MLP vs SVM for Binary Classification

1. Introduction

Heart disease causes devasting effects around the world and has taken many lives with the CDC reporting approximately 659,000 deaths in the United States each year [1]. However, there has been large volumes of research over the last century regarding prevention and it has been proven there are numerous measures most individuals can take to reduce their likelihood of suffering heart disease, such as regular exercise [2]. Therefore, if it possible to predict if someone has or is at risk of heart disease, preventative and remedial measures can then be implemented to improve the person’s health. This dataset uses the Kaggle ‘Heart Disease Health Indicators Dataset’ [3] which contains data from survey questions regarding American citizen’s health and the target variable describing whether the patient has heart disease or has had a heart attack. Data from surveys are relatively easily obtainable; therefore, if the target variable can be predicted reliably, similar data could be collected in other countries and these predictions reproduced. This study implements Multilayer Perceptrons and Support Vector Machines to predict if a patient has had a heart disease or heart attack.

2. Dataset

This dataset contains attributes regarding the health and lifestyle of American patients and an attribute describing whether the target has had a heart disease or heart attack (HDoA). Only 9% of citizens had an experienced heart disease or a heart attack; **Figure 1** further illustrates this class imbalance. Synthetic Minority Oversampling Technique (SMOTE Technique) was used to address the class imbalance (**Section 5**). The ‘age’ variable had been encoded to values from 1 to 15 inclusive. The remaining six numeric variables consist of values input by individuals based on a scale provided on the survey. The fifteen remaining variables are binary and based on the responses provided by individuals.

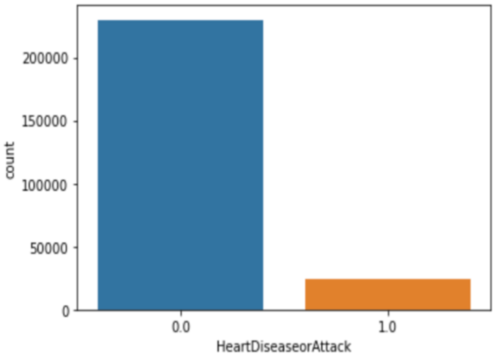


Figure 1. Count plot comparing frequency of each class of the target variable ‘HeartDiseaseorAttack’.

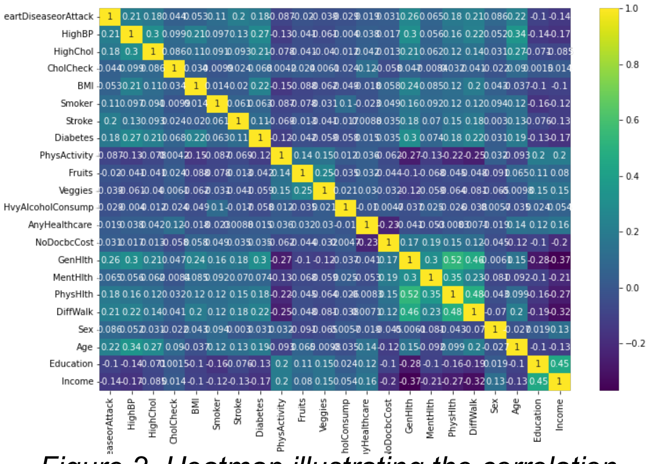
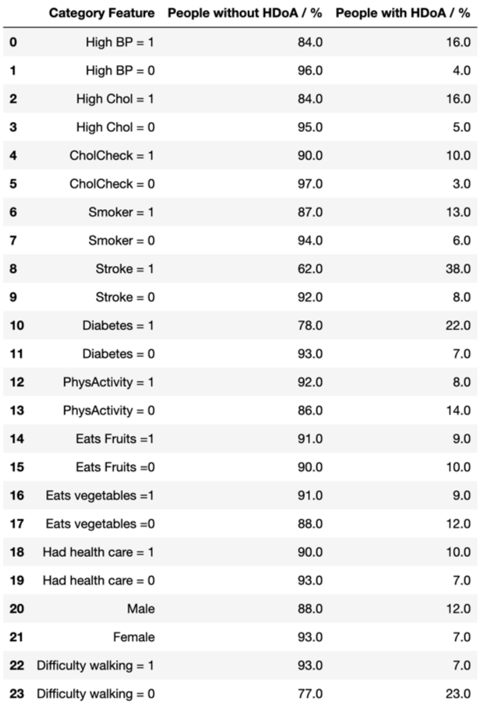
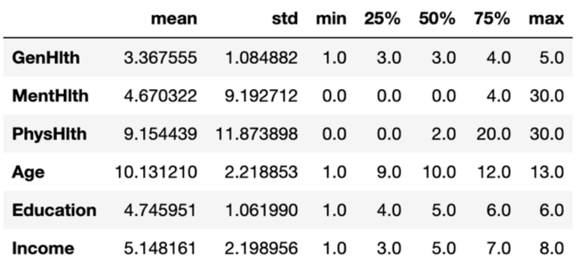


Figure 2. Heatmap illustrating the correlation between all variables.



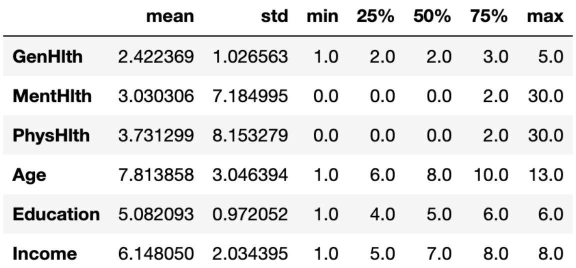
	Category Feature	People without HDoA / %	People with HDoA / %
0	High BP = 1	84.0	16.0
1	High BP = 0	96.0	4.0
2	High Chol = 1	84.0	16.0
3	High Chol = 0	95.0	5.0
4	CholCheck = 1	90.0	10.0
5	CholCheck = 0	97.0	3.0
6	Smoker = 1	87.0	13.0
7	Smoker = 0	94.0	6.0
8	Stroke = 1	62.0	38.0
9	Stroke = 0	92.0	8.0
10	Diabetes = 1	78.0	22.0
11	Diabetes = 0	93.0	7.0
12	PhysActivity = 1	92.0	8.0
13	PhysActivity = 0	86.0	14.0
14	Eats Fruits =1	91.0	9.0
15	Eats Fruits =0	90.0	10.0
16	Eats vegetables =1	91.0	9.0
17	Eats vegetables =0	88.0	12.0
18	Had health care = 1	90.0	10.0
19	Had health care = 0	93.0	7.0
20	Male	88.0	12.0
21	Female	93.0	7.0
22	Difficulty walking = 1	93.0	7.0
23	Difficulty walking = 0	77.0	23.0

Figure 3. % of those afflicted with Heart Disease or Heart Attack (HDoA) which fall into each category.



	mean	std	min	25%	50%	75%	max
GenHlth	3.367555	1.084882	1.0	3.0	3.0	4.0	5.0
MentHlth	4.670322	9.192712	0.0	0.0	0.0	4.0	30.0
PhysHlth	9.154439	11.873898	0.0	0.0	2.0	20.0	30.0
Age	10.131210	2.218853	1.0	9.0	10.0	12.0	13.0
Education	4.745951	1.061990	1.0	4.0	5.0	6.0	6.0
Income	5.148161	2.198956	1.0	3.0	5.0	7.0	8.0

Figure 4a. Statistical summary for patients who had Heart Disease or Heart Attack.



	mean	std	min	25%	50%	75%	max
GenHlth	2.422369	1.026563	1.0	2.0	2.0	3.0	5.0
MentHlth	3.030306	7.184995	0.0	0.0	0.0	2.0	30.0
PhysHlth	3.731299	8.153279	0.0	0.0	0.0	2.0	30.0
Age	7.813858	3.046394	1.0	6.0	8.0	10.0	13.0
Education	5.082093	0.972052	1.0	4.0	5.0	6.0	6.0
Income	6.148050	2.034395	1.0	5.0	7.0	8.0	8.0

Figure 4b. Statistical summary for patients who didn't have Heart Disease or Heart Attack.

Figure 2 illustrates the most correlated independent variables to the target variables are age and general health. Furthermore, there are clear differences in the attributes of those with HDoA and those without. For example, a higher proportion of people with high blood pressure experienced HDoA than those without high blood pressure (**Figure 3**). Furthermore, there is a difference in the continuous attributes between the two classes; people who experienced HDoA had a greater mean Age and lower mean income than those who didn't (**Figure 4a** and **4b**).

3. Model Summaries

3.1 SVM

Support vector machines are used for regression and classification task. Originally, data could only be classified using linear decision boundaries [4]. The introduction of the kernel functions allowed for data to be projected from a low dimensionality to spaces of higher dimensionality [5] where the classes are separated by a hyperplane. Common Applications of SVMs include credit scoring [6] and facial emotion classification [7].

Correct use of SVM parameters C and r can lead the SVMs becoming robust with good generalisation [8]. Furthermore, the convex nature of the loss function often leads to quicker training and hyper-parameter tuning than other models such as MLPs which have many local minima [8, 9].

However, when using larger datasets, SVMs use had a significantly greater computational cost [10]. Due to the large dataset size in this investigation, this greater computational cost limited the use of ensemble methods. Additionally, the performance of SVMs is significantly affected when trained on imbalanced data [10]. This can be overcome using Synthetic Minority Over-sampling Technique [11].

3.2 MLP

Neural networks are models which are loosely associated with biological motivations [12]. The multilayer perceptron consists of layers with a give number of neurones; the input layer, followed by hidden layers and the output layer. The neurones in one layer are linked to the subsequent layer [12]. Each layer may be followed by an activation function and the model adapts through changes to the weights and bias between layers in a direction which decreases the loss function.

MLPs can identify complex non-linear relationships between explanatory and response variables [13] and can work with incomplete data once trained [14].

It is often difficult to explain the behaviour of MLPs which can impact trust in the model [13]. There are few guidelines for determining an appropriate architecture of the network which can lead to extensive trial and error, and consequently longer training times [13]. MLPs can often overfit and require greater computational resources than alternative models such as logistic regression [14].

4. Hypothesis

Despite the use of SMOTE, both models will identify a greater proportion of the majority class correctly as the synthetic data points generated in training rely on the existing data points, which may not capture sufficiently similar characteristics as the data points in the hold out set. The SVM will have a greater F1 score than MLP, as it is being optimised with respect to F1 score in contrast to the MLP which is optimised with respect to binary cross entropy loss. Additionally, the SVM will therefore identify a greater proportion of the minority class as F1 score accounts for class imbalances more than binary cross entropy loss. Final training of the SVM on the Sample 2 training set will take longer as the time taken to train SVMs increases rapidly when trained on larger datasets [10].

5. Train and Evaluation Method

Due to the large dataset size, it was computationally not feasible in the given time frame to perform extensive hyper-parameter optimisation using a large training set. Consequently, the dataset was initially partitioned thrice, into Sample 1, Sample 2 and Test set. All sets initially had the same proportions of classes as the original dataset. Sample 1 was used for extensive hyper-parameter tuning, Sample 2 for slight hyper-parameter adjustments and for the final training of the model, and the Test set was used for final evaluation of the models.

Sample 1 and 2 were split into a training and validation sets. The subsequent training sets had class imbalance; models trained on class imbalanced data can result in significant misclassifications of the minority classed [15]. SMOTE was used to balance the classes as it was notes by Chawla et al. this technique can lead to improved minority class classification [11]. SMOTE was performed on the training set after the partition of each Sample to avoid data leakage. The final versions of each model were retrained on the balanced Sample 2 training set using the best hyper-parameter values found (**Section 6**).

Confusion matrixes were created describing each models' correct and incorrect classifications on the test set. Given the class imbalance in the holdout set, the accuracy metric was a less reliable descriptor of model performance as misclassification in the minority class could be overshadowed. Precision, recall and F1 score are considered for evaluation, allowing for the cost of false negatives, false positives and both combined to be evaluated.

6. Parameters and Experimental results.

For both models, using Sample 1, multiple grid searches were performed. Sample 2 was then used for slight hyper-parameter adjustments and final training.

6.1 MLP

Adam Optimizer was chosen; Liu et al. found Adam to be superior to SGD for binary classification problems [16]. The criterion was binary cross entropy. Early stopping was implemented during hyperparameter tuning. The number of epochs for the final model was 38.

Data Sample Trained and Validated on	No. of Neurones in Hidden Layer 1	No. of Neurones in Hidden Layer 2	Weight Decay	Learning Rate	Non-terminal activation function	Validation Loss
1	8	16	0.001	0.001	Sigmoid	0.44
1	256	512	0.001	0.001	Relu	0.31
2	**256**	**512**	**0.08**	**0.001**	**Relu**	**0.56**

Table 1. Best MLP models trained, final model in bold.

6.2 SVM

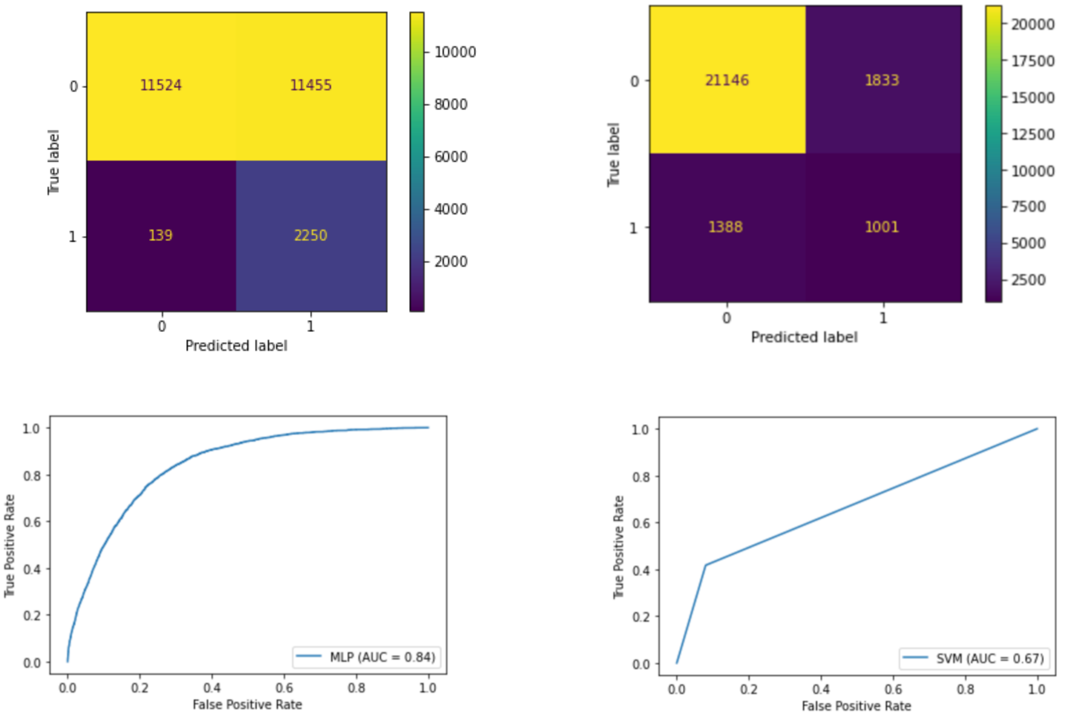
Given SVMs are deterministic and there was a class imbalance in the validation sets, the hyper-parameters were optimised with respect to F1 score.

Data Sample Trained and Validated on	Kernel	C	Gamma	Degree	F1 Score
1	rbf	1	0.01	N/A	0.41
1	rbf	0.1	0.05	N/A	0.41
1	poly	0.1	0.01	3	0.39
2	**rbf**	**1**	**0.01**	**N/A**	**0.38**

Table 2. Best SVM models trained, final model in bold.

SVM ensemble voting classifiers were made from the best ensembles, however, they could not be fitted on the final training set in the given time frame.

6.3 Results



MLP

SVM

Model	Accuracy	Precision	Recall	F1 Score
MLP	0.54	0.16	0.94	0.27
SVM	0.87	0.35	0.31	0.38

Figure 5. Results

7. Analysis and Critical Evaluation

The SVM identified the minority class poorly despite the use of SMOTE in training. It is unlikely the minority class data points differed significantly from those in the hold-out since the MLP had excellent minority class classifications. The MLP had very poor majority class classifications, suggesting it overfitted to the training set, despite the use of early stopping. Furthermore, K-fold cross-validation was not used, which can make potential overfitting harder to identify [17]. Given the poor results, the use k-fold cross validation may have been valuable for reducing overfitting despite the extra computational and time cost associated. Another preventative measure to reduce potential overfitting of the MLP could have been the use of a drop-out layer during training which randomly ignores certain nodes in a layer to increase the noisiness of the training process [19]. The SVM identified a lesser number of the minority class, despite being optimised with the goal of maximising F1 score which is more sensitive to minority class misclassifications than binary cross entropy which does not directly account for class imbalances. An alternative to SMOTE is the RUSBoost algorithm, which involves ensembling models by training each model the full number of instances of the minority class and the same number of instances randomly selected from the majority class, Seiffert et al. found this to be better performing than SMOTE [18] This may improve the minority class recognition of the SVM. Additionally, outliers were not removed in this investigation and SVMs are sensitive to outliers in training, potentially explaining the poor F1 score of the SVM.

The training time was significantly greater for the SVM than the MLP. If the user decides the cost of false negatives is high, the MLP which had a higher recall should be considered However, if both false positives and false negatives are equally costly, the SVM which had a greater F1 score should be used despite the extra training time.

The SVM had a greater execution time on the test set, meaning if the models would need to be used frequently and in short time frames, the MLP would be preferable.

Xiang et al. stated choosing the optimal number of layers in an MLP involved significant trial and error [20], two hidden layers may not have been significant enough to model the relationships between the dependent and independent variables, hence a poor F1 score.

The MLP has a greater AUC than the SVM suggesting the MLP has better discriminatory abilities.

8. Conclusions, lessons learned and future work

The MLP which had a very high recall would be preferable to the SVM if the cost of False Negatives is high. The MLP can be executed more frequently in a given timeframe than the SVM. Outliers should be removed before training the SVM. RUSBoost should be attempted for the SVM to overcome the class imbalance issue. A drop-out layer in the MLP should be implemented to reduce potential overfitting. MLP are more likely to overfit.

References:

[1] 2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm#:~:text=One%20person%20dies%20every%2036,United%20States%20from%20cardiovascular%20disease.&text=About%20659%2C000%20people%20in%20the,1%20in%20every%204%20deaths.&text=Heart%20disease%20costs%20the%20United,year%20from%202016%20to%202017. [Accessed: 07- May- 2022].

[2] F. Hu, "Optimal Diets for Prevention of Coronary Heart Disease", *JAMA*, vol. 288, no. 20, p. 2569, 2002. Available: 10.1001/jama.288.20.2569.

[3] "Heart Disease Health Indicators Dataset", *Kaggle.com*, 2022. [Online]. Available: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset. [Accessed: 07- May- 2022].

[4] S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

[5] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," *2013 International Conference on Advances in Technology and Engineering (ICATE)*, 2013, pp. 1-9, doi: 10.1109/ICAdTE.2013.6524743.

[6] C. Huang, M. Chen and C. Wang, "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, vol. 33, no. 4, pp. 847-856, 2007. Available: 10.1016/j.eswa.2006.07.007.

[7] J. Ghent and J. McDonald, Facial Expression Classification using a OneAgainst-All Support Vector Machine, proceedings of the Irish Machine Vision and Image Processing Conference, Aug 2005.

[8] L. Auria and R. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis", *SSRN Electronic Journal*, 2008. Available: 10.2139/ssrn.1424949.
[9] 2022. [Online]. Available: https://www.researchgate.net/profile/Krzysztof-Siwek/publication/4095905_MLP_and_SVM_networks_-_a_comparative_study/links/0f31753a58e1a37d71000000/MLP-and-SVM-networks-a-comparative-study.pdf. [Accessed: 07- May- 2022].
[10] Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends", *Neurocomputing*, vol. 408, pp. 189-215, 2020. Available: 10.1016/j.neucom.2019.10.118.

[11] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

Available: 10.1613/jair.953.
[12] A. Pinkus, "Approximation theory of the MLP model in neural networks", *Acta Numerica*, vol. 8, pp. 143-195, 1999. Available: 10.1017/s0962492900002919.
[13] 2022. [Online]. Available: https://www.researchgate.net/publication/323665827_Artificial_Neural_Networks_Advantages_and_Disadvantages. [Accessed: 07- May- 2022].
[14] J. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225-1231, 1996. Available: 10.1016/s0895-4356(96)00002-9.
[15] D. Devi, S. Biswas and B. Purkayastha, "Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique", *Connecti on Science*, vol. 31, no. 2, pp. 105-142, 2019. Available: 10.1080/09540091.2018.1560394.
[16] Liu, Z., Shen, Z., Li, S., Helwegen, K., Huang, D. and Cheng, K., 2022. *How Do Adam and Training Strategies Help BNNs Optimization?*. [online] arXiv.org. Available at: <https://arxiv.org/abs/2106.11309> [Accessed 7 May 2022].
[17] 2022. [Online]. Available: https://www.researchgate.net/profile/Daniel-Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf. [Accessed: 07- May- 2022].
[18] Seiffert, C., Khoshgoftaar, T., Van Hulse, J. and Napolitano, A., 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1), pp.185-197.
[19] Brownlee, J., 2022. *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/> [Accessed 7 May 2022].
[20] Cheng Xiang, S. Q. Ding and Tong Heng Lee, "Geometrical interpretation and architecture selection of MLP," in *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 84-96, Jan. 2005, doi: 10.1109/TNN.2004.836197.

Appendix

Glossary

Overfitting: When a model fits to closely to a limited group of datapoints such that it does not generalise well to external data [1].

Hyper-parameter: An adjustable value which influences the learning process of a model.

Gridsearch: a method of hyper-parameter optimisation. A range of values for each hyper-parameter are provided and a model is built based on each combination [2].

Ensemble: A collection of models whose output are considered together. This investigation considered the use of a voting classifier.

Precision: True Positive / (True Positive + False Positive).

Recall: True Positive / (True Positive + False Negative).

F1 Score: Harmonic mean of precision and recall.

Binary Cross Entropy Loss: A loss function which punishes predicted probabilities based on distance from the true class output values [3].

[1] “Overfitting and Optimism in Prediction Models”, Available: https://link.springer.com/chapter/10.1007/978-3-030-16399-0_5 [Accessed: 07-May-2022].

[2] "Grid Searching in Machine Learning: Quick Explanation and Python Implementation", Medium, 2022. [Online]. Available: https://elutins.medium.com/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596. [Accessed: 07- May- 2022].

[3] "Binary Cross Entropy/Log Loss for Binary Classification", Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/. [Accessed: 07- May- 2022].

Implementation

Imbalanced-learn was used to implement SMOTE. Smote works by projected the minority class data points to a feature space and generating new samples along the line segments joining those points [1].

Pytorch was used to implement the neural network, specifically using the ‘nn.Module’ class [2]. Skorch was used to wrap the neural network in, allowing compatibility with scikit-learn. Scikit-learn was used to implement the support vector machine using ‘SVC()’. Scikit-learn was used to optimised hyper-parameters for both models via ‘GridSearchCV’.

The hyper-parameters tuned for the MLP were weight decay, learning rate, non-terminal activation functions, number of neurones in hidden layer 1 and number of neurones in hidden layer 2. Learning rate determines the steps to take each iteration when seeking to minimizing the loss function [3]. Weight decay is a hyper-parameter which can be used for regularisation and prevent overfitting from having too high weights. Early-stopping was used to save time during training however, it also can reduce overfitting and improve generalisation by stopping the network training as validation loss starts to plateau or increase [3].

The SVM hyper-parameters tuned were kernel, C, gamma and degree. It was not possible to fit ensemble SVMs in the timeframe given due to the computational demands of each SVM. All kernels have the hyper-parameter C; a large C seeks to classify all points correctly.

[1] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. Available: 10.1613/jair.953.
[2] "PyTorch", Pytorch.org, 2022. [Online]. Available: https://pytorch.org/. [Accessed: 07- May- 2022].
[3] 2022. [Online]. Available: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/ [4]. [Accessed: 07- May- 2022].