Student Number: 251122

1. Approach

The machine learning approach utilised in this project was a multi-stage pipeline involving multiple pre-processing techniques. Initially, missing values were imputed using the median of existing values, assuming the data to be Missing Completely at Random (MCAR). This robust statistical approach reduced bias.

To better understand the data distribution and feature relationships, we visualised the data using scatter plots and T-SNE techniques.

Next, the data was standardised to bring all features to the same scale. This is critical for certain algorithms that are sensitive to feature scales, helping to eliminate any undue influence from features with larger scales.

PCA was applied to address the challenge of high dimensionality. This technique transformed the original features into a new set of uncorrelated features, arranged in descending order of variance retention. By reducing the data set's dimensions, it mitigated the curse of dimensionality and helped prevent overfitting.

A grid search was then conducted to optimise the hyperparameters of the Support Vector Classifier (SVC). This exhaustive search technique operates under the assumption that the explored parameters lead to the best model performance. The SVC was selected due to its ability to handle high dimensional data and its effectiveness in dividing the data set into clear margins of separation using hyperplanes. Here, we utilised an SVC with a radial basis function (RBF) kernel, which has the ability to handle non-linear data by mapping it into higher dimensional space.

2. Methods

2.1. Combining Training data & Filling NaN values

First and foremost, we amalgamated two training data files: the initial file comprised 500 samples with complete data, while the subsequent file contained 2,500 samples.

Subsequently, we addressed the missing values in the training data by replacing them with the median of the corresponding feature, a method which preserves the overall distribution of each attribute. Furthermore, by merging these files earlier, we obtained a comprehensive data set of 3,000 samples, which enhanced the imputation method for handling missing values. As this larger data set enabled the imputer to make more accurate predictions.

Also, increasing the number of samples in data improves model's accuracy by capturing more diverse patterns and reducing over-fitting.

2.2. Model Selection

SVM showed the best results in compared with others as shown in Table 1. [1] [2] [3]. The simulation was done by using 5 fold cross-validation, it was chosen over a simple training-validation split for its advantages in providing a robust performance estimate and hyper-parameter tuning assistance. Moreover, cross-validation was employed to ensure a more accurate performance evaluation across multiple tests, which helps to reduce over-fitting and identify models that generalise well to unseen data.

2.3. Scaling the data

The data set contains two distinct feature types, CNN and GIST, with varying magnitudes. As some algorithms, like SVM [1], are sensitive to feature scales, applying feature scaling was essential to improve performance and convergence. Feature scaling brings features into comparable ranges, easing the learning process for algorithms and enhancing model interpret-ability.

Both Min-Max Scaling (normalisation) and Standard Scaling (standardisation) were evaluated using an SVC model. The accuracy obtained for Standard Scaling was 76.67%. Standard Scaling offered slightly better performance and was selected as the feature scaling method for the final model. Standardisation helps in giving equal importance to all features, preventing features with larger values from dominating the model's learning process [4].

2.4. Feature Selection & CNN vs GIST features

The dataset used for training encompasses 2,304 features derived from Convolutional Neural Networks (CNNs), introducing a high level of dimensionality. PCA (Principal Component Analysis) [5] was used for CNNs features, to reduce the dimensionality while preserving as much of the original data's variance as possible. PCA works by finding new directions (principal components) in the feature space along which the data variance is maximised [6]. The transformed data in the new directions is then used as the reduced-dimensionality features. This is often an effective technique for high-dimensional, dense features like from CNNs.

Contrarily, GIST features, being more structured and fewer in number, do not necessitate any dimensionality reduction. Also, use of GIST features in various applications, including image classification, have been proven beneficial in a large variety [7].

To assess the relative importance of CNN and gist features, we fed different classifier algorithms; only GIST features (F_{GIST}), only CNN features (F_{CNN}), and lastly, combination of CNN and GIST features $F_{Combined}$ =

$concatenate(F_{CNN}, F_{GIST})$. A five-fold cross-validation revealed that (F_{CNN}) and $F_{Combined}$ performed equally well, particularly with the SVM classifier. However, since GIST features proved that they perform good in classification scene [8], $F_{Combined}$ was decided to be used.

2.5. Confidence labels

The process involved utilising the "confidence labels" to adjust the weights of training samples in a Support Vector Machine (SVM) model. The confidence labels were used to modify the 'C' hyper-parameter, which in turn adjusts the soft margin of the SVM. The model emphasised high-confidence samples during training, but it had no impact on performance. Since it doesn't harm performance and may enhance the classification algorithm, leading to improved prediction accuracy on the test set, it was retained.

3. Results & Discussion

3.1. Selection of Algorithm Classifier & Features

The Support Vector Classifier consistently showed the best performance across different feature sets, achieving the highest accuracy. The CNN features alone resulted in higher accuracy compared to GIST features alone. However, combining GIST and CNN features did not significantly improve the performance, as the accuracy remained similar to using CNN features alone. Random Forest and XGBoost performed relatively well but had slightly lower accuracy compared to SVC and showed similar performance across different feature sets.

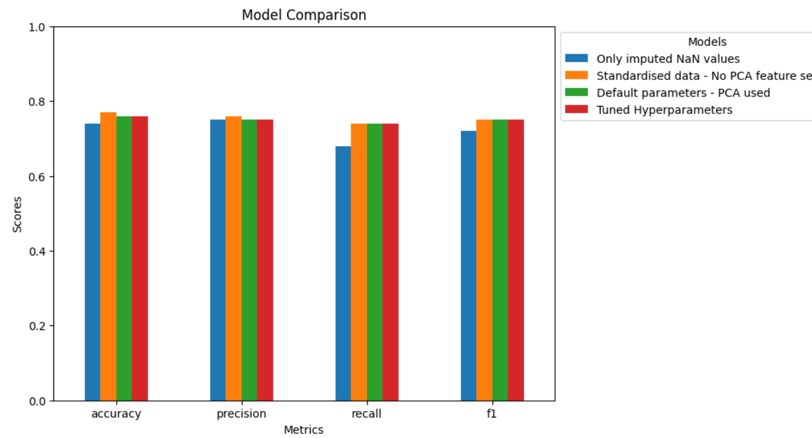
	KNN	SVM	Random Forest	XGB
F_{GIST}	60.9%	65.97%	66.57%	66.53%
F_{CNN}	69.80%	74.27%	73.73%	72.93%
$F_{Concatenate}(F_{GIST}, F_{CNN})$	69.80%	**74.27%**	73.70%	74.17%

Table 1, Accuracy rates of different algorithms and features

3.2. Data set pre-processing comparisons

The "Only Imputed NaN values" approach focuses solely on imputing missing values without additional pre-processing. It has moderate performance but potential for improvement. The "Standardised data" approach shows enhanced performance compared to the previous method. Precision and recall are both approximately 0.75, indicating balanced prediction performance. In the "Default parameters - PCA used" approach, PCA is applied to the data without hyperparameter tuning. The results resemble those of the previous approach, suggesting that PCA may not significantly enhance the model's performance in this case. The "Tuned hyperparameters with PCA" approach utilises grid search for hyperparameter optimization. However, the results remain unchanged from the previous approach, implying that the hyperparameters might not have significantly impacted the model's performance.

Comparing the approaches, it seems that the standardized data without PCA feature selection achieved the highest accuracy (76.67%). However, the differences in performance between these approaches are small, suggesting that the choice of approach may depend on other factors such as computational efficiency, and the specific characteristics of the data set.



graph 1, The presentation of accuracy, precision, recall, and f1 for a classifier's performance on training sets before, during, and after feature engineering using SVC classifier

3.3. Ways of getting better performance

Using more advanced imputation techniques such as multiple imputation, KNN imputation, or using models like random forests to predict missing values. These methods might have potentially provided more accurate imputations, leading to better model performance.

could have tried different pre-processing techniques: Besides standardization, other pre-processing techniques like normalization, logarithmic transformation, or discretization could be applied to the data. Experimenting with different pre-processing strategies may lead to better results.

Lastly, having more data means improved model performance by capturing diverse patterns, mitigating outliers, enabling deeper analysis.

3.4. Ways of getting a better job of evaluation

ROC curves could have provided a comprehensive evaluation of the model's performance by considering various classification thresholds. Plotting the ROC curves and calculating the corresponding area under the curve (AUC) values could have given a more detailed understanding of the data's trade-offs between true positive rate and false positive rate.

References

[1] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011. 1

[2] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, 1999. 1

[3] V. Vapnik, *The nature of statistical learning theory*. Springer New York, 1995. 1

[4] FAQs.org, "Ai faq: Neural networks, part 2," *Journal of Foo*, vol. 12, no. 1, p. 2, 2002. 1

[5] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987. 1

[6] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008. 1

[7] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, (New York, NY, USA), pp. 19:1–19:8, ACM, 2009. 1

[8] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, May 2001. 2