[Annexure] Predicting Anti-Microbial Peptides (AMP) through ensemble-based machine learning approach

Atharva Mandar Phatak[1,†], **G K Harish Balaji**[2,†], **Sanjeev M**[3,†], **Hrutik Ravindra Pawar**[4,†], **Kritika M**[5,†] **and Ramavat Yamuna**[6,†]

[1] *BE21B009*

[2] *BE21B017*

[3] *BE21B034*

[4] *BE21B026*

[5] *BE21B022*

[6] *BE21B030*

[†] *These authors contributed equally to this work*

Tables

Table 1 : Feature Information

Feature Name	Feature Description
APAAC (1-28)	The Pseudo Amino Acid Composition, that reflect some sequence-order information via a series of rank-different correlation factors
A	Percentage of Alanine residue present in the sequence
R	Percentage of Arginine residue present in the sequence
N	Percentage of Asparagine residue present in the sequence
D	Percentage of Aspartic Acid residue present in the sequence
C	Percentage of Cysteine residue present in the sequence
Q	Percentage of Glutamine residue present in the sequence
E	Percentage of Glutamic Acid residue present in the sequence
G	Percentage of Glycine residue present in the sequence

H	Percentage of Histidine residue present in the sequence
I	Percentage of Isoleucine residue present in the sequence
L	Percentage of Leucine residue present in the sequence
K	Percentage of Lysine residue present in the sequence
M	Percentage of Methionine residue present in the sequence
F	Percentage of Phenylalanine residue present in the sequence
P	Percentage of Proline residue present in the sequence
S	Percentage of Serine residue present in the sequence
T	Percentage of Threonine residue present in the sequence
W	Percentage of Tryptophan residue present in the sequence
PolarizabilityC1	Percentage of Class 1 (Value 0 - 0.108) residues based on polarizability present in the sequence
PolarizabilityC2	Percentage of Class 2 (Value 0.128 - 0.186) residues based on polarizability present in the sequence
PolarizabilityC3	Percentage of Class 3 (Value 4.03 - 8.08) residues based on polarizability present in the sequence
SolventAccessibilityC1	Percentage of Class 1 (Buried) residues based on solvent accessibility present in the sequence
SolventAccessibilityC2	Percentage of Class 2 (Exposed) residues based on solvent accessibility present in the sequence
SolventAccessibilityC3	Percentage of Class 3 (Intermediate) residues based on solvent accessibility present in the sequence

SecondaryStrC1	Percentage of Class 1 (Helix) residues based on occurrence in protein secondary structure prediction present in the sequence
SecondaryStrC2	Percentage of Class 2 (Strand) residues based on occurrence in protein secondary structure prediction present in the sequence
SecondaryStrC3	Percentage of Class 3 (Intermediate) residues based on occurrence in protein secondary structure prediction present in the sequence
ChargeC1	Percentage of Class 1 (Positive) residues based on charge present in the sequence
ChargeC2	Percentage of Class 2 (Neutral) residues based on charge present in the sequence
ChargeC3	Percentage of Class 3 (Negative) residues based on charge present in the sequence
PolarityC1	Percentage of Class 1 (Value 4.9 - 6.2) residues based on polarity present in the sequence
PolarityC2	Percentage of Class 2 (Value 8.0 - 9.2) residues based on polarity present in the sequence
PolarityC3	Percentage of Class 3 (Value 10.4 - 13) residues based on polarity present in the sequence
NormalizedVDWVC1	Percentage of Class 1 (Volume 0 - 2.78) residues based on normalized van der Waals volume present in the sequence
NormalizedVDWVC2	Percentage of Class 2 (Volume 2.95 - 4.0) residues based on normalized van der Waals volume present in the sequence
NormalizedVDWVC3	Percentage of Class 3 (Volume 4.03 - 8.08) residues based on normalized van der Waals volume present in the sequence
HydrophobicityC1	Percentage of Class 1 (Polar) residues based on polarity present in the sequence

HydrophobicityC2	Percentage of Class 2 (Neutral) residues based on polarity present in the sequence
HydrophobicityC3	Percentage of Class 3 (Hydrophobic) residues based on polarity present in the sequence

Table 2 : Descriptive Statistics Results

Statistic	APAAC13	APAAC6	M	E	APAAC17	D	APAAC23	APAAC15	APAAC24	Q
Mean	2.083663	4.478175	2.129065	4.591795	4.230492	3.341569	-0.198766	3.975153	-0.169504	2.893020
Median	1.613500	3.931000	1.653000	4.000000	3.996500	2.857000	-0.025000	3.146000	-0.010000	2.461000
Mode	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Skewness	2.210122	1.208986	2.111351	1.313039	1.428017	4.066732	-0.317174	4.380212	-0.843988	2.427990
Kurtosis	12.459285	2.914227	10.814811	3.448823	5.919100	72.435296	11.785018	36.967099	12.889921	20.843188
25th Percentile	0.000000	0.000000	0.000000	0.000000	1.168000	0.000000	-0.716000	1.168000	-0.614000	0.000000
50th Percentile	1.613500	3.931000	1.653000	4.000000	3.996500	2.857000	-0.025000	3.146000	-0.010000	2.461000
75th Percentile	3.269250	6.929250	3.333000	7.042000	6.280000	5.333000	0.556000	5.368000	0.436000	4.396000

Table 3 : Inferential Statistics Results

Feature	Z-score	P-value	T-statistic	P-value	F-statistic	P-value	U-statistic	P-value
APAAC 13	88.573790	0.000000e+00	90.186249	0.000000e+00	8133.559572	0.000000e+00	74522391 8.0	0.0
APAAC 6	109.713045	0.000000e+00	110.061200	0.000000e+00	12113.467668	0.000000e+00	73879710 2.0	0.0
M	89.480398	0.000000e+00	91.027609	0.000000e+00	8286.025534	0.000000e+00	74464402 8.0	0.0
E	106.938889	0.000000e+00	107.379989	0.000000e+00	11530.462142	0.000000e+00	73877120 5.5	0.0
APAAC 17	50.047290	0.000000e+00	51.309220	0.000000e+00	2632.636099	0.000000e+00	63630016 9.5	0.0
D	78.639742	0.000000e+00	79.885856	0.000000e+00	6381.750063	0.000000e+00	70201787 2.0	0.0
APAAC 23	42.920225	0.000000e+00	44.788424	0.000000e+00	2006.002926	0.000000e+00	57896871 3.0	0.0
APAAC 15	-6.227769	4.731246e-10	-6.466632	1.009584e-10	41.817336	1.009584e-10	56555036 6.0	0.0
APAAC 24	50.267043	0.000000e+00	52.389683	0.000000e+00	2744.678875	0.000000e+00	60471252 7.0	0.0
Q	45.464309	0.000000e+00	46.313594	0.000000e+00	2144.948958	0.000000e+00	65347108 5.5	0.0

Table 4 : Top 10 features(obtained using Recursive Feature Elimination)

Feature	RFE_Rank	Column_Number
APAAC13	1	13
APAAC6	2	6
M	3	41
E	4	34
APAAC17	5	17
D	6	32
APAAC23	7	23
APAAC15	8	15
APAAC24	9	24
Q	10	35

Table 5: Groupings of amino acids based on the physical properties

Property	Class 1 (C1)	Class 2 (C2)	Class 3 (C3)
Hydrophobicity	Polar	Neutral	Hydrophobic
Peptides	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals	Volume range 0-2.78	Volume range 2.95-4.0	Volume range 4.03-8.08
Peptides	G, A, S, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	Polarity value 4.9-6.2	Polarity value 8.0-9.2	Polarity value 10.4-13
Peptides	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	Polarizability value	Polarizability value 186	Polarizability value .409
Peptides	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
Peptides	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, V	D, E
Secondary structure	Helix	Strand	Coil
Peptides	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent accessibility	Buried	Exposed	Intermediate
Peptides	A, L, F, C, G, I, V, W	P, K, Q, E, N, D	M, P, S, T, H, Y

Figures

Figure 1 : Violin Plots

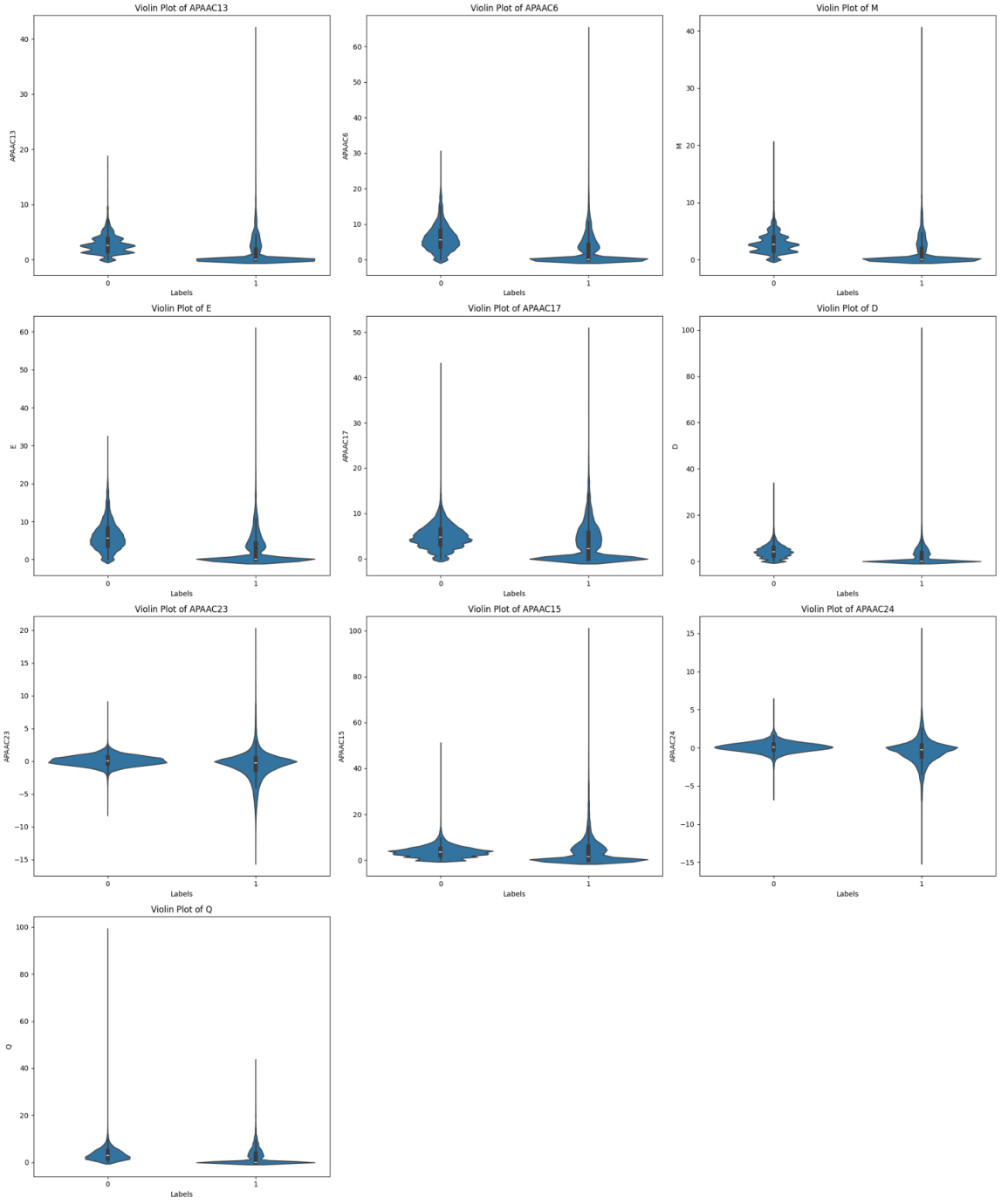


Figure 2 : Pair plots

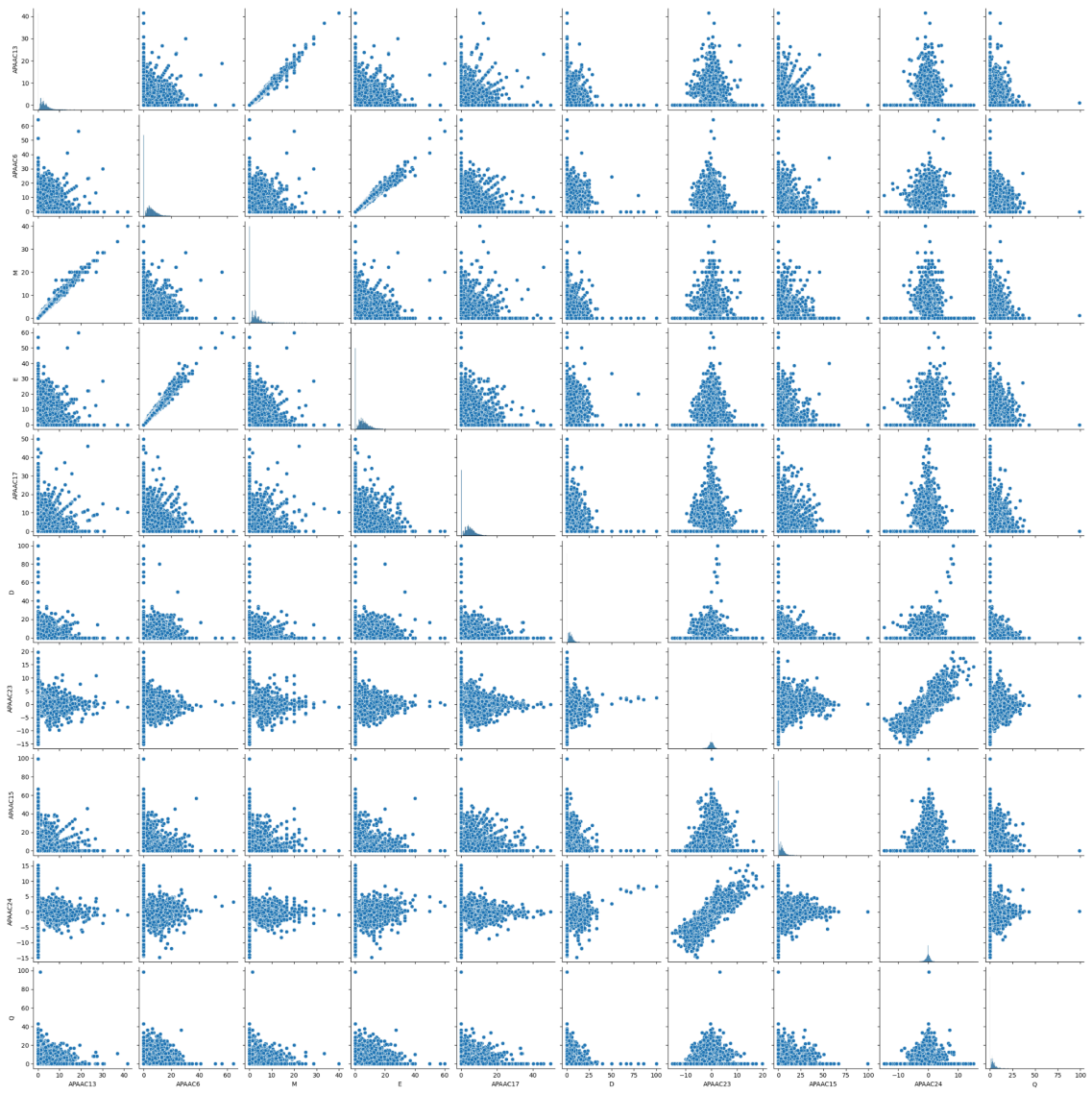


Figure 3 : Jitter plots

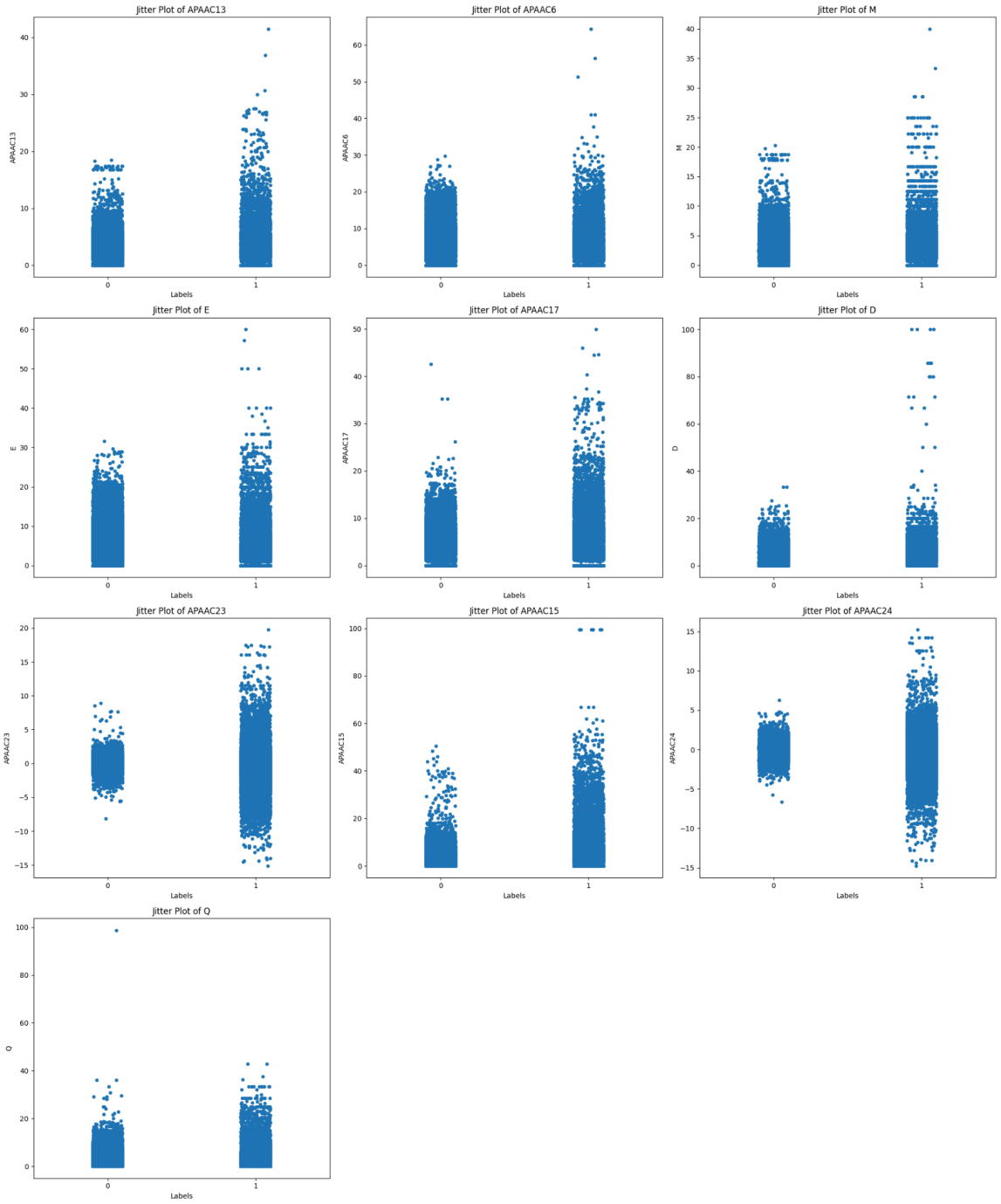


Figure 4 : Histograms

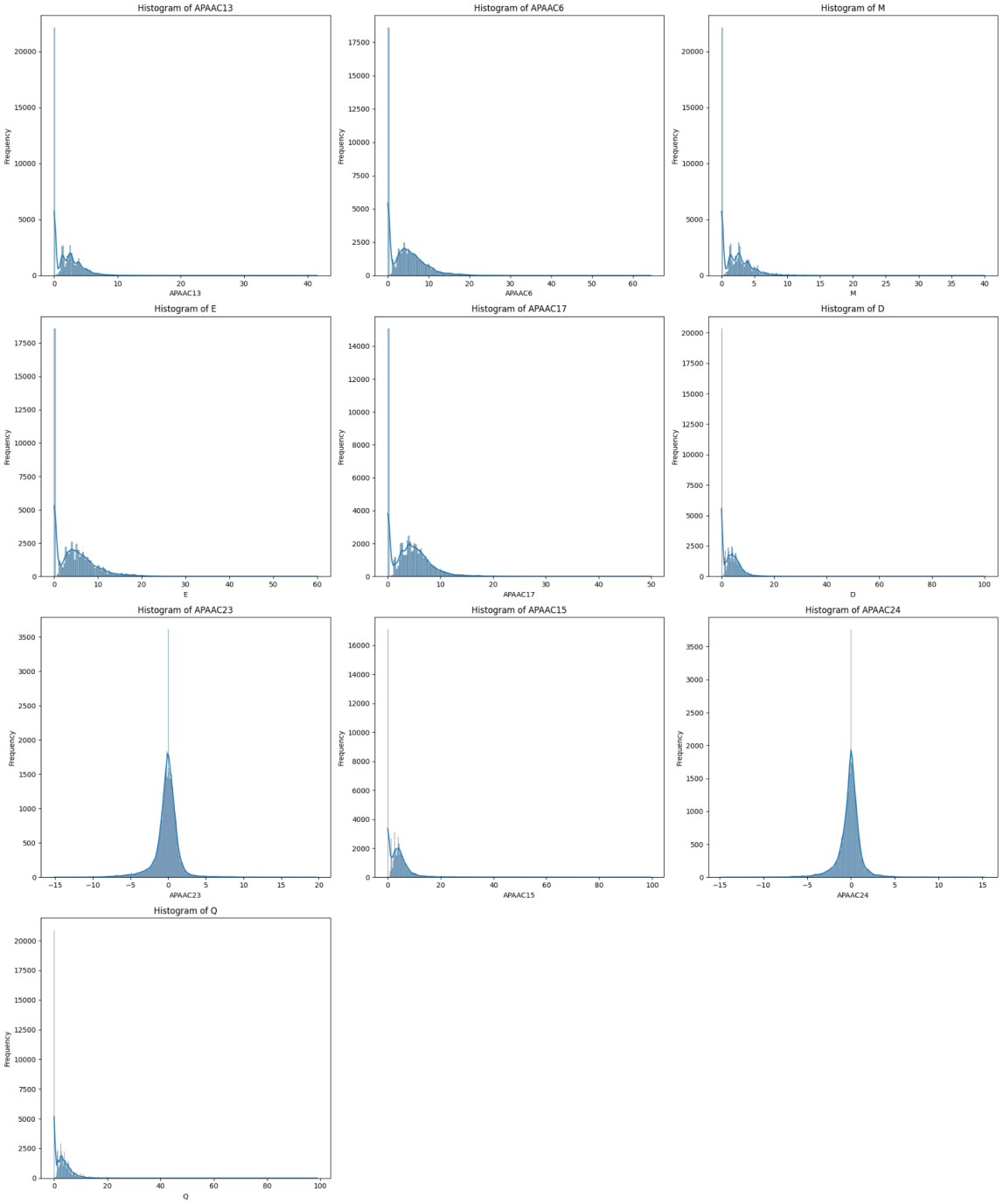


Figure 5 : Density Plots

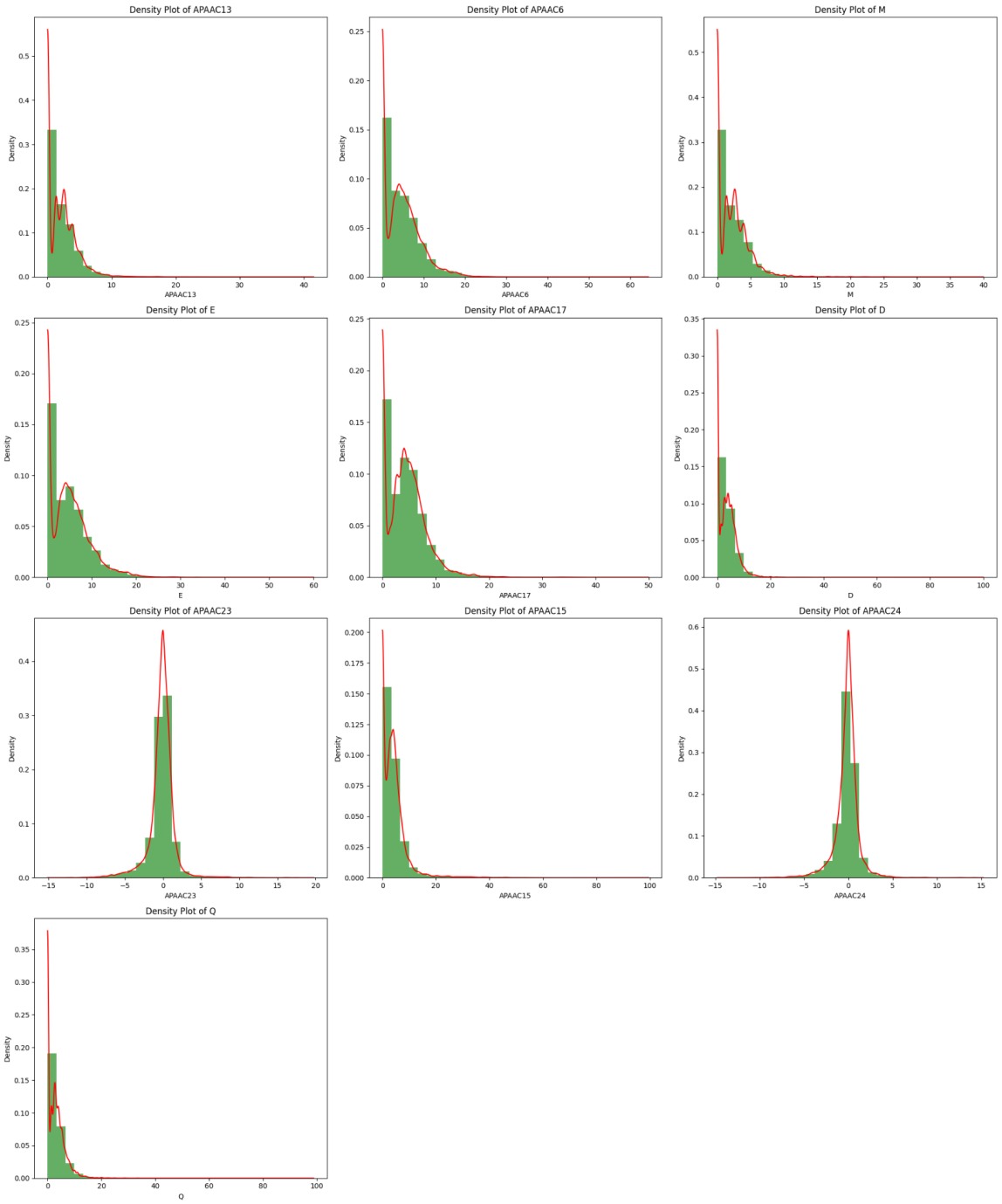
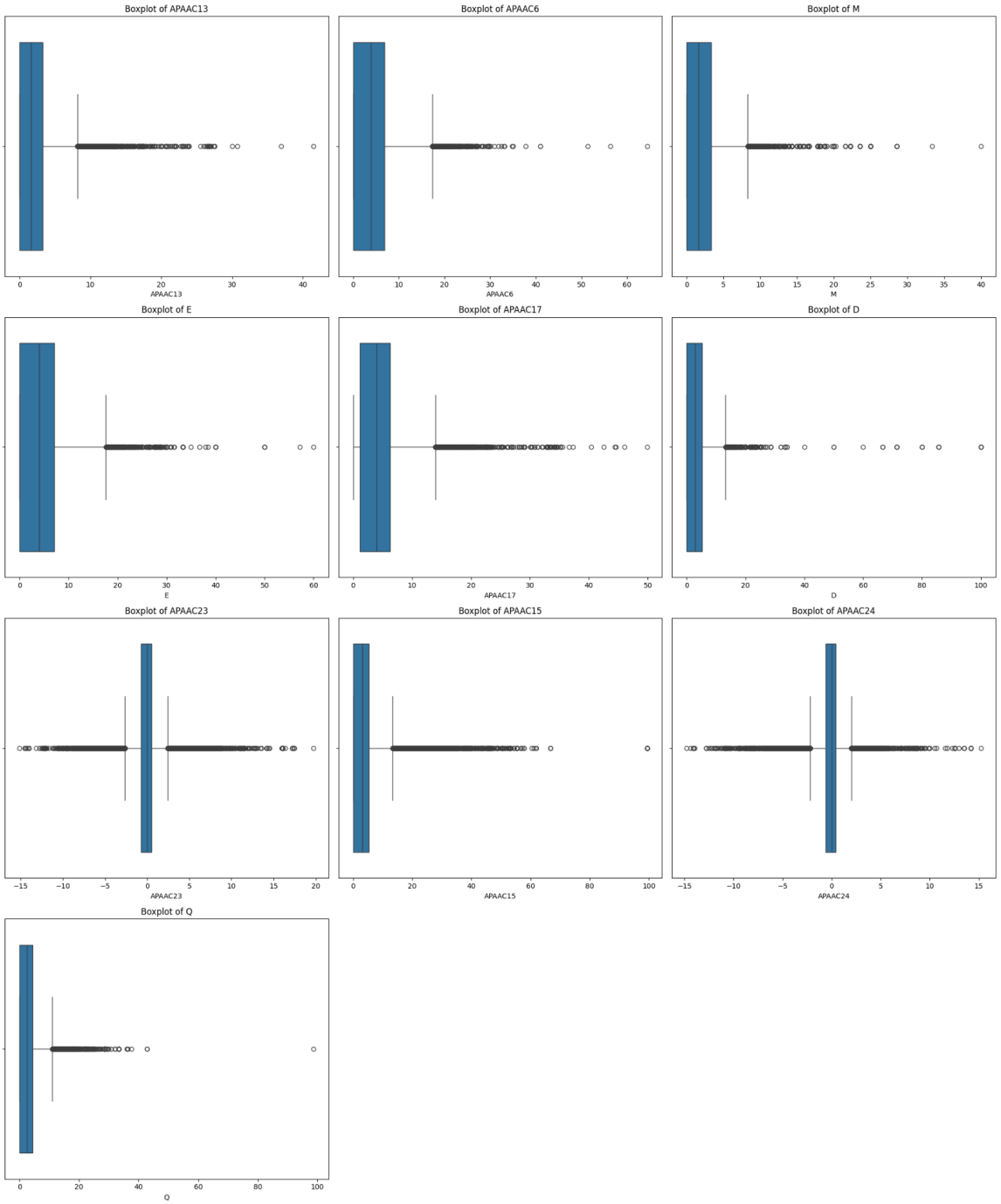


Figure 6 : Box plots



Contributions

Team Member	Roll No.	Contribution
Atharva Mandar Phatak	BE21B009	Literature review, Feature Compilation, Feature Extraction, Feature selection, Report Writing, Report Formatting, Presentation.
G.K Harish Balaji	BE21B017	Literature review, Report writing , Formatting, Presentation, Data Preprocessing, Dataset Compilation, Data Analysis (Descriptive, Visual and Inferential Statistics) and Interpretation.
Krithika M	BE21B022	Literature Review, Visual analysis, Data Compilation, Presentation, Report writing, Report Formatting.
Hrutik Ravindra Pawar	BE21B026	Literature review, Data collection and preprocessing, Features reduction, Report Writing, Report Formatting, Preliminary Analysis, Presentation
Yamuna Ramavath	BE21B030	Literature review, Descriptive statistics, Report writing, Report Formatting, Compilation and Analysis
Sanjeev M.	BE21B034	Literature review, Report writing, Report formatting , Model development, Sequence-descriptor extraction, Managed Github repository, Developed streamlit website tool