CS-550 - Spring 2023 Project: Detection of Pneumonia-Induced Lung Diseases with an Imbalanced Dataset of Chest X-ray Lung Images

Salih Deniz Uzel (deniz.uzel@bilkent.edu.tr)
22201382

***Abstract*—Diagnosing lower respiratory diseases through X-ray images requires experienced professionals and often times diagnose accuracy of the newer or unknown diseases lacks behind of the other test methods. We are aiming to use ensemble learning methods to increase the diagnostic success rate and help radiologists and doctors. For this purpose, I trained machine learning techniques using ensemble learning methods on balanced and in-balanced Covid19, Normal and Bacterial & Viral Pneumonia chest X-ray dataset. Single model Support Vector Machine (SVM) followed by bagging SVM achieved highest test accuracy among all the training models for the balanced and unbalanced dataset. Experiments conducted on 5 different data distribution for balanced an imbalanced dataset. Results were calculated by averaging 5 runs with a standard deviation between 0% and 2% for the overall accuracy metric. The purpose of this research is to provide an comparative overview of common ensemble learning techniques and their performance for the medical radiology image classification task.**

***Index Terms*—ensemble learning, medical AI, machine learning, medical image classification.**

I. Introduction

In the field of medicine, imaging tools play a important role in diagnosing and treating patients. Medical doctors rely on these devices, along with other complementary diagnostic methods, to make better decisions. Medical professionals and computer scientists are collaborating to enhance software and develop post-processing tools, aiming to achieve better results from imaging devices. As artificial intelligence (AI) techniques become more prevalent and the number of AI practitioners increases, researchers are conducting studies to explore the utilization of these techniques in medicine for purposes such as diagnosis, treatment regulation, and risk assessment. The applications being developed in this relatively new and evolving field aim to expand diagnostic services, address the shortage of experts and financial burden, and provide solutions to existing challenges in the medical field. In the field of AI, ensemble learning techniques have gained popularity in the medical image classification field to improve prediction accuracy. Methods such as bagging, boosting, and stacking are widely utilized to select the most suitable model for a given problem. When the ensemble model appropriately configured with hyperparameters, they can increase the success rate of given task. However, it is important to note that ensemble models require substantial amount of computational power. Unsuitable hyperparameter or technique selections may lead to overfitting issues. In the field of medical image classification, ensemble methods such as bootstrapping, bagging, random forests, and boosting have demonstrated their effectiveness in improving decision model performance [1], [2].

Lower respiratory diseases are mostly viral and bacterial diseases that affect the world population every year. Chest x-ray imaging devices are used to help the diagnosing these diseases. Intensive research has been carried out on the detection, treatment and prevention of the disease with the spread of SARS-COV-2 DELTA also known as Covid19-Delta(C19-D), which was encountered in 2019 and severely affected the whole world. To increase the success of the correct diagnosis and treatment of the lower respiratory symptoms and compensate for the lack of manpower, AI techniques are utilized. For this purpose, X-ray imaging devices which are available in almost all hospitals, was targeted. In this study, the effect of ensemble learning methods on the detection of normal, covid19 and viral or bacterial induced pneumonia with machine learning techniques is investigated. For this purpose, publicly available and labeled data belonging to these classes were collected and combined. Data for normal and non-covid pneumonia were also obtained from data published before 2019 to eliminate the possibility of mislabeling. Models were trained on balanced and unbalanced data sets. As a result of these trainings, no significant change was observed between the two distributions. The single SVM model achieved the highest success with 94.5 percent overall accuracy and 0.36 percent standard deviation. Bagging SVM took the second place. A 2% improvement was observed between machine learning models and ensemble learning techniques in favor of ensemble learning techniques, except for the SVM method. This may be due to the fact that the estimation is more successful in the class with more samples, since the distribution of training and test data is the same.

II. Setup of the Experiment

Experiments were set up for two dataset distributions. In a balanced distribution, all classes have an equal amount of but randomly selected samples, whereas in an unbalanced distribution, classes have a randomly selected amount of samples between 20 percent and 80 percent of their maximum

sample number. In this way, the distribution is proportional to the total number of samples that the classes have and they have different samples in different numbers. For the balanced and unbalanced distributions, 5 different dataset were created with the shuffling method for the given seed values. Each method trained on these 5 different distribution and their 5 run average accuracy metric along with their standard deviations were calculated. The models were run in parallel on 24 processors due to the high computational power required.

III. Data

For the task chest X-ray radiology images collected from publicly available resources. Many publicly available dataset for pneumonia detection task partly sharing the same x-ray images due to Covid19 detection competitions. For this reason we used dataset published before Covid19 pandemic for normal and Viral & Bacterial Pneumonia frontal chest X-ray images to eliminate the possibility of duplicate images or mislabelling [3]. The collected dataset contains 3616 Covid19 [4], 1583 Normal, 4273 Bacterial or Viral Pneumonia frontal chest X-ray radiology images which is known to be labeled by radiologists.

A. Data Pre-processing

All the images were resized to $(224 \times -1 \times n)$ or $(-1 \times 224 \times n)$ where n is the number of channels of the image and -1 is the any number of pixels for non square images. Then images were center cropped and transformed into 1 channel gray images using Python Pillow library.

B. Balanced and Unbalanced Distribution

For the experiments two different type of dataset were created. Balanced dataset Table I, contains equal amount of samples for each class. The amount of samples to be selected for the classes in the unbalanced dataset Table II, is randomly selected to resemble the original distribution. Each class is represented by a randomly selected amount between 20% and 80% of the total number of samples included in the original dataset . In this way, it is aimed to test whether the initial distribution affects the success of the ensemble learning method. The same seed values 26, 43, 738, 984, and 1437 were used for the balanced and unbalanced datasets. Training, validation and test split ratio is as follows: Training set contains approximately 90%, validation set approximately 10% of the data. Test Set contains exactly 10% of the data.

TABLE I: Class Sample Sizes of Each Run for the Balanced Dataset.

Run No	Covid	Normal	Pneumonia
1	405	405	405
2	567	567	567
3	810	810	810
4	972	972	972
5	1282	1282	1282

TABLE II: Class Sample Sizes of Each Run for Unbalanced Dataset

Run No	Covid	Normal	Pneumonia
1	955	884	2170
2	2178	511	1487
3	1351	881	1165
4	1567	758	2565
5	1038	1020	2550

IV. Models

In the experiments, machine learning methods and models obtained by applying ensemble learning techniques to these methods were compared. Class accuracy and overall accuracy metrics were used for comparisons. The parameters used for each model are given in the Table III, Table IV. Due to the high computing power requirement, the widely used python scikit-learn library with the highest performance optimization was used for the tests. The same random state value 42 was used for each model.

TABLE III: Running Parameters of the Machine Learning Models

Method	Random State
SVM	42
DT	42
kNN	-

TABLE IV: Running Parameters of the methods with Bagging

Method	Random State	Estimators	Bagging Estimators
AdaBoost	42	100	-
Bagging AdaBoost	42	100	10
Bagging SVM	42	-	10
Bagging DT	42	-	10
Random Forest	42	-	100
Bagging kNN	42	-	10

V. Results and Analysis

A total of 100 runs were obtained with 5 different balanced dataset distributions Table I for 9 different models, and 5 different unbalanced dataset distributions Table II for 11 different models. The methods used do not have implementations running on the GPU. Thereforey CPUs are used. Training of 100 models was with the parallel distribution of the tasks to 24 processors.

A. Training on balanced dataset

Among all methods, Single SVM gives the highest test dataset overall accuracy, Table VI. The SVM method is also the most balanced and high-accuracy yielding method among class-based accuracies. The second highest method, Bagging SVM method, has similar features. It is observed that Desicion Trees overfit during training. This is due to the implementation of the desicion tree. With a sufficiently complex desicion tree model, the task of 3 class classification can be overfitted easily.

TABLE V: Ensemble Model Accuracy trained on Unbalanced Datasets

	Overall Accuracy(%)			Class Accuracy(%) Test Dataset		
Method	Train.	Val.	Test	Covid	Normal	Pneumonia
SVM	97.06	94.69	94.50	97.83	85.00	96.18
DT	100.00	81.74	82.38	86.99	66.80	86.02
kNN	93.52	91.95	90.78	93.72	70.60	97.07
AdaBoost	93.08	88.47	88.77	92.92	79.00	90.00
Bagging SVM	96.84	94.65	94.46	97.95	85.00	96.10
Bagging DT	99.55	89.38	88.86	93.61	76.00	91.22
Random Forest	100.00	92.94	92.13	96.92	78.40	94.55
Bagging kNN	93.41	91.55	90.94	93.72	71.20	97.15
Bagging AdaBoost	94.11	91.17	90.8	93.95	78.20	93.82
SVM+kNN+DT-voting	97.76	93.79	93.3	97.49	78.4	96.59
SVM+AdaBoost-voting	94.66	91.41	91.56	98.40	86.8	88.86

* The data was sampled 5 times for 5 different seeds [26, 43, 738, 984, 1437] for 5 different label sample distributions. The reported accuracies are the mean values of the 5 runs. Standard deviation of the overall train, validation, and test accuracies are between 0% and 2%.

Although the data is balanced Table I, it is observed that the success of Normal Chest X-ray image in class based accuracies is almost always lower than the others. This may be due to the quality of the data. Or, the fact that the total number of samples belonging to the Covid and Pneumonia class is higher than the Normal Chest X-ray images may cause the model to see proportionally more samples of the disease. It can be assumed that the Covid dataset contributed to Pneumonia's training process, vice-versa.

TABLE VI: Ensemble Model Accuracy trained on Balanced Datasets

	Overall Acuracy(%)			Class Acuracy(%) Test Dataset		
Method	Train.	Val.	Test	Covid	Nor.[a]	Pne.[b]
SVM	96.36	94.33	93.31	98.59	90.96	91.38
DT	100.00	76.68	77.77	80.72	75.10	78.36
kNN	92.54	90.05	89.40	91.97	82.73	93.59
AdaBoost	92.93	87.94	86.64	89.36	84.74	86.37
Bagging SVM	96.10	94.50	92.97	98.59	90.16	91.18
Bagging DT	99.48	88.66	86.62	91.57	84.74	84.97
Random Forest	100.00	92.19	90.25	94.58	89.16	88.18
Bagging kNN	92.55	89.61	89.48	92.97	81.93	93.99
B. Adaboost[c]	93.63	90.83	89.32	92.57	86.55	89.18

[a,b,c] Normal, Pneumonia, Bagging Adaboost

* The data was sampled 5 times for 5 different seeds [26, 43, 738, 984, 1437] for 5 different sample sizes [500, 700, 1000, 1200, 1282]. The reported accuracies are the mean values of the 5 runs. Standard deviation of the overall train, validation, and test accuracies are between 0% and 2%.

B. Training on unbalanced dataset

For the unbalanced dataset, the Single SVM model has the highest accuracy values among all methods. The boosting method did not make a statistically significant contribution compare to single models. Models using more than one technique and voting methods, could not pass the single SVM model performance.

Looking at the unbalanced dataset, it is observed that the randomly selected sample size and samples sizes of the original dataset have the similar distribution. By looking at the test results of the models, it can be interpreted that the accuracy is similar to this ratio. It can be said that class accuracy ratio is in parallel with the sample size distribution. However, the same behavior is observed in the results we obtained in our experiment for the balanced data set. Therefore, the low accuracy values for the class of Normal chest X-ray images cannot be explained with this observation only. In addition, looking at the Table V, it is observed that the Majority Voting method does not make a statistically significant difference. It can be hypothesized that the method increased the accuracy of better represented classes. More runs should be performed to test this hypothesis.

The results obtained for 5 different runs can be observed in Figure 1 before the mean value is calculated. The standard deviation values of the Table V is given in Table VII for training, validation, and test set accuracies.

TABLE VII: Ensemble Model Accuracy Standard Deviation values trained on Unbalanced Datasets

	Model Accuracy Std.(%)		
Method	Train.	Val.	Test
SVM	0.36	1.23	0.92
DT	0.00	2.15	2.49
kNN	0.99	1.12	0.88
AdaBoost	1.65	1.68	1.91
Bagging SVM	0.47	1.29	1.33
Bagging DT	0.09	1.40	2.29
Random Forest	0.00	0.75	1.06
Bagging kNN	1.25	1.04	0.95
Bagging AdaBoost	1.05	1.88	1.67
SVM+kNN+DT-voting	0.33	1.15	0.92
SVM+AdaBoost-voting	1.31	2.24	1.27

* The data was sampled 5 times for 5 different seeds [26, 43, 738, 984, 1437] for 5 different label sample distributions. The reported standard deviation values belongs to mean calcualtion of the 5 runs.

VI. Conclusion

It can be seen that the SVM technique is the most successful technique for both balanced and unbalanced datasets. SVMs have been successfully used for feature extraction in image

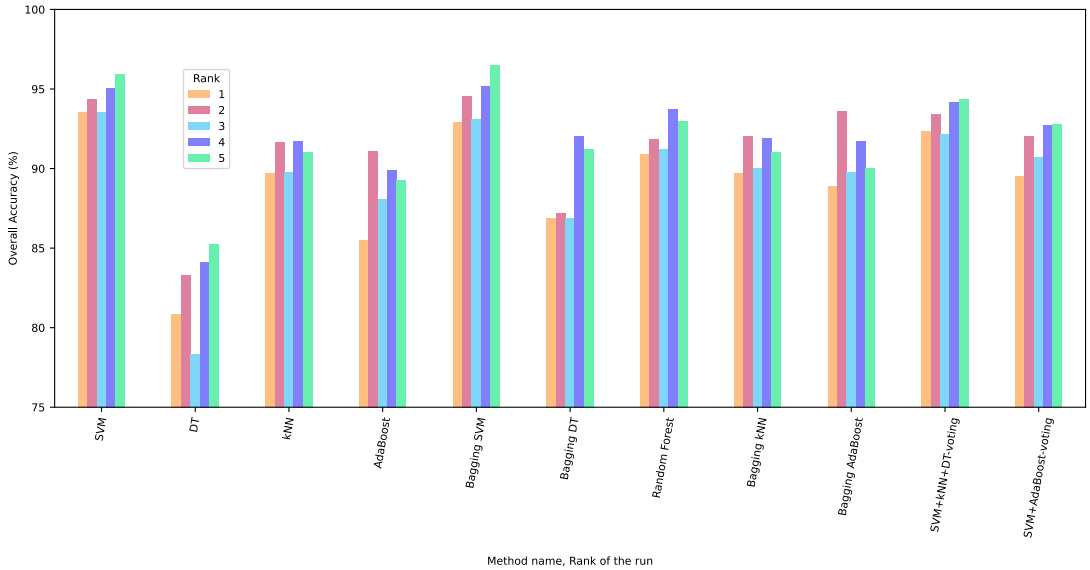

Fig. 1: Test set accuracies of each run for all methods.

classification and they appear to extract distinctive features for the chest X-ray image classification task [1]. Decision tree, and the Random Forests have the lowest accuracy due to their nature. Decision tree can overfit without proper pruning techniques. However, it can be seen that the random forest solves the generalization problem caused by training a single decision tree to a great extent. The boosting method couldn't outperform the single models alone. There was no statistically significant increase in the techniques in which the Bootstrap Aggregation method was applied. The results obtained with the Bagging method and the Majority Voting method did not amortize the computational cost of the technique used. The SVM method produced better results. It has been observed that the SVM gives good results in comparison studies for image classification and feature extraction. The superiority of the methods other than SVM over each other was not observed in this study.

REFERENCES

[1] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden and S. J. Johnston, "Random Subspace Ensembles for fMRI Classification," in IEEE Transactions on Medical Imaging, vol. 29, no. 2, pp. 531-542, Feb. 2010, doi: 10.1109/TMI.2009.2037756

[2] Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021, April). CheXtransfer. Proceedings of the Conference on Health, Inference, and Learning. doi:10.1145/3450439.3451867

[3] Mooney, P. (2018) Chest X-ray images (pneumonia), Kaggle. Avail- able at: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia?resource=download (Accessed: May 30, 2023).

[4] Rahman, T. (2022) Covid-19 radiography database, Kaggle. Available at: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (Accessed: May 30, 2023)