Trusting Classifiers with Interpretable Machine Learning Based Feature Selection Backpropagation

Saikat Das
Department of Computer Scinece
Utah Valley University
Orem, UT, USA
Saikat.Das@uvu.edu

Raktim Ranjan Das
Department of Computer Science and Engineering
Stamford University, Bangladesh
RaktimDas16@gmail.com

Frederick T. Sheldon
Department of Computer Science
University of Idaho
Moscow, ID, USA
sheldon@uidaho.edu

Sajjan Shiva
Department of Computer Science
The University of Memphis
Memphis, TN, USA
sshiva@memphis.edu

***Abstract*— In a machine learning classification problem, feature selection is a required pre-processing phase which identifies important and relevant features from the dataset to potentially reduce the computational complexity and improves the overall classification performances. Feature reduction mechanisms, such as Information Gain, Gain Ratio, Chi-squared, ReliefF, Deep Learning, etc. along with domain knowledge are used to find the appropriate features from a dataset. In this paper, we propose a novel feature selection process based on interpretable machine learning technique (IMLFS) to find the optimal relevant features in detecting DDoS cyber-attacks. Based on the effectiveness of critical features, this technique is also used to explain a detected DDoS attack. These relevant features are used in the feature selection phase to retrain the model for better accuracy. The benchmark dataset, NSL-KDD is used to evaluate the proposed approach. Moreover, using the extracted features obtained from this dataset, we investigated our recently developed ensemble supervised framework. This investigation confirms the efficacy of the IMLFS approach by producing both higher detection accuracy and lower false positive alarms. A significant improved accuracy and model training times compared to earlier studies that compared various IML methods are reported here.**

APPENDIX

In this section, ROC AUC curves for nine selection methods and the overview of all experimental results are shown graphically and in tabular form, respectively.

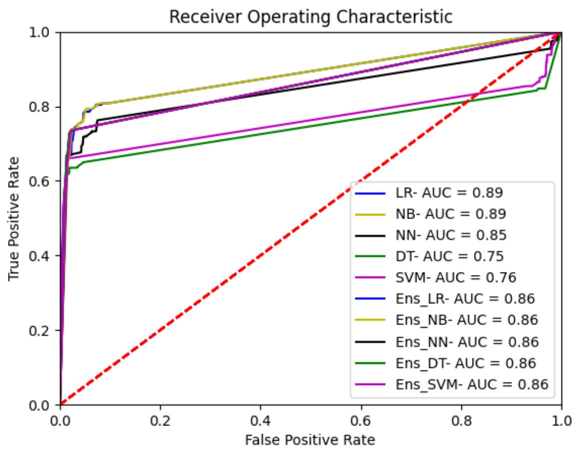

Fig. 1. ROC AUC using Anova Method

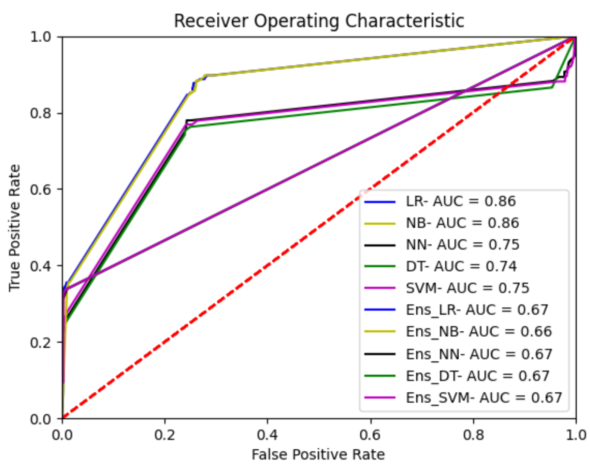

Fig. 2. ROC AUC using Chi-Square Method

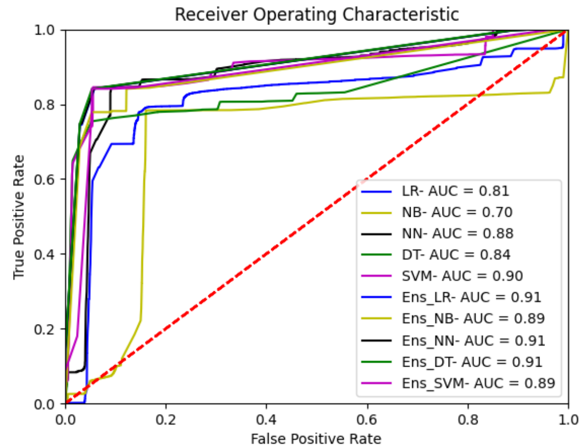

Fig. 3. ROC AUC using LASSO Method

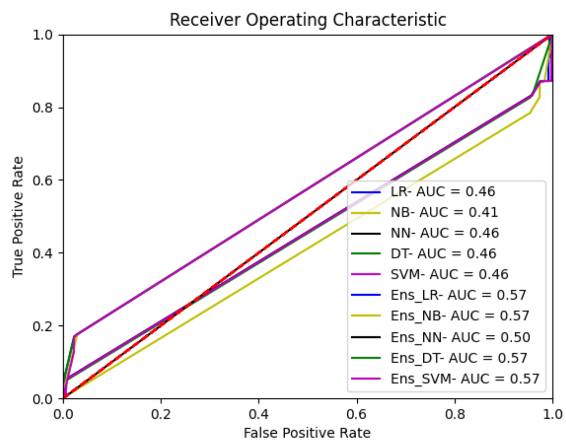

Fig. 4. ROC AUC using LR with L1 penalty Method

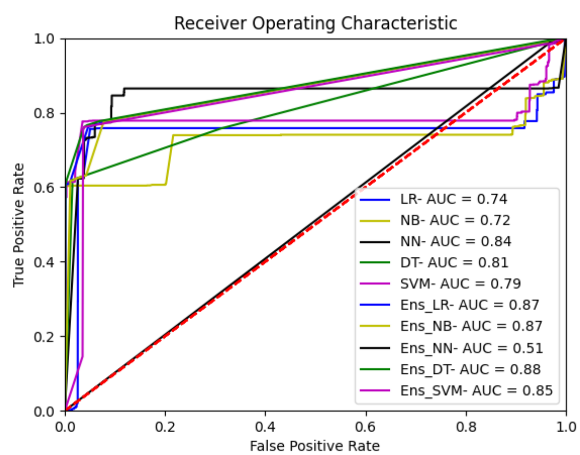

Fig. 5. ROC AUC using Mutual Information Method

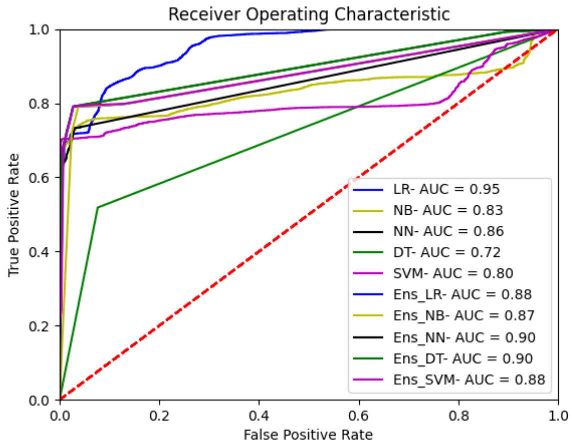

Fig. 6. ROC AUC using PCA Method

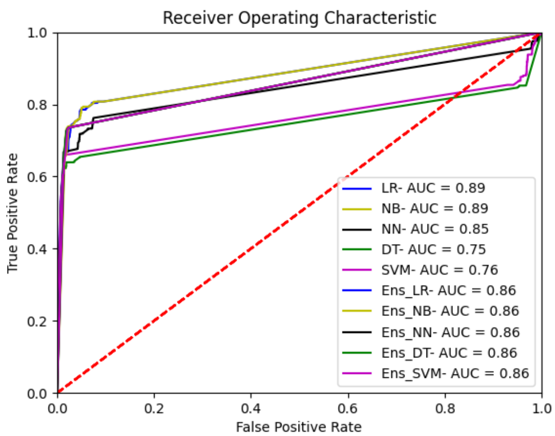

Fig. 7. ROC AUC using Pearson Method

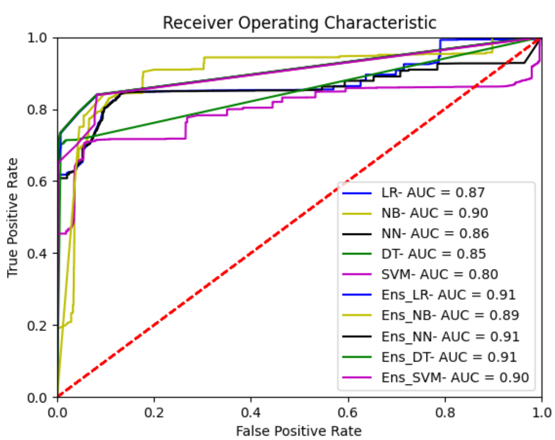

Fig. 8. ROC AUC using Random Forest Method

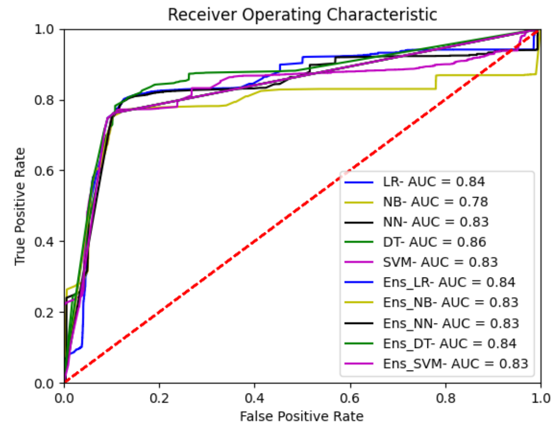

Fig. 9. ROC AUC using Recursive Feature Elimination Method

TABLE I. DATA CLASSIFICATION OVERVIEW WITH ENSEMBLE SUPERVISED FRAMEWORK [3] USING EXTRACTED FEATURES FROM SEVEN SELECTION METHODS AND FROM ENFS, AND WITHOUT USING ANY SELECTION METHOD.

Method	Classifier Category	Classifier Name	F-1 Score	Accuracy	Precision	Recall	FPR
Without any feature selection (Full Feature Set)	Individual	LR	0.846	0.877	0.930	0.775	0.045
		NB	0.807	0.856	0.971	0.690	0.016
		NN	0.840	0.873	0.933	0.763	0.042
		DT	0.875	0.895	0.928	0.832	0.021
		SVM	0.866	0.897	0.990	0.770	0.006
	Ensemble	Ens_MV	0.858	0.891	0.988	0.759	0.007
		Ens_LR	0.804	0.857	0.938	0.722	0.010
		Ens_NB	0.870	0.892	0.925	0.821	0.052
		Ens_NN	0.872	0.901	0.930	0.835	0.013
		Ens_DT	0.884	0.900	0.878	0.890	0.011
		Ens_SVM	0.834	0.845	0.882	0.791	0.012
Anova (F#1)	Individual	LR	0.782	0.842	0.971	0.655	0.015
		NB	0.831	0.871	0.968	0.728	0.019
		NN	0.744	0.820	0.976	0.601	0.011
		DT	0.753	0.825	0.971	0.615	0.014
		SVM	0.763	0.831	0.975	0.627	0.012
	Ensemble	Ens_MV	0.770	0.835	0.975	0.636	0.012
		Ens_LR	0.831	0.871	0.964	0.730	0.021
		Ens_NB	0.833	0.872	0.959	0.736	0.024
		Ens_NN	0.833	0.872	0.959	0.736	0.024
		Ens_DT	0.833	0.872	0.959	0.736	0.024
		Ens_SVM	0.833	0.872	0.959	0.736	0.024
Chi-Square (F#2)	Individual	LR	0.488	0.705	0.993	0.324	0.002
		NB	0.488	0.704	0.981	0.325	0.005
		NN	0.396	0.670	0.970	0.249	0.006
		DT	0.398	0.671	0.965	0.251	0.007
		SVM	0.385	0.668	0.989	0.239	0.002
	Ensemble	Ens_MV	0.395	0.672	0.988	0.247	0.002
		Ens_LR	0.495	0.707	0.980	0.331	0.005
		Ens_NB	0.501	0.708	0.963	0.339	0.010
		Ens_NN	0.501	0.708	0.963	0.339	0.010
		Ens_DT	0.498	0.709	0.981	0.334	0.005
		Ens_SVM	0.501	0.709	0.973	0.337	0.007
LASSO (F#3)	Individual	LR	0.750	0.799	0.815	0.694	0.121
		NB	0.786	0.815	0.789	0.783	0.161
		NN	0.855	0.877	0.876	0.836	0.091
		DT	0.797	0.848	0.951	0.686	0.027
		SVM	0.842	0.873	0.916	0.778	0.055

Method	Classifier Category	Classifier Name	F-1 Score	Accuracy	Precision	Recall	FPR
	Ensemble	Ens_MV	0.821	0.853	0.869	0.778	0.090
		Ens_LR	0.876	0.897	0.918	0.838	0.057
		Ens_NB	0.822	0.854	0.867	0.782	0.091
		Ens_NN	0.878	0.899	0.918	0.841	0.057
		Ens_DT	0.880	0.901	0.923	0.841	0.053
		Ens_SVM	0.880	0.901	0.923	0.841	0.053
LR with L1 (F#4)	Individual	LR	0.013	0.565	0.462	0.006	0.006
		NB	0.553	0.390	0.406	0.869	0.978
		NN	0.092	0.583	0.858	0.049	0.006
		DT	0.093	0.583	0.861	0.049	0.006
		SVM	0.092	0.583	0.854	0.049	0.006
	Ensemble	Ens_MV	0.092	0.583	0.858	0.049	0.006
		Ens_LR	0.290	0.627	0.824	0.176	0.029
		Ens_NB	0.290	0.627	0.822	0.176	0.029
		Ens_NN	nan	0.567	nan	0.000	0.000
		Ens_DT	0.283	0.628	0.850	0.170	0.023
		Ens_SVM	0.290	0.627	0.824	0.176	0.029
Mutual Information (F#5)	Individual	LR	0.032	0.560	0.354	0.017	0.024
		NB	0.550	0.399	0.408	0.844	0.942
		NN	0.752	0.821	0.947	0.623	0.027
		DT	0.755	0.826	0.966	0.620	0.017
		SVM	0.758	0.830	0.991	0.613	0.004
	Ensemble	Ens_MV	0.756	0.828	0.981	0.615	0.009
		Ens_LR	0.836	0.870	0.926	0.762	0.046
		Ens_NB	0.749	0.816	0.919	0.632	0.043
		Ens_NN	nan	0.567	nan	0.000	0.000
		Ens_DT	0.840	0.875	0.937	0.762	0.039
		Ens_SVM	0.840	0.875	0.937	0.762	0.039
PCA (F#6)	Individual	LR	0.818	0.862	0.962	0.711	0.022
		NB	0.790	0.848	0.983	0.661	0.009
		NN	0.802	0.853	0.961	0.689	0.021
		DT	0.641	0.748	0.839	0.519	0.077
		SVM	0.792	0.838	0.897	0.709	0.063
	Ensemble	Ens_MV	0.793	0.848	0.970	0.671	0.016
		Ens_LR	0.866	0.894	0.957	0.790	0.027
		Ens_NB	0.860	0.888	0.941	0.791	0.038
		Ens_NN	0.866	0.894	0.956	0.791	0.028
		Ens_DT	0.866	0.894	0.957	0.791	0.027
		Ens_SVM	0.866	0.894	0.957	0.790	0.027
Pearson	Individual	LR	0.782	0.842	0.971	0.655	0.015
		NB	0.831	0.871	0.968	0.728	0.019
		NN	0.744	0.820	0.976	0.601	0.011

Method	Classifier Category	Classifier Name	F-1 Score	Accuracy	Precision	Recall	FPR
		DT	0.756	0.827	0.971	0.619	0.014
		SVM	0.763	0.831	0.975	0.627	0.012
	Ensemble	Ens_MV	0.770	0.835	0.975	0.636	0.012
		Ens_LR	0.831	0.871	0.964	0.730	0.021
		Ens_NB	0.833	0.872	0.959	0.736	0.024
		Ens_NN	0.833	0.872	0.959	0.736	0.024
		Ens_DT	0.833	0.872	0.959	0.736	0.024
		Ens_SVM	0.833	0.872	0.959	0.736	0.024
RF (F#8)	Individual	LR	0.782	0.833	0.906	0.688	0.055
		NB	0.832	0.861	0.875	0.793	0.087
		NN	0.762	0.822	0.909	0.656	0.050
		DT	0.819	0.866	0.987	0.700	0.007
		SVM	0.763	0.825	0.927	0.649	0.039
	Ensemble	Ens_MV	0.764	0.823	0.912	0.657	0.049
		Ens_LR	0.861	0.883	0.883	0.840	0.085
		Ens_NB	0.844	0.875	0.916	0.783	0.055
		Ens_NN	0.863	0.884	0.887	0.840	0.082
		Ens_DT	0.863	0.884	0.886	0.840	0.082
		Ens_SVM	0.862	0.884	0.885	0.840	0.083
RFE (F#9)	Individual	LR	0.701	0.784	0.878	0.583	0.062
		NB	0.701	0.783	0.874	0.585	0.065
		NN	0.772	0.819	0.852	0.707	0.095
		DT	0.760	0.813	0.860	0.681	0.085
		SVM	0.717	0.789	0.858	0.616	0.078
	Ensemble	Ens_MV	0.714	0.790	0.872	0.605	0.068
		Ens_LR	0.801	0.839	0.862	0.748	0.091
		Ens_NB	0.775	0.822	0.854	0.709	0.093
		Ens_NN	0.803	0.838	0.849	0.761	0.103
		Ens_DT	0.800	0.839	0.862	0.747	0.091
		Ens_SVM	0.801	0.839	0.862	0.748	0.091
Explanation Based Learning (F#10)	Individual	LR	0.823	0.853	0.864	0.785	0.095
		NB	0.824	0.853	0.862	0.788	0.097
		NN	0.106	0.570	0.553	0.058	0.036
		DT	0.879	0.897	0.895	0.863	0.078
		SVM	0.913	0.925	0.926	0.900	0.055
	Ensemble	Ens_MV	0.827	0.856	0.867	0.790	0.093
		Ens_LR	0.938	0.945	0.921	0.955	0.064
		Ens_NB	0.888	0.901	0.877	0.900	0.099
		Ens_NN	0.938	0.944	0.921	0.955	0.064
		Ens_DT	0.940	0.946	0.925	0.955	0.060
		Ens_SVM	0.940	0.946	0.925	0.955	0.060