1. **Introduction**

**Student Number: 237707 & Kaggle Team Name: “world\_Hello”**

This report aims to investigate the various classification models and underlying techniques to purposefully realize fruitful features in our datasets. To address this purpose, we are provided three sets of training data based on photos taken in the summer time along with testing data packed with photos taken all year round. In this respect, we intend to not only understand each dataset as whole, but also identify the most important features in our datasets to increase the cross-validation score.

As a starting point we will be focusing on machine learning classifiers such as Support Vector Machines and Random Forest. Subsequently, experimentations will attempt to understand how the dataset(s) behave in response to our classifiers. Thereafter, experimentations will aim to determine potential feature selection and pre-processing techniques that can increase accuracy.

As with any machine learning project we anticipate trade-offs in respect to selecting the most effective classifier(s) as well as pre-processing and feature selection techniques. Nevertheless, we intend to use our experimentation to realize our core objective of attaining the highest possible accuracy score in the Kaggle competition.

## **Approach**

## **Choice of Classifier**

Tasked with classifying each and every instance in the dataset, classification can be considered as one of the fundamental tasks in machine learning. In this respect, the choice of classifiers can be seen as a vital aspect of our approach. Considering the trade-off that may be associated with various classifiers, we acknowledge that no single approach is perfect. Therefore, we intend to firstly establish the strengths and weaknesses of each classifier in respect to datasets provided before selecting one or a combination of classifier(s) for our final model.

* + 1. **Support Vector Machine**

Support Vector Machine is a classifier that has vastly yielded compelling outcomes in solving a variety of computer vision and pattern recognition problems. In the training procedure, the hyperplane is used to separate the training data associated with two classes with positions identified from a subset of the training data referred to as support vectors [3].

Despite the fact that the hyperplane is linearly separated, techniques can be used to solve highly non-linear problems through the kernel function [4]. The non-linear classifier can be represented as follows: , where represents kernel and represents the lagrange multiplier [3]. Among key parameters also include the choice of noise observation levels which will let the model make some misclassification yet increase the overall prediction accuracy.

One of the key drawbacks of SVM is that it requires substantial durations and memory when dealing with data sets with large dimensions [3]. Therefore, a greater may on quality over quantity may be effective in our approach.

* + 1. **Random Forest**

Random Forest (RF) is an unsupervised machine learning algorithm that uses an ensemble learning approach and can be used for both regression and classification problems. In contrast to other alternatives that rely on one single classifier, ensemble learning uses multiple models effectively resulting in delivering more accurate predictions [6].

Using random samples from the dataset, a decision tree is established for each sample. It is then tasked with the categorization of the sample. Thereafter, every single tree makes its own prediction and the classification occurs based on the category with the highest number of votes [5]. In this respect, one key characteristic of RF is it doesn’t treat data in a biased manner considering it relies on the overall number of votes in finalising its predictions. Therefore, the incorporation of additional data would not significantly affect the overall prediction. On the other hand, larger number of trees can also increase the performance [7]. Moreover, diversity within the data provided has shown to increase the performance [6] suggesting that it can effectively deal with combinations of features within various datasets.

RF can select many kernels. The non-linear kernel referred to as “Gini” can be represented as follows: , where represents the relative frequency of the observed class and represents the number of classes. Accordingly, the number of trees can be pre-determined with a higher number of trees increasing the accuracy.

RF can be computationally expensive due to its complexity compared to its counterparts. Regardless, the dimensionality reduction method embedded in RF can deal with ambiguous values such as blank data and can generally be known to convey compelling results [5]. Having said that, RF is also highly compatible and can be combined with other classifiers such as SVM through a stacking classifier.

1. **Method**

Considering we have 3 kinds of data, we will attempt to use each data on a separate model and use the one that best fits for our classification task. Having said that, we intend to use cross validation scores to estimate the accuracy of our different classifiers. Cross validation impartially rearranges the training dataset into subsets which are trained against the combination of all other subsets [2]. The average error of all the subsets can then effectively project the accuracy of our classifiers. Therefore, it will give us an indication of how accurate the classifier is.

## **Datasets**

Among the datasets provided include the CNN (4096 x 657) and GIST (512 x 657) features as well as an additional dataset (4608 x 5909) covering both GIST and CNN features and includes blank features. GIST features can be articulated as representation of “low dimensional images containing enough information to identify the scene of an image” [1].

Although the GIST features represent a mere 11% of the total data, it may well be with exploring considering our classification tasks focus on the setting of images to identify if the images are taken in summer time.

Among the 657 rows in the CNN and GIST dataset only186 of 657 have confidence scores 100% with the rest of the data at only 66%. Having said that, deleting parts of the training data with low confidence scores will certainly be an avenue to explore. Similarly, in the additional dataset, only 1567 of the 4342 rows have a confidence score of 100% with rest at 66%. Dubbed as an unavoidable problem in real-life datasets, missing data can be a cause of forgotten/misplaced data, inapplicability of the feature or the purposeful neglect by the observer [2]. In this respect, separation of the data with 100% confidence can be an avenue to explore.

* 1. **– Classifiers**

Considering we have two datasets which including two feature types (CNN & GIST), we will attempt selectively use numerous combinations dataset(s) to against our classifiers to determine their strength and weaknesses in respect to our task. Having said that, we intend to use cross validation scores to estimate our performance. Cross validation impartially rearranges the training dataset into subsets which are trained against the combination of all other subsets [2]. The average error of all the subsets can then effectively project the accuracy of our classifiers. Therefore, it will give us an indication of how accurate the classifier is. We will experiment with the following datasets on both of our classifiers:

1. CNN & GIST separately.
2. CNN & GIST together.
3. CNN with inclusion of additional data.
4. GIST with inclusion of additional data.
5. CNN & GIST with inclusion of additional data with 100% confidence.
6. GIST with inclusion of additional data with 100% confidence.
7. Combination of all the data with 100% confidence (CNN, GIST, additional data).
8. Commination all the data with 100% confidence (CNN, GIST, additional data).
   * 1. **– Support Vector Machine**

The kernels available to support vector machines include Linear, Polynomial as well as Gaussian. We aim to identify the most suitable kernel, using the cross-validation accuracy. Subsequently, we also aim to determine whether to implement various noise observation levels and try a variety of values in order to determine its impact on the cross-validation score.

Among various pre-processing techniques include normalisation, binarization as well as standardization and scaling. Respectively, we intend to identify the most effective technique. Therefore, each technique will be implemented against the CNN and GIST feature considering they are composed of any blank suggesting greater reliability.

Feature selection can also play an instrumental role in removing unwanted/unimportant features in the dataset. In this respect, we will be exploring the use of univariate feature selection based on filtering. A narrowed down approach in respect to the additional data can also see us with only 1567 rows consisting of 100% confidence rates. This may increase the accuracy of the predictions by focusing on higher quality and more reliable data as opposed to the quantity of data.

* + 1. **– Random Forest**

The number of trees selected in RF can be highly influential in the accuracy. Given our aim to have the optimal accuracy for our model, experimentations of various figures can be helpful in finding the optimal parameters.

In contrast to the SVM model which can be biased, RF is not biased meaning that a diverse set of data may increase the accuracy score. Hence, the experimentation of various mixtures of the datasets can be beneficial. Furthermore, RF is known to handle blank/ambiguous data very well. Seeing that the additional data contains 829 – 1019 bank points in each row, various experimentations can be conducted using the additional data. Having said that, we aim to use RF in order to foresee the ambiguity in our datasets from a different angle.

**3.2.3 – Combining RF and SVM Classifiers**

Once analysis is concluded we will attempt to use our two best configurations with both the RF and SVM classifiers. Considering our datasets include the extra GIST feature as well the additional data, RF’s ability to deal with ambiguity may add to the overall accuracy. This can be done using the stacking classifier. However, it would only limit us to the linear kernel for SVM. Moreover, we would not be able to implement our feature selection technique for SVM without impacting RF.

1. **Results and Discussions**

In establishing the best parameters for the SVM classifier, both Polynomial and Gaussian kernels gave similar cross-validation scores of 73% and 74%. In contrast, the linear kernel only scored 67%. On the other hand, we found to be the optimal noise observation parameter. Meanwhile, experiments with the RF classifier saw the increased number of trees as a performance boost and we witnessed an increased performance when 500 trees were applied. Subsequently, various strategies were used for filling the missing data and we found the “mean” approach to be most effective.

Although pre-processing techniques and feature selection do not cohere to RF, promising results emerged in our experiments using SVM. The “K-Best” based on “chi2” feature selection technique increased the cross-validation by 1% and the use of 2000 sample produced the best results. Although binarization and standardization were attempted, the best results came from binarization and scaling with cross-validation scores of 76%.

Both classifiers performed similarly in response to the combination of CNN and GIST features at 76%, SVM led with an additional 1% scoring 76% with the CNN features alone. The GIST features alone saw RF scoring 70% in contrast to SVM at 54% further suggesting the GIST features yielded better results on RF. Similarly, the additional data also attained greater accuracies with RF at 81% compared to the 75% SVM scored. However, the combination of both the additional dataset and GIST features saw a decline in performance with cross-validation reduced to 68% compared to 70% scored without the additional data on RF. Likewise, a small loss in performance was also witnessed when the additional dataset was combined with the CNN features.

Despite superior results with RF on the additional dataset and GIST features, we concluded the datasets consist of more unreliable than reliable data. Therefore, we submerged the best 1567 rows with 100% percent confidence along with CNN and GIST features. Improved scores of 85% on RF and 86% on SVM were attained. Considering the drastic boost in performance, we also extracted 186 rows from the CNN and GIST features effectively using all data with 100% confidence. Hence, increasing accuracy on RF to 89% and SVM to 90%. Drawing back on our earlier remark which drastically suggested that RF outperforms SVM on the GIST dataset, we decided to combine the two models together using the stacking classifier. Respectively, we were able to attain an accuracy score of 91% addressing our core objective of maximizing accuracy.

Overall, this experiment addressed its core aim of assessing the suitability of the datasets in response to the RF and SVM classifiers. Although the trade-offs incurred were far from convenient, our approach to the task was insightful in finding the practicality of the data with high confidence in respect to both classifiers. Nevertheless, our model could increase its performance given the explorations of combined classification methods. Such that non-linear kernel are included and less limitations seen in respect to feature selection as well as pre-processing (despite the availably of scaling was used in our model).

**References (in APA style)**

1. Oujaoura, M., Minaoui, B., Fakir, M., El Ayachi, R., & Bencharef, O. (2014). Recognition of isolated printed tifinagh characters. International Journal of Computer Applications, 85(1).
2. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159-190.
3. Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*, *52*(2), 857-900.
4. Nieto, P. G., Combarro, E. F., del Coz Díaz, J. J., & Montañés, E. (2013). A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. *Applied Mathematics and Computation*, *219*(17), 8923-8937.
5. More, A. S., & Rana, D. P. (2017, October). Review of random forest classification techniques to resolve data imbalance. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 72-78). IEEE.
6. Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, *2*(1), 602
7. Devetyarov, D., & Nouretdinov, I. (2010, October). Prediction with confidence based on a random forest classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 37-44). Springer, Berlin, Heidelberg.
8. Bernard, S., Heutte, L., & Adam, S. (2010, August). A study of strength and correlation in random forests. In *International Conference on Intelligent Computing* (pp. 186-191). Springer, Berlin, Heidelberg.

|  |
| --- |
|  |
|  |
| Chicago |  |

|  |
| --- |
|  |
| Chicago |  |