IDENTIFICATION OF MICROBIAL BIOMARKERS FOR HUMAN GUT DISEASES USING MICROBIAL INTERACTION NETWORK EMBEDDED DEEP LEARNING

ANUSHKA SIVAKUMAR
SYAMA K
J. ANGEL ARUL

01. RESEARCH STATEMENT

To identify prominent and meaningful biomarkers from metagenomic datasets of Inflammatory Bowel Disease and Colorectal Cancer by constructing an informative Microbial Interaction Network using the tool MAGMA, and embedding it into a Deep Feed-Forward Neural Network model.

02.

LITERATURE REVIEW

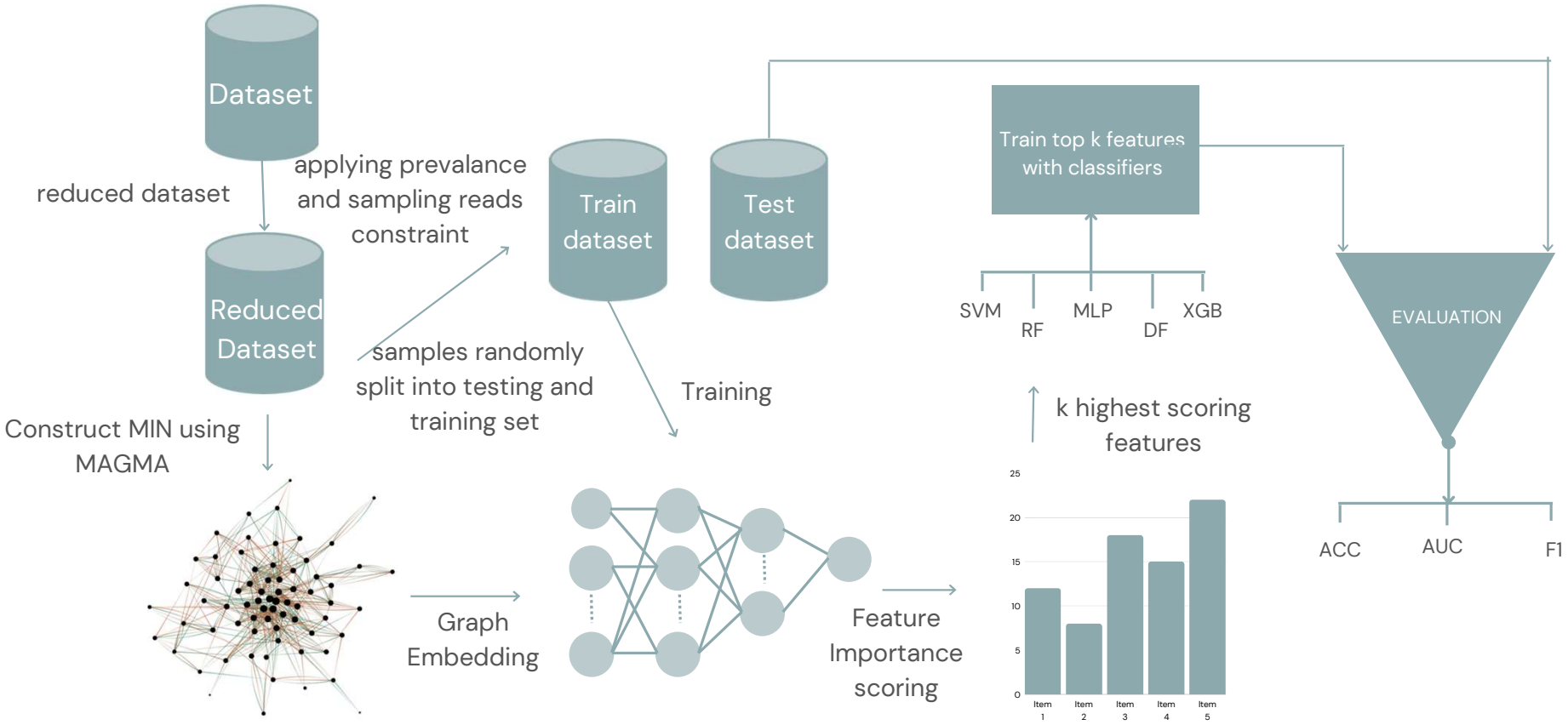
FEATURE SELECTION BASED

PAPER	METHOD
Fleuret(2004)	Feature selection: 'Fast CMIM' Classification: perceptron, Naïve Bayesian classifier, Nearest neighbor, Adaboost, and SVM.
L. Yu and H. Liu	Feature selection: FCBF. Classification: C4.5 and NBC.
H Peng, C Ding (2005)	Feature selection: MRMR Classification: NBC, LDA, and SVM, Logistic regression.
B Wajid et al.(2019)	Hybrid feature selection: (BSS/WSS)+(embedded classifier) NCC

BIOMARKER IDENTIFICATION BASED

PAPER	METHOD
Zhu Q, et al (2019)	Network construction: SparCC+Spiec-Easi Feature selection: MIN + MIC Feature Importance+classification: GEDFN
Abbas M et al. (2019)	Network construction: SparCC, MB (Spiec-Easi), RMT, CoNet, and Proxi. Feature Selection: Betweenness Centrality, Closeness Centrality, Average Neighbor Degree, Clustering Coefficient, Node Clique Number, Core Number, and critical attack set- NBR-Clust. Classification: RF
Bakir-Gungor B et al. (2021)	Feature selection: CMIM, MRMR,CBF,SelectKBest.; K-means to generate subgroups. Classification: RF
A Acharjee et al. (2020)	RF based Feature selection: Boruta, Recursive feature elimination, permutation based feature selection with and without correction, and backward elimination based feature selection. Classification: RF
H Hacilar, et al (2020)	Feature selection: FCBF, CMIM, mRMR, and XGBoost. Subgroups: K-means, PCA, hierarchical clustering. Classification: RF, Decision trees, Logiboost, AdaBoost, K-means + Logiboost, and SVM.
Q Zhu, et al. (2020)	Feature selection + Classification: Deep Forest (data perturbation method for feature selection)

03. METHODOLOGY



04.

FINDINGS & RESULTS

A. IBD DATASET RESULTS

a) All Features	RF			SVM			MLP			DF			XGB		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
	0.804	0.756	0.855	0.728	0.750	0.856	0.604	0.741	0.850	0.855	0.829	0.890	0.829	0.797	0.872
b) MAGMA	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.819	0.785	0.872	0.707	0.763	0.865	0.544	0.757	0.861	0.814	0.818	0.887	0.806	0.801	0.879
200	0.785	0.773	0.866	0.675	0.760	0.861	0.552	0.757	0.862	0.839	0.822	0.889	0.813	0.802	0.880
300	0.811	0.782	0.868	0.726	0.755	0.860	0.533	0.750	0.857	**0.863**	**0.839**	**0.897**	0.850	0.807	0.881
400	0.813	0.794	0.879	0.705	0.772	0.870	0.525	0.763	0.865	0.848	0.816	0.884	0.822	0.806	0.880
500	0.809	0.789	0.875	0.773	0.775	0.872	0.528	0.753	0.859	0.840	0.819	0.884	0.845	0.796	0.871
c) Sparcc+SpiecEasi+MIC	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.799	0.793	0.877	0.647	0.750	0.857	0.512	0.747	0.855	0.824	0.805	0.876	0.829	0.789	0.867
200	0.814	0.802	0.883	0.689	0.751	0.857	0.541	0.750	0.857	0.817	0.810	0.879	0.841	0.788	0.866
300	0.809	0.754	0.853	0.713	0.746	0.855	0.532	0.759	0.863	0.847	0.832	0.891	0.851	0.806	0.879
400	0.828	0.779	0.870	0.732	0.741	0.851	0.555	0.743	0.852	0.842	0.830	0.892	0.835	0.790	0.868
500	0.814	0.775	0.867	0.695	0.764	0.866	0.537	0.754	0.859	0.826	0.820	0.885	0.840	0.799	0.872
d) SpiecEasi	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.761	0.773	0.864	0.692	0.747	0.855	0.611	0.763	0.865	0.815	0.790	0.867	0.780	0.788	0.871
200	0.825	0.785	0.872	0.711	0.749	0.856	0.589	0.748	0.855	0.838	0.820	0.884	0.835	0.805	0.881
300	0.815	0.788	0.874	0.694	0.751	0.857	0.542	0.755	0.860	0.813	0.793	0.870	0.816	0.779	0.860
400	0.849	0.771	0.865	0.713	0.750	0.857	0.517	0.755	0.860	0.824	0.815	0.882	0.832	0.798	0.874
500	0.831	0.787	0.875	0.723	0.768	0.868	0.540	0.741	0.851	0.821	0.811	0.880	0.840	0.806	0.880
e) SparCC	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.811	0.779	0.868	0.706	0.761	0.864	0.539	0.726	0.841	0.851	0.824	0.889	0.837	0.802	0.876
200	0.839	0.801	0.881	0.729	0.749	0.856	0.537	0.739	0.850	0.845	0.809	0.877	0.851	0.809	0.879
300	0.818	0.802	0.884	0.748	0.767	0.867	0.505	0.752	0.858	0.858	0.819	0.884	0.844	0.790	0.867
400	0.810	0.775	0.866	0.734	0.761	0.864	0.545	0.764	0.866	0.854	0.818	0.883	0.856	0.801	0.876
500	0.829	0.788	0.874	0.742	0.764	0.865	0.549	0.742	0.852	0.855	0.823	0.889	0.840	0.802	0.875

B. CRC DATASET RESULTS

a) All Features	RF			SVM			MLP			DF			XGB		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
	0.801	0.730	0.717	0.758	0.697	0.720	0.697	0.627	0.604	0.767	0.746	0.731	0.709	0.649	0.652
b) MAGMA	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.680	0.611	0.580	0.766	0.665	0.683	0.637	0.551	0.425	0.752	0.692	0.690	0.773	0.692	0.692
200	0.709	0.649	0.647	0.693	0.638	0.625	0.633	0.649	0.574	0.727	0.638	0.619	0.635	0.578	0.549
300	0.732	0.665	0.643	0.803	0.714	0.694	0.600	0.503	0.527	0.777	0.670	0.658	0.696	0.616	0.616
400	0.819	0.724	0.715	0.743	0.649	0.695	0.679	0.627	0.589	**0.837**	**0.768**	**0.757**	0.737	0.632	0.601
500	0.728	0.665	0.615	0.812	0.719	0.745	0.667	0.622	0.645	0.803	0.730	0.732	0.672	0.605	0.633
c) Sparcc+SpiecEasi+MIC	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.627	0.568	0.551	0.621	0.573	0.589	0.549	0.514	0.452	0.643	0.584	0.586	0.659	0.584	0.571
200	0.793	0.735	0.742	0.670	0.573	0.573	0.593	0.573	0.474	0.758	0.719	0.727	0.684	0.627	0.616
300	0.808	0.686	0.673	0.707	0.611	0.609	0.690	0.627	0.593	0.734	0.692	0.686	0.649	0.600	0.583
400	0.706	0.676	0.685	0.707	0.605	0.536	0.621	0.600	0.611	0.763	0.719	0.699	0.664	0.605	0.612
500	0.759	0.676	0.691	0.707	0.643	0.616	0.623	0.584	0.575	0.760	0.670	0.637	0.673	0.649	0.640
d) SpiecEasi	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.695	0.643	0.622	0.715	0.616	0.628	0.593	0.546	0.349	0.730	0.697	0.694	0.720	0.654	0.674
200	0.727	0.649	0.657	0.704	0.659	0.701	0.509	0.497	0.338	0.700	0.638	0.623	0.656	0.643	0.647
300	0.667	0.622	0.610	0.692	0.627	0.623	0.609	0.535	0.494	0.713	0.659	0.656	0.738	0.681	0.681
400	0.802	0.735	0.729	0.745	0.670	0.681	0.647	0.611	0.593	0.779	0.719	0.728	0.595	0.600	0.586
500	0.803	0.730	0.716	0.765	0.692	0.699	0.669	0.638	0.556	0.757	0.692	0.685	0.668	0.627	0.633
e) SparCC	**RF**			**SVM**			**MLP**			**DF**			**XGB**		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.732	0.643	0.634	0.684	0.595	0.656	0.519	0.535	0.444	0.714	0.600	0.577	0.648	0.611	0.592
200	0.769	0.681	0.667	0.800	0.724	0.752	0.576	0.568	0.422	0.765	0.665	0.652	0.745	0.659	0.652
300	0.717	0.659	0.634	0.785	0.714	0.747	0.585	0.546	0.508	0.776	0.703	0.702	0.708	0.659	0.659
400	0.714	0.627	0.608	0.789	0.719	0.748	0.512	0.476	0.364	0.815	0.724	0.724	0.766	0.697	0.689
500	0.724	0.643	0.622	0.788	0.719	0.705	0.584	0.524	0.478	0.772	0.686	0.694	0.694	0.665	0.684

C. COMPARATIVE RESULTS

IBD DATASET

a) SparCC+Spiec-Easi+MIC_gedfn	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.812	0.782	0.869	0.688	0.766	0.867	0.509	0.761	0.864	0.830	0.794	0.866	0.812	0.784	0.865
200	0.831	0.769	0.864	0.676	0.753	0.859	0.570	0.757	0.861	0.843	0.819	0.885	0.821	0.788	0.866
300	0.823	0.761	0.856	0.710	0.775	0.873	0.513	0.754	0.860	0.857	0.826	0.888	0.846	0.804	0.876
400	0.831	0.780	0.870	0.699	0.769	0.869	0.541	0.738	0.849	0.853	0.821	0.886	0.832	0.818	0.888
500	0.828	0.801	0.884	0.702	0.749	0.856	0.511	0.757	0.861	0.819	0.815	0.883	0.842	0.813	0.882

MAXIMUM by MAGMA+DFNN (300 features): AUC = 0.863, ACC = 0.839, F1 = 0.897

CRC DATASET

b) SparCC+Spiec-Easi+MIC_gedfn	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.683	0.605	0.563	0.658	0.595	0.522	0.600	0.568	0.547	0.724	0.686	0.668	0.741	0.654	0.645
200	0.667	0.643	0.641	0.722	0.697	0.705	0.548	0.535	0.412	0.727	0.686	0.675	0.734	0.681	0.682
300	0.703	0.692	0.649	0.722	0.622	0.635	0.590	0.600	0.579	0.789	0.681	0.672	0.717	0.638	0.643
400	0.736	0.670	0.666	0.736	0.659	0.658	0.560	0.545	0.451	0.659	0.622	0.618	0.717	0.654	0.645
500	0.715	0.681	0.697	0.683	0.616	0.636	0.655	0.578	0.498	0.717	0.670	0.644	0.693	0.622	0.609

MAXIMUM by MAGMA+DFNN (400 features): AUC = 0.837, ACC = 0.768, F1 = 0.757

D. BIOMARKER VALIDATION

IBD DATASET

a) Čipčić Paljetak, Hana et al. (2022)		b) Gevers, Dirk et al. (2014)
Enterobacteriaceae	**F. prausnitzii**	**Enterobacteriaceae**
Eubacterium	**Turicibacteriaceae / Turicibacter**	**Pasteurellacaea**
Lactobacillaceae	**Haemophilus**	**Veillonellaceae**
Dialister	**R. gnavus**	**Fusobacteriaceae**
Christensenellaceae	**Erysipelotrichaceae**	**Erysipelotrichales**
Ruminococcus	**Blautia**	**Bacteroidales**
Anaerostipes	**Coprococcus**	**Clostridiales**
A. muciniphila	**Veillonellaceae**	
Adlercreutzia	**Phascolarctobacterium**	
Lactobacillus		

CRC DATASET

a) M. Oudah and A. Henschel (2018) - Top 20 features CRC1	b) Zeller, Georg et al. (2014)
Fusobacteriaceae	Fusobacteriaceae
Clostridiales	Peptostreptococcus
Bacteroides	Eubacterium
Eubacterium biforme	**Streptococcus**
Ruminococcus	
Prevotella	
Rikenellaceae	
S24-7	
Veillonellaceae	
Coprococcus	
Dorea	

05. CONCLUSION

- *MAGMA MIN emphasizes the underlying biological process, through the inclusion of covariates*
- *MAGMA considers multivariate associations and partial correlations*
- *MAGMA showed most tempered output and least negative links. The spurious negative links were eliminated by taking the covariate measure into account.*
- *Embedding suitable MINs helped to deal with over-dispersion and high levels of noise in the dataset – reliable feature selection*
- *The proposed methodology achieved the highest AUC, accuracy, and F1-score when classified using DF across both the IBD and CRC datasets.*

THANK YOU

The floor is now open to questions