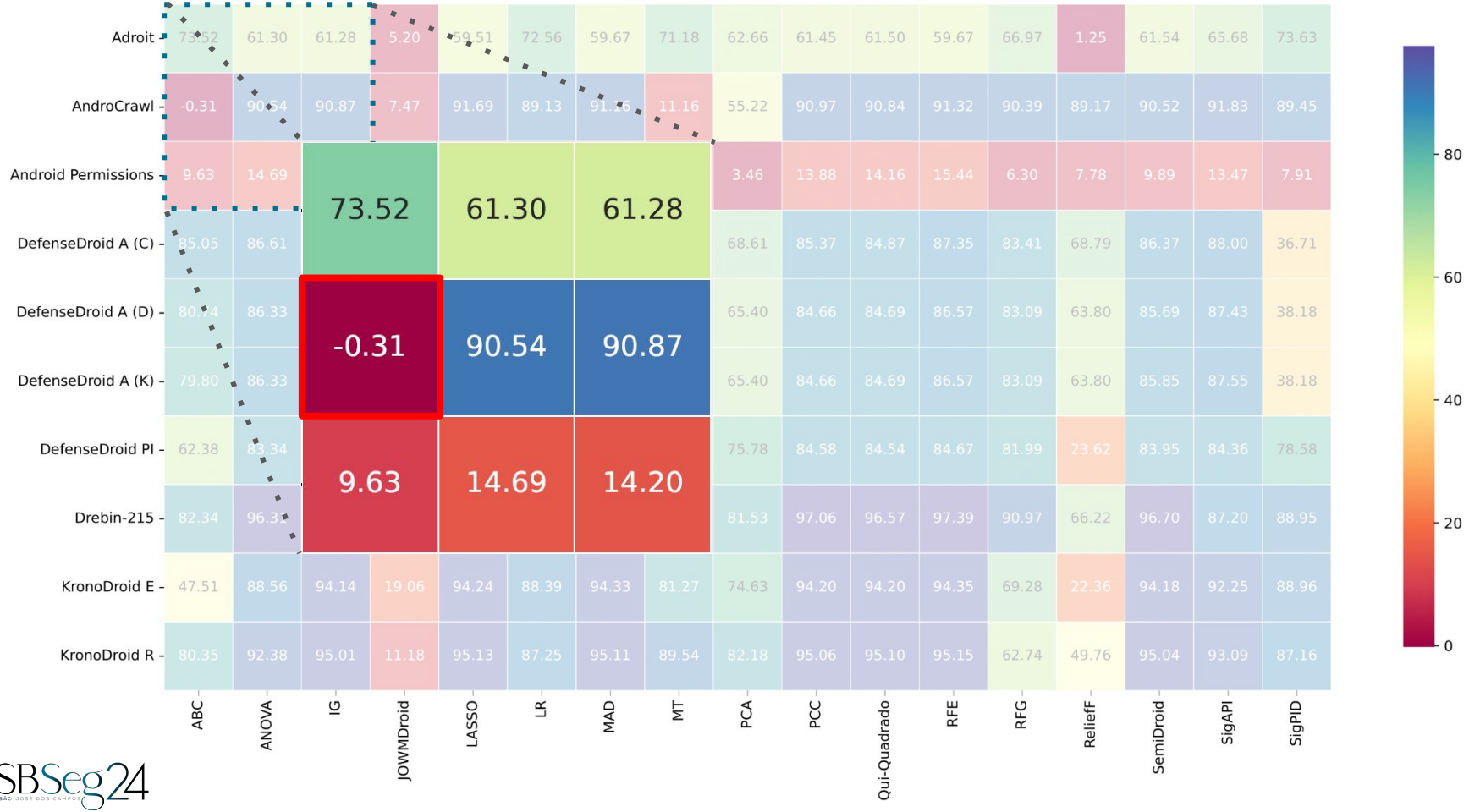
Desempenho dos Métodos: MCC

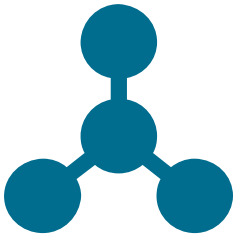


Desempenho dos Métodos: MCC



Desempenho dos Métodos: MCC





Diferentes métodos
de seleção de características

\+

Quantidade mais representativa
e significativa de conjuntos de dados

=

Modelos eficazes de classificação
de aplicativos maliciosos

Trabalhos Futuros

- Avaliar outras abordagens para seleção de características;
- Combinar diferentes métodos;
- Incorporar datasets com multiclasse na avaliação;
- Incluir métodos de seleção de características;
- Agrupamento das amostras por famílias;
- Avaliar o desempenho, a escalabilidade e eficiência computacional dos métodos em conjuntos de dados significativamente grandes.

Obrigado!

Vanderson Rocha

vanderson@ufam.edu.br

ppgi.ufam.edu.br

CAPES

MOTOROLA MOBILITY