Feature Selection, Reduction and Classifiers using Histogram of Oriented Gradients: How important is Feature selection?

Abstract

Facial Expressions are one of the main methods we use to express our emotions to others. Yet Facial Expression Recognition (FER) remains a difficult topic for machines to intrepret. While Computer Vision can extract features quite easily from imagery, there is still the difficult step of recognizing what emotion those features belong to. Many have taken to Deep Learning to bridge this learning gap. However this paper shows that with selected features, even classic techniques without modification can achieve high accuracy. This paper demonstrates how select features, taken from ANOVA, LDA and PCA, enhances the accuracy of HOG without further processes.

Keywords: Facial Expression Recognition, Feature Selection, Feature Reduction, Machine Vision, Machine learning

1 Introduction

Facial expressions, within an emotional context, are an essential aspect to communication [Mehrabian, 1981]. For this reason, attempts to expand Human-Computer Interaction with facial expressions have existed for nearly three decades [Mase, 1991]. Several different methods have developed over the years. Wavelet based methods were a common technique in early methods [Mase, 1991] [Zhang et al., 1998]. Since then, Local Binary Patterns (LBP) and its variations, have largely superseded or used in combination with wavelets for some time with high accuracy [Tyagi et al., 2017]. Histogram of Oriented Gradients (HOG) is another example of a feature extraction method, utilising the direction of pixel gradients as features for the image. This is the method settled for this paper for testing due to the high accuracy that could be obtained using HOG alone [Gritti et al., 2008]. Many contemporary research has focused on Deep learning, but this paper intends to find if it is possible for older methods to use feature selection or feature reduction to gain comparable results. Is it possible to obtain higher accuracy by analysing and reducing the features to a select few with the most variation that corresponds to their emotional label, and can that have a learning time reduction benefit?

Feature selection and reduction have been common techniques in Bioinformatics for some time due to the large number of factors involved in biology systems [Saeys et al., 2007]. The size of the feature set in FER is dependent of the extractor and its parameters. Nevertheless, the size of the feature set will affect the computational time, but it could also have an affect on the accuracy as well, particularly with significant noisy or irrelevant features. The goal of the feature selection or reduction is thus, reduction of the feature set, improvement in accuracy and reduction to computional time. These are the factors that are tested for in this paper and how the performance of the methods are measured. For this the three feature selection and reduction methods chosen were Linear Discriminant Analysis (LDA), Analysis of Variance (ANOVA), and Principle Component Analysis (PCA) for close relation to each other, giving us an understanding of the data based on their output. LDA and ANOVA each attempt to maximise variance between the classes while reducing variance within the class [James et al., 2014], while PCA maximises the overall variance [James et al., 2014]. LDA and PCA are similar in that they reduce the data by transforming it as opposed to selecting a subset of features from the original feature set.The final selected features are then passed into one of the classifiers. The seven classifiers that are investigated are K-Nearest Neigbour (K-NN), three different kernels for a Support Vector Machine (SVM), Decision Tree, Random Forest, and the Gradient boosting Tree.

2 Feature Selection and Feature Reduction

Feature selection and Feature reduction are an important aspect in data analysis and machine learning. For Feature Selection, it assumes that a smaller subset of the feature set can produce better accuracy due to the high variation of the smaller subset. Feature Reduction instead projects the feature set onto a lower dimension. From this projected feature set a number of components can be selected which has high variation that represent the unique aspects of the output label, in this case the emotion of the image. For testing three different methods to select or reduce the data, ANOVA, LDA and PCA are used on features extracted from facial images using HOG. These three methods are closely linked mathematically but look at the data in different ways. Due to the close similarity these methods have, the results from classification can be used to interpret the feature set to conclude an optimal method at reducing or selecting features.

ANOVA F-test was used, with n-features selected which has the most variance between the seven different classes. ANOVA is similar to LDA in that it also attempts to express the dependent variable, in this case the emotion, through the particular features that exist in the image. ANOVA calculates the f-value which is used to rank features. First, it calculates the mean of each sample and compares it to the mean of everything as the between-group variability. This is calculated in equation 1 below.

$$\sum_{i=1}^{K} n_i \left(\frac{(\bar{Y}_i - \bar{Y})^2}{K-1} \right) \tag{1}$$

$\bar{Y}_i$ being the sample mean from a sample size of 151, $\bar{Y}$ the total mean of everything which is dependent of the image database, n_i the number of features in each sample, and K is the number of samples.

Variation within the samples, called within-group variability, is calculated as in equation 2.

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(\frac{(Y_{ij} - \bar{Y}_i)^2}{N-K} \right) \tag{2}$$

N being the number of features in a sample. Finally the f-value is calculated by dividing the between-group variance (equation 1) by the within-group variance (equation 2). The value selected for n-features, is the number of components that explain 95% of the variance.

LDA, is a method that is also used for classification [Close et al., 2016]. However it also ranks its components, giving us a useful tool to select the features that has the most variance between the different classes. LDA takes in the features as linear data and treats them as continuous independent variables that point to a dependent variable, the emotion these features relate to. It is assumed that the features will has a normal distribution, where other methods, such as ANOVA, do not make that assumption. LDA uses Bayes' rule to make predictions. $P(X|y)$ has a density expressed in equation 3, and is modelled as a multivariate Gaussian distribution.

$$P(X|y=k) = \frac{1}{(2\pi)^{n/2}|\sum_k|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu_k)^t \sum_k^{-1} (X-\mu_k) \right) \tag{3}$$

With μ being the mean, k the number of samples and n the number of features. The LDA is only used to transform the data, rather than to fully classify it. The transformed data are then passed to one of the classifiers explained in the next section.

PCA is similiar to LDA but instead attempts to maximise the overall variance rather than maximising the difference between classes. Data are fitted to a new plane, coordinate system reduces the number of components to those that explain 95% of the variance as is the case in ANOVA.

3 Classifiers

A large number of classifiers can be selected for testing but three main groups were chosen: SVM, Decision Trees, and K-NN. Three versions of SVM were selected as an SVM classifier is commonly used to classify HOG features [Tyagi et al., 2017][Gritti et al., 2008][Alfalou et al., 2017], while the Decision Trees, and K-NN were selected based on their simplicity and speed. Both are fast methods which given the likely goal of FER is real-time application could be of use. Finally we also use two ensemble methods based on the decision tree to see if the accuracy can be improved.

SVM is a discriminate binary classifier that separates data using a hyper-plane to produce a single output. Since it is a binary classifier and the data is multiclass, this approach uses the one-versus-rest approach to classification, which trains one class as positive and all others as negative, repeating until classified. Tests use and compare three different kernels to find the most effective. These three kernels are the Linear, Radial Basis Function (RBF), and Polynomial. The linear kernel produces a straight boundary line between the two classes. The second kernel is the RBF which produce circular areas of classification. Finally a 2nd order polynomial is employed to produce another non-linear decision boundary to test accuracy if the data isn't linearly separable.

Decision trees are one of the simplest classification methods, they split the data into smaller and smaller groups based on differences until a clear class is found for the training data. It has a higher tendency for over fitting, but its ability to take in data without the need for a lot of preprocessing is highly advantageous. The simplicity of the decision trees is what contributed to its inclusion in this comparison.

To improve performance, the random forest takes splits the data up into groups in a divide and conquer approach, and then a decision tree is created for these data subsets. This allows for a boost in speed. The input data is run through all of the trees and the resulting output is based on what the majority of the subset data returned as the classification.

Gradient boosting makes use of multiple weak tree classifiers to generate a strong classifier. The weak tree classifiers are regression models as this outputs real values that can be added together. Weak classifier trees are added to minimise the loss function, in this case a logarithmic loss function.

K-NN is an instance based classifier, where 'training' is storing the feature vectors and labels. Output is based on the largest number of classes closest in the k neighbourhood. In this case the Euclidean distance is used to determine the seven closest neighbours, which was found to be large enough for good classification based on testing over various k-values. As this is a multi-class problem, odd numbered k-value will not prevent a situation where two or more k's of different labels have identical distance. There is no ideal method for deciding which to select, but the Scikit-learn implementation used to test, uses the ordering, in other words the first value in the list to select the winner.

4 Methodology

For testing three databases were selected, one of which was modified bringing the total to four. These are the CK+ (Extended Cohn-Kanade) database [Lucey et al., 2010], JAFFE (Japanese Female Facial Expression) [Lyons et al., 1998], and KDEF(Karolinska Directed Emotional Faces)[Goeleven et al., 2008]. For the fourth, KDEF, the category of fear was removed from KDEF. The makeup of the categories in KDEF gives emphasis to the fear emotion due to the number of fear images compared to other emotions. This leads to a miscalculation of facial images towards fear. Since KDEF without fear is included, testing is also performed on the original KDEF as well so that the reader can make their own judgment from the data. KDEF also includes the facial emotion from five different angles, but for testing this paper only uses the front angle directly facing the subject. CK+ differs slightly from JAFFE and KDEF by including contempt as a seventh emotional classification, based on the expanded emotional list Ekman proposed in the 90s [Ekman, 1999]. JAFFE and KDEF share the original Ekman's six emotions [Ekman and Friesen, 1971], anger, sad, fear, disgust, surprise and happy. With neutral this comes to seven categories for KDEF and JAFFE, and eight categories for CK+. For CK+, images are a sequence showing the change in expression. The image selected is the last image of each emotion, effectively at its' height for the data.

First the image is converted to gray scale before the face is found using the Viola-Jones method of Haar cascades [Viola and Jones, 2001] [Ding et al., 2017]. The image is then cropped to this facial area and resized to 130 by 130 pixels as the real size of the face varies from image to image. While different sizes can be used, where larger images typically increase accuracy, this size was used as it was smaller than all of the cropped JAFFE images which was the smallest image set, thus avoiding introducing any artifacts from scaling.

The next stage is feature extraction. A Histogram of Oriented Gradient is used as this has been shown in the past to achieve high accuracy by itself [Gritti et al., 2008]. HOG calculates the gradient orientation and intensity of 'cells' in the image, these are portion of the image, the size of which can be adjusted. This is with six orientation bins, with 24x24px cells. Then a block of cells, 5x5, is normalised giving the final histogram for the image. The whole set of histograms are then normalised giving the final data that will be used. ANOVA, PCA and LDA are employed separately to the same feature set to produce a new set, subset for ANOVA and transformed data for PCA and LDA. This new data is fed into each classifier in a K-fold manner, with the time for each fold recorded and averaged to measure the computational time.

5 Results

Machine Learning		LDA	ANOVA	PCA	None
KNN	**Time**	0.0013	0.0015	0.0019	0.0074
	Accuracy	97.77%	67.24%	64.28%	66.98%
SVM (Linear)	**T**	0.0021	0.0061	0.0158	0.0262
	A	98.21%	52.14%	66.02%	81.56%
SVM(RBF)	**T**	0.0059	0.0092	0.0225	0.0499
	A	96.38%	26.53%	26.54%	71.49%
SVM(Polynomial)	**T**	0.0024	0.0072	0.0180	0.0543
	A	96.84%	26.51%	26.52%	57.78%
Decision Tree	**T**	0.0655	0.0017	0.0116	0.2241
	A	97.98%	59.08%	47.83%	73.24%
Random Forest	**T**	0.0931	0.0929	0.1467	0.0364
	A	97.98%	72.57%	62.21%	60.44%
Gradient boosting Tree	**T**	0.1809	0.3114	0.5647	1.3842
	A	93.93%	67.88%	61.39%	72.14%

Table 1: CK+

Machine Learning		LDA	ANOVA	PCA	None
KNN	**T**	0.0008	0.0010	0.0009	0.0026
	A	100.00%	61.02%	42.77%	33.25%
SVM (Linear)	**T**	0.0008	0.0020	0.0041	0.0090
	A	100.00%	8.51%	44.11%	94.78%
SVM(RBF)	**T**	0.0039	0.0028	0.0051	0.0146
	A	100.00%	6.06%	8.51%	86.34%
SVM(Polynomial)	**T**	0.0008	0.0016	0.0034	0.0134
	A	100.00%	8.96%	10.74%	65.26%
Decision Tree	**T**	0.0547	0.0016	0.0016	0.1270
	A	100.00%	65.61%	43.23%	84.61%
Random Forest	**T**	0.0812	0.0710	0.0910	0.0120
	A	100.00%	67.12%	80.74%	54.46%
Gradient boosting Tree	**T**	0.1088	0.1373	0.2156	0.4970
	A	98.57%	63.94%	68.53%	70.84%

Table 2: JAFFE

Machine Learning		LDA	ANOVA	PCA	None
KNN	**T**	0.0034	0.0022	0.0196	0.0433
	A	70.19%	40.30%	47.14%	43.62%
SVM (Linear)	**T**	0.0110	0.0208	0.0715	0.1974
	A	74.50%	31.42%	46.40%	49.00%
SVM(RBF)	**T**	0.0210	0.0448	0.1308	0.2329
	A	71.99%	31.14%	31.16%	54.58%
SVM(Polynomial)	**T**	0.0242	0.0338	0.0999	0.2650
	A	65.35%	31.13%	31.15%	41.48%
Decision Tree	**T**	0.1139	0.0053	0.0360	0.5608
	A	56.90%	35.27%	34.10%	43.71%
Random Forest	**T**	0.1602	0.1605	0.3321	0.1006
	A	56.46%	31.42%	40.23%	37.07%
Gradient boosting Tree	**T**	0.2953	0.3447	1.2039	3.0452
	A	54.50%	28.18%	37.07%	42.19%

Table 3: KDEF

Machine Learning		LDA	ANOVA	PCA	None
KNN	**T**	0.0019	0.0022	0.0070	0.0191
	A	95.44%	64.25%	71.70%	68.07%
SVM (Linear)	**T**	0.0027	0.0115	0.0354	0.0476
	A	96.23%	51.61%	79.51%	87.49%
SVM(RBF)	**T**	0.0087	0.0208	0.0704	0.0951
	A	95.44%	19.05%	15.63%	86.32%
SVM(Polynomial)	**T**	0.0063	0.0115	0.0562	0.1188
	A	90.35%	11.73%	12.13%	73.52%
Decision Tree	**T**	0.0749	0.0031	0.0221	0.3463
	A	95.05%	54.50%	58.16%	83.45%
Random Forest	**T**	0.0993	0.0031	0.2199	0.0613
	A	95.05%	65.72%	77.98%	61.26%
Gradient boosting Tree	**T**	0.2042	0.2129	0.7626	1.9072
	A	94.00%	62.86%	76.80%	83.83%

Table 4: NF-KDEF

All classifications used a 10 k-fold cross-validation, making sure that different training and testing data was used for each test. Results in Tables 1 through 4 show accuracy and speed for each data set using the different feature selection/reduction and machine learning combinations. Times are the average of the folds.

Figure 4 displays the average times of machine learning algorithms for every feature selection and reduction method, by the average time taken for the folds. This shows the CK+ database but all databases showed similarity between their respective points. In CK+, PCA achieved 51.53% at 0.09s. LDA and ANOVA outperform

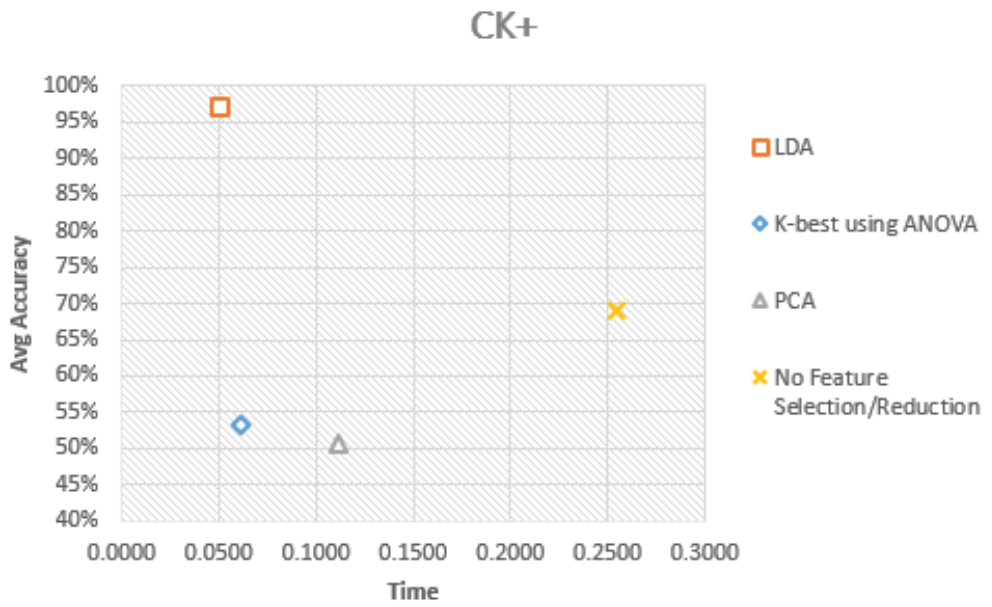

Figure 1: CK+ - Average accuracies by time taken.

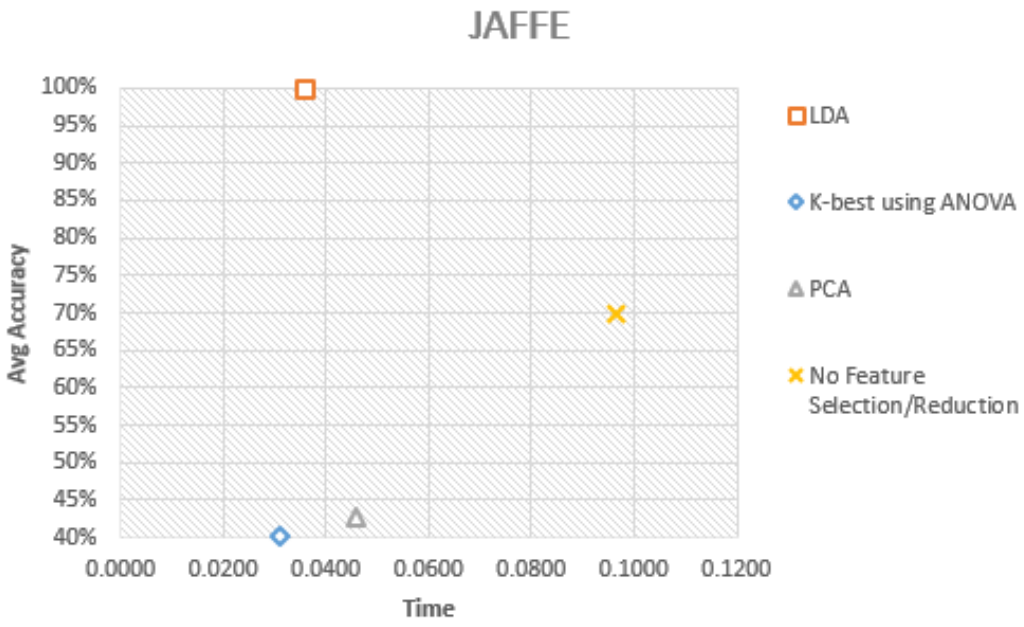

Figure 2: JAFFE - Average accuracies by time taken.

PCA in terms of speed and accuracy. Using CK+ again, averaging the accuracies and times gives LDA 96.31% accuracy within 0.032s, while ANOVA achieves 95.09% accuracy within 0.034s.

Figures 1 through 4 shows the average times of the machine learning algorithms for every feature selection/reduction method and the average accuracies they achieved for each of the three databases, and KDEF without fear. While K-Best using ANOVA was, on average, the fastest method, it was also typically the least accurate, followed by PCA. While both of these methods were at least one tenth of a second faster than no feature selection/reduction, it was also at least 5% less accurate. This is quite the contrast to the average LDA which not only was consistently less than one tenth of a second in per image classification speed, but also at least 10% more accurate than the average no feature selection/reduction method.

In terms of the data, the JAFFE database, the smallest of those tested, produced the highest accuracies for LDA, with an average of 99.8% for JAFFE and an average of 94.5% for KDEF without fear. However this

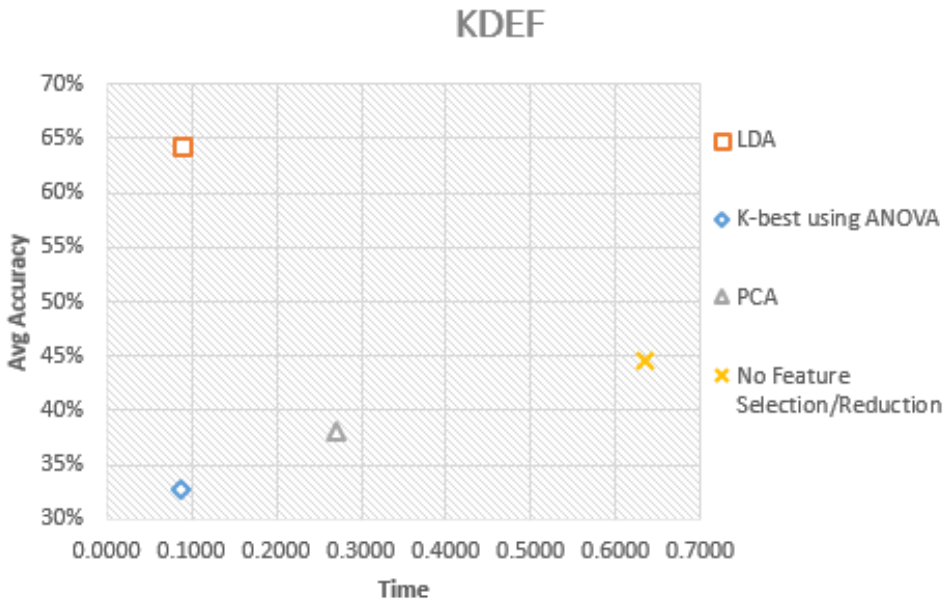

Figure 3: KDEF - Average accuracies by time taken.

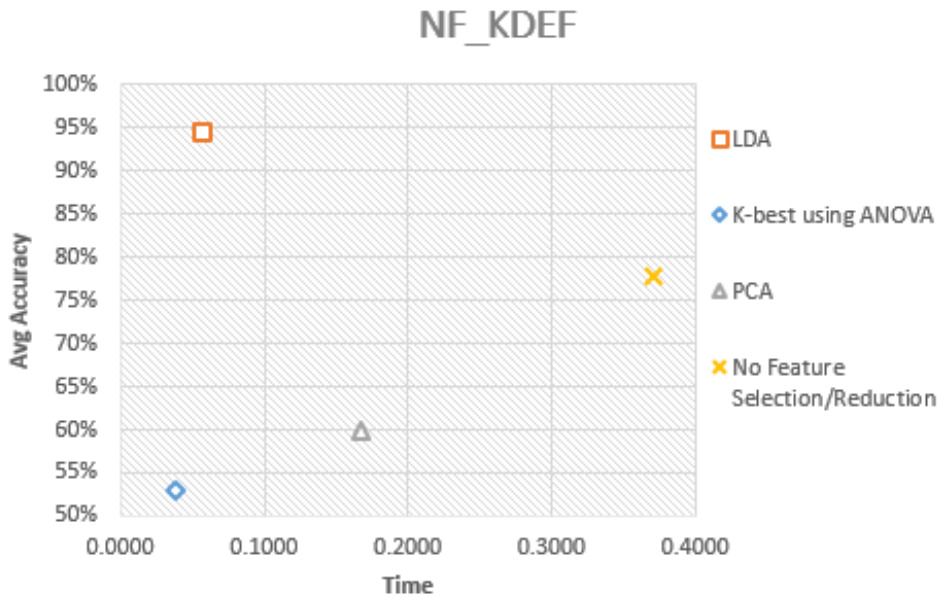

Figure 4: NF-KDEF - Average accuracies by time taken.

wasn't the case for PCA which performed better with more samples which had an average of 42.7% for JAFFE and an average of 59.9% for KDEF without Fear. Looking at the machine learning classifiers in Figure 5, the decision tree, on average, performed the best for JAFFE but was one of the least effective in KDEF. Taking into account overall accuracies, linear SVM was shown to be the best overall classifier for all databases. It also can be concluded that the data is linearly separable.

Looking at the machine learning classifiers in figure 5. Decision tree on average performed the best for JAFFE but was one of the least effective in KDEF. Linear SVM was thus the best overall classifier for all databases.

In terms of time, all the feature selection and reduction methods create an improvement, however, as seen by Figures 1 - 4 and Tables 1 - 4 only LDA creates an improvement consistently in both time and accuracy. Other studies [Gritti et al., 2008] have shown similar accuracy with linear SVM as to our HOG without feature selection or reduction, helping to establish our baseline. Even when compared to deep learning methods,

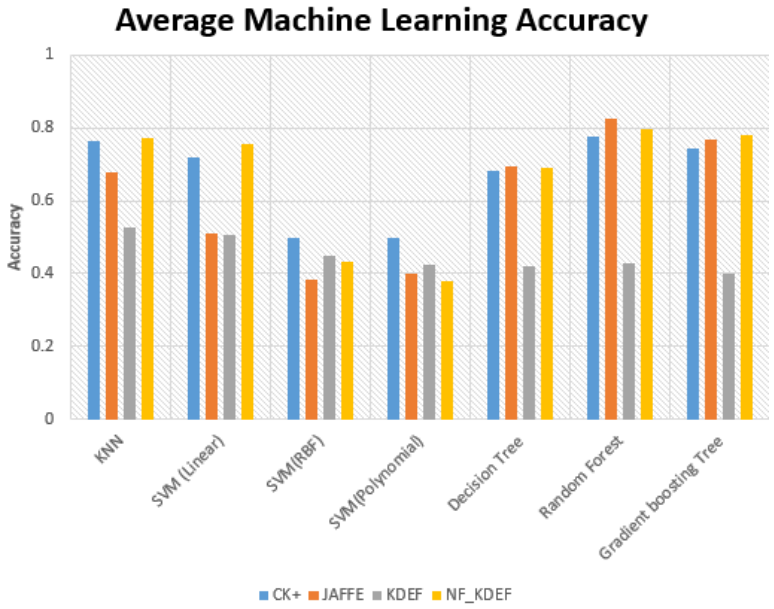

Figure 5: Accuracy by dataset and machine learning classifier.

with only the addition of feature selection,tests were are able to achieve accuracy results close to others. In [Ding et al., 2017] they developed 'FaceNet2ExpNet'. They used more data from the CK+ database, using more of the image frames available, common for deep learning methods, and achieved 96.8% on the eight emotion labeled CK+ database. Further testing would need to be done for a true comparison, but it's likely their robust setup would perform better feature selection alone. However the point of this study is to show that proper treatment of data can be enough to push performance without needing to change tools. In a non-deep learning comparison, in [Tyagi et al., 2017] , they were able to achieve with CLBP with Gabor filter and SVM, 95.23% with JAFFE and 97.61% with CK+.

Given that LDA performs the best compared with PCA and K-best using ANOVA, it can be concluded that there is enough variance between classes to make the distinction for classification. It can also be concluded that the features which create the most variance, which PCA transforms based on, are not the same as the features which are the most distinct between classes, which LDA transforms based on. The data also suggests that the transformation and reduction of data, such as in LDA, produces a better accuracy than simply selecting the best features, such as in ANOVA. Being that LDA out preformed the non-reduced data, PCA, and ANOVA in both speed and accuracy, it is the preferable feature reduction method. All testing was conducted with an Intel core i7-6700 3.40Ghz processor, 16GB ram on windows 10. Times are fast compared to learning times without any feature selection or reduction and well within the requirements that would be needed for real-time application. This makes it a suitable addition for embedded computing that uses limited resources.

6 Conclusion

Data is at the core of machine learning, and this study hopes to show the importance of utilising that data to the maximum extent. Feature selection is a key aspect to locating and isolating the core unique features that contribute to classification. For example in our CK+ linear SVM test, feature reduction increased accuracy by 18% and the average of each fold speed up by 150%, there is no doubt of the essential role that analysis of features has. As too the importance of the classifier itself. HOG is commonly used in conjunction with SVM, and this study showed that its use is appropriate but not without suitable alternatives. Often the decision tree and K-NN were able to achieve similar results. Regardless of classifier, where there is redundant or noisy data in a feature set, feature selection has a benefit, particularly if the end goal is real-time application in which as

few features as possible for classification is desirable.

References

[Alfalou et al., 2017] Alfalou, A., Ouerhani, Y., and Brosseau, C. (2017). Road mark recognition using hog-svm and correlation. *Optics and Photonics for Information Processing XI*.

[Close et al., 2016] Close, M., Abraham, P., Humphries, B., Lilburne, L., Cuthill, T., and Wilson, S. (2016). Predicting groundwater redox status on a regional scale using linear discriminant analysis. *Journal of Contaminant Hydrology*, 191:19–32.

[Ding et al., 2017] Ding, H., Zhou, S. K., and Chellappa, R. (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*.

[Ekman, 1999] Ekman, P. (1999). *Basic Emotions*, pages 45–60. John Wiley & Sons.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

[Goeleven et al., 2008] Goeleven, E., Raedt, R. D., Leyman, L., and Verschuere, B. (2008). The karolinska directed emotional faces: A validation study. *Cognition & Emotion*, 22(6):1094–1118.

[Gritti et al., 2008] Gritti, T., Shan, C., Jeanne, V., and Braspenning, R. (2008). Local features based facial expression recognition with face registration errors. *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*.

[James et al., 2014] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An introduction to statistical learning: with applications in R*. Springer.

[Lucey et al., 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*.

[Lyons et al., 1998] Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205.

[Mase, 1991] Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, E74-D(10):3474–3483.

[Mehrabian, 1981] Mehrabian, A. (1981). *Silent messages: implicit communication of emotions and attitudes*. Wadsworth Pub. Co.

[Saeys et al., 2007] Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[Tyagi et al., 2017] Tyagi, D., Verma, A., and Sharma, S. (2017). An improved method for facial expression recognition using hybrid approach of clbp and gabor filter. *2017 International Conference on Computing, Communication and Automation (ICCCA)*.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*.

[Zhang et al., 1998] Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*.