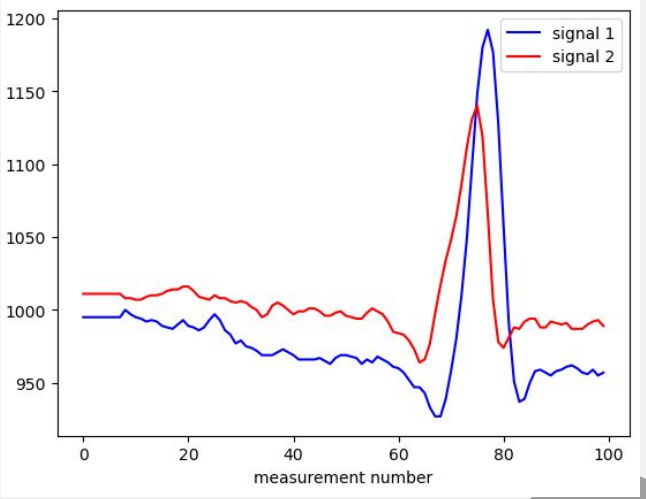
Gateway Data Science Project:

Detecting Abnormal ECG Signals

Anna An, Jon Tchounguen, Victor Wu

The MIT-BIH Arrhythmia Dataset

- Contains 48 30 minute ECG heartbeat samples
  - 100-series collected randomly from Boston hospital
  - 200-series selected to include clinically important rare arrhythmias
  - Beats have been annotated by cardiologists
- 20 specific standard classifications grouped as **5 broad categories:**
  - Normal beats (N)
  - Fusion beats (F)
  - Supraventricular ectopic beats (S)
  - Ventricular ectopic beats (V)
  - Unclassified beats (Q)



About the Data: Record 100

```
[ ] record = wfdb.rdsamp('mitdb/100')
    annotation = wfdb.rdann('mitdb/100', 'atr')
```

```
[ ] record
```

```
(array([[-0.145, -0.065],
        [-0.145, -0.065],
        [-0.145, -0.065],
        ...,
        [-0.675, -0.365],
        [-0.765, -0.335],
        [-1.28 ,  0.   ]]),
 {'fs': 360,
  'sig_len': 650000,
  'n_sig': 2,
  'base_date': None,
  'base_time': None,
  'units': ['mV', 'mV'],
  'sig_name': ['MLII', 'V5'],
  'comments': ['69 M 1085 1629 x1', 'Aldomet, Inderal']})
```

```
[ ] annotation.sample
```

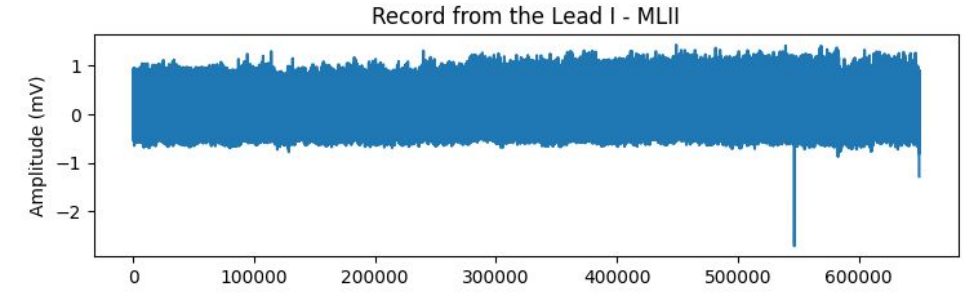
```
array([    18,     77,    370, ..., 649484, 649734, 649991])
```

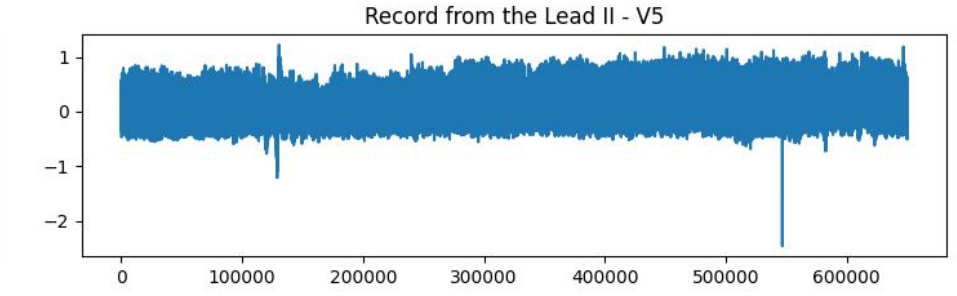
```
[ ] print(annotation.symbol)
```

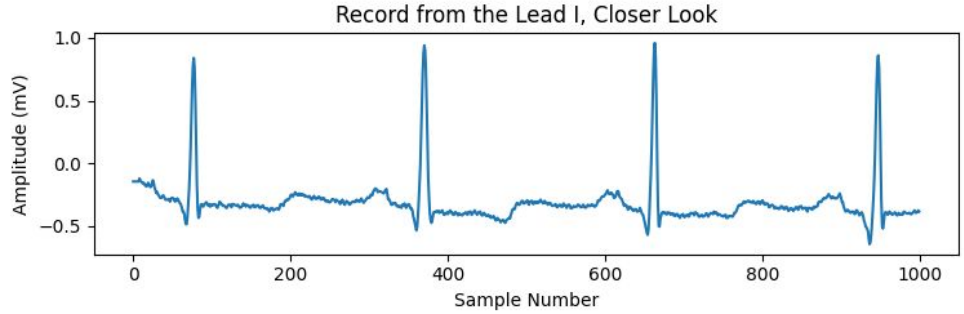
```
['+', 'N', 'N', 'N', 'N', 'N', 'N', 'N', 'A', 'N', 'N', 'N', 'N', 'N',
```

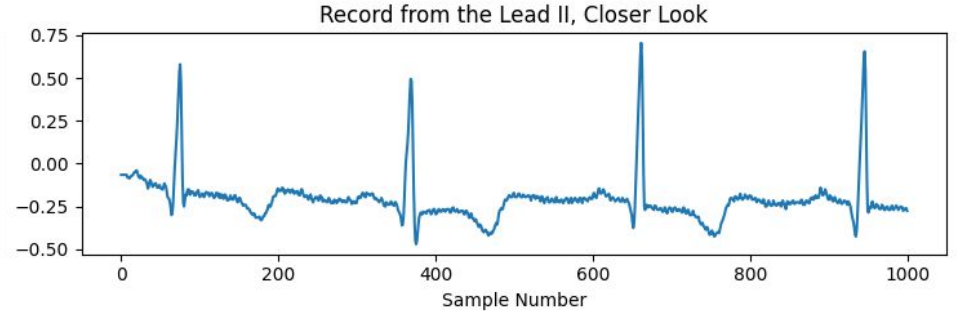
- First, we used a single file for EDA
  - A single file was found to have around 650,000 samples
    - 2274 R peaks
  - The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range

EDA of Record 100









EDA of Record 100

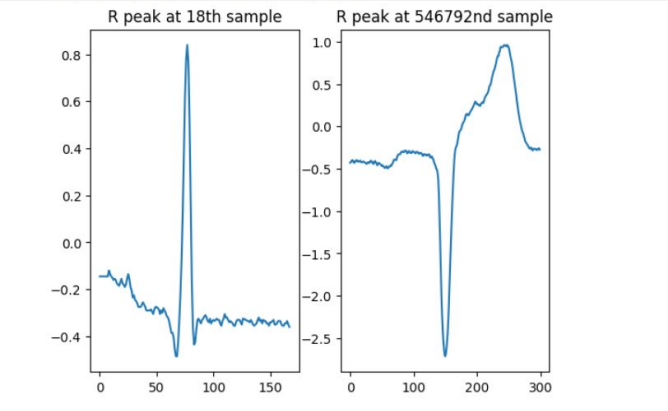
Descriptive analysis
Measures of Central Tendency
Mean = -0.24866670384615372
Median = -0.27

Measures of Dispersion
Minimum = -2.715
Maximum = 1.435
Range = 4.15
Variance = 0.032968090379058436
Standard Deviation = 0.18157117166295544

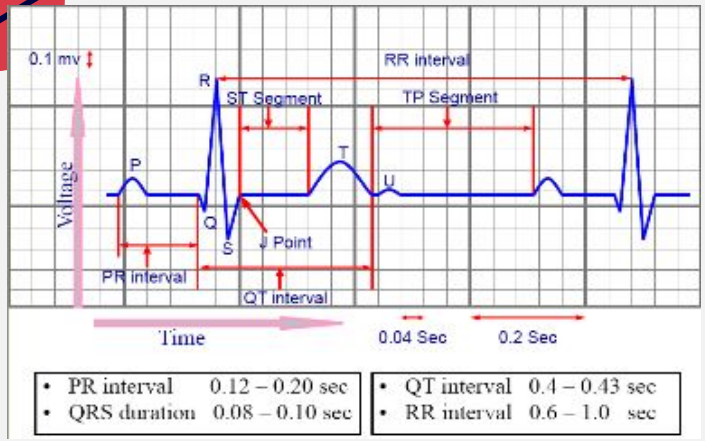
- Are there missing values?

```
Number of NaN values
Lead I: 0
Lead II: 0
annotation.sample: 0
annotation.sample: 0
Lead II: 0
annotation.sample: 0
annotation.sample: 0
```

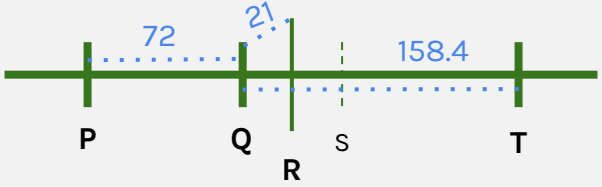
- Outliers?



About the Dataset: Data Wrangling



- Data Segmentation (Darmawahyuni et al., 2022)
  - The recordings were digitized at **360 samples per second** per channel.
  - This implies that for every 36 samples, 0.1 seconds have elapsed.
  - Based on the time elapsed between different parts of an ECG wave, we were able to calculate that a single ECG wave should be 230 samples long (corresponding to 0.63 s)



Step 1. Preprocessing

Data set	N			S						V		F	Q	
	N	L	R	e	j	A	a	J	S	V	E	F	f	Q
DS1[a]	38052	3946	3779	16	16	807	100	32	2	3664	105	414	0	8
	45777			973						3769		414	8	
DS2[b]	36413	4123	3475	0	213	1735	50	51	0	3215	1	388	0	7
	44011			2049						3216		388	7	

Table 1. Size of data sets DS1 and DS2 and the mapping between AAMI and MIT-BIH-AR labels.

- Records from patients 102, 104, 107 and 217 were excluded due to the majority of their beats being **paced beats** *(is controlled by an electrical impulse from an artificial cardiac pacemaker)*
- **Segment**: Data was segmented into individual arrays with 1 beat each
  - If an upside-down peak is observed, we neglected. (Outlier)
  - We observed that each record's baseline differed with each other. Therefore, we normalized the dataset (Kachuee et al., 2018).
- **Classify**: There are 42 unique symbols in annotation.symbol across the whole datasets *(See the appendix 1 for more details)*
  - They were classified based on AAMI labels mapping rules (Zhang et al., 2014)

	0	1	2	3	4	5	6	7	8	9	...	220	221	222	223	224	225	226	227	228	229
0	0.149153	0.152542	0.155932	0.155932	0.152542	0.159322	0.155932	0.166102	0.159322	0.159322	...	0.145763	0.149153	0.155932	0.155932	0.162712	0.145763	0.138983	0.155932	0.152542	0.159322
1	0.143791	0.143791	0.150327	0.156863	0.147059	0.147059	0.140523	0.143791	0.150327	0.156863	...	0.173203	0.166667	0.163399	0.160131	0.169935	0.173203	0.173203	0.163399	0.160131	0.173203
2	0.192691	0.192691	0.192691	0.179402	0.172757	0.192691	0.189369	0.192691	0.189369	0.189369	...	0.232558	0.239203	0.242525	0.245847	0.232558	0.225914	0.229236	0.232558	0.232558	0.239203
3	0.133574	0.148014	0.148014	0.158845	0.158845	0.155235	0.151625	0.158845	0.169675	0.169675	...	0.194946	0.184116	0.176895	0.184116	0.180505	0.194946	0.191336	0.187726	0.191336	0.202166
4	0.150350	0.150350	0.150350	0.143357	0.143357	0.143357	0.153846	0.153846	0.153846	0.139860	...	0.174825	0.178322	0.185315	0.188811	0.188811	0.178322	0.178322	0.181818	0.192308	0.192308
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
100694	0.076555	0.076555	0.078947	0.083732	0.083732	0.081340	0.074163	0.069378	0.074163	0.071770	...	0.064593	0.059809	0.062201	0.052632	0.052632	0.045455	0.050239	0.057416	0.062201	0.066986
100695	0.095694	0.088517	0.090909	0.083732	0.083732	0.083732	0.086124	0.086124	0.088517	0.095694	...	0.059809	0.062201	0.059809	0.059809	0.064593	0.066986	0.071770	0.071770	0.071770	0.059809
100696	0.096447	0.088832	0.081218	0.078680	0.081218	0.081218	0.083756	0.083756	0.081218	0.091371	...	0.078680	0.081218	0.083756	0.088832	0.088832	0.093909	0.096447	0.096447	0.088832	0.096447
100697	0.054455	0.061881	0.056931	0.056931	0.061881	0.071782	0.069307	0.066832	0.059406	0.056931	...	0.071782	0.079208	0.076733	0.079208	0.079208	0.076733	0.071782	0.066832	0.061881	0.064356
100698	0.121655	0.121655	0.116788	0.114355	0.111922	0.114355	0.116788	0.116788	0.119221	0.119221	...	0.048662	0.048662	0.051095	0.055961	0.060827	0.063260	0.065693	0.063260	0.058394	0.063260

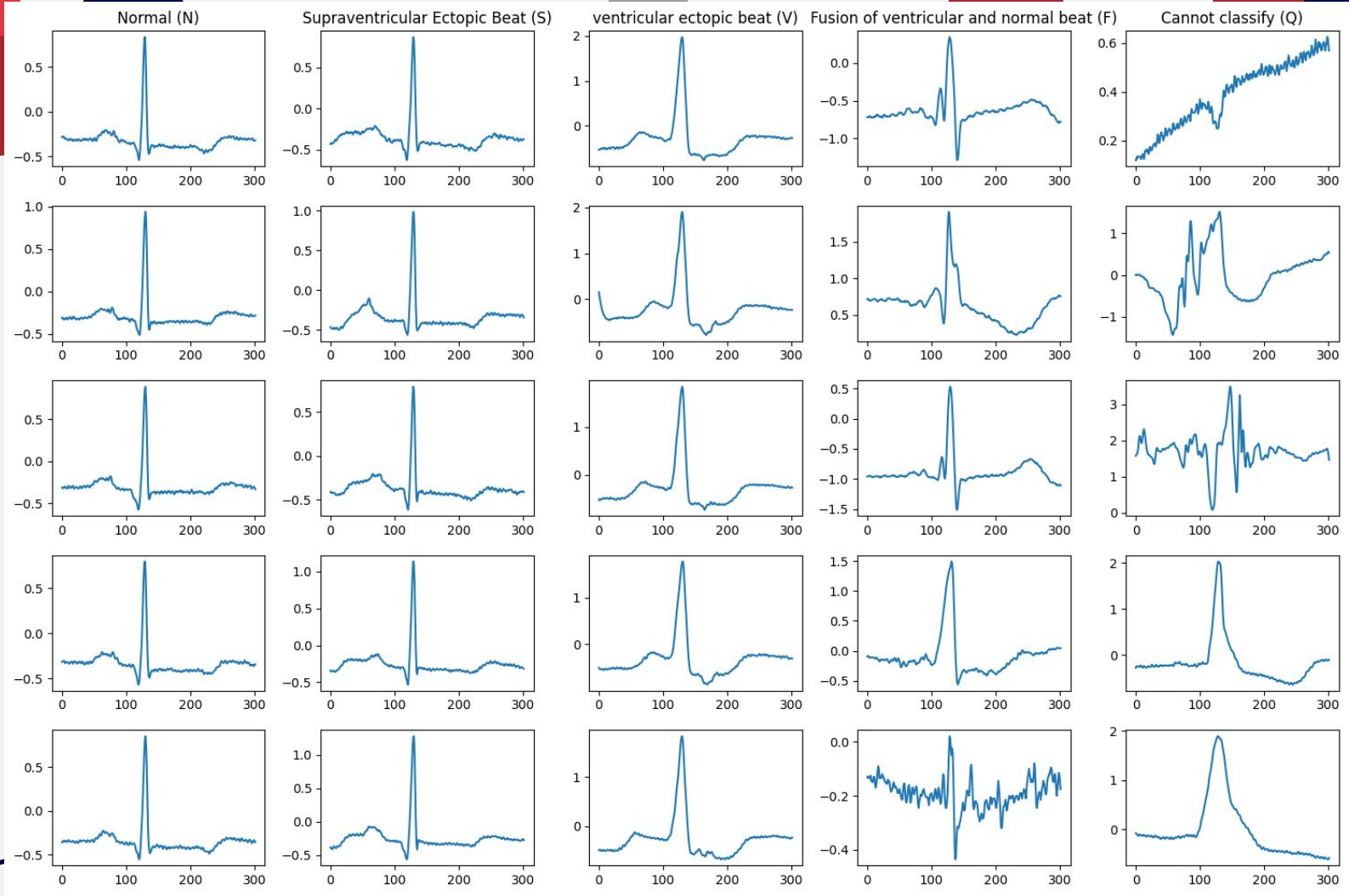
Variable 1. 230*100698 Table Formed; Columns represent a sample (230 samples = 0.63 seconds), rows represent one wave (N, S, V, F, or Q), and values represent an amplitude for a wave at a specific sample (mV).

```
['N' 'N' 'N' ... 'N' 'N' 'N']
(100699,)
```

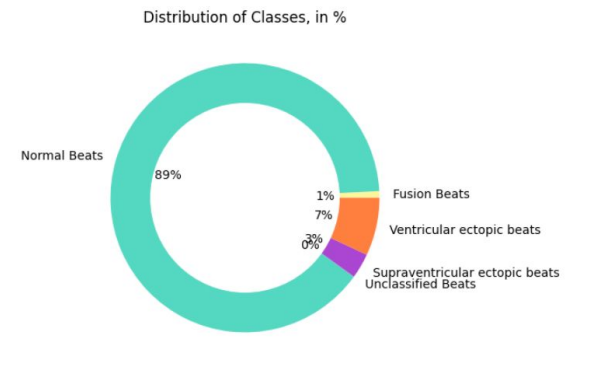
Variable 2. Wave label

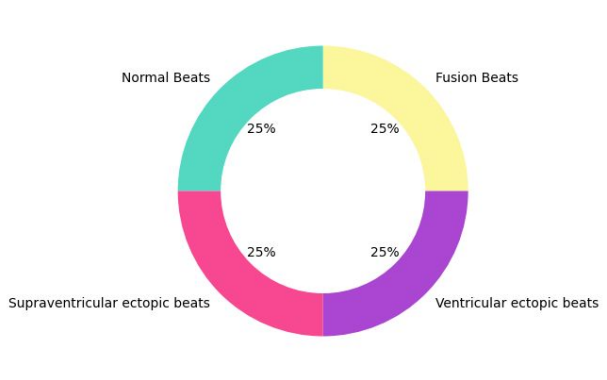
```
[3.70000e+02 6.62000e+02 9.46000e+02 ... 6.49292e+05 6.49536e+05
 6.49772e+05]
(100699,)
```

Variable 3. Wave's R peak location


Normal (N)
Supraventricular Ectopic Beat (S)
ventricular ectopic beat (V)
Fusion of ventricular and normal beat (F)
Cannot classify (Q)

Step 2. Balancing Dataset



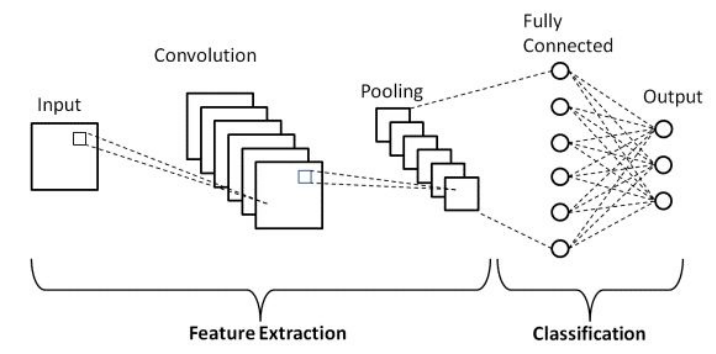


- Dataset is imbalanced across 5 classes
  - Normal: 89848
  - Supraventricular: 3026
  - Ventricular: 7008
  - Fusion: 802
  - Unclassified: 15
- Smaller classes were resampled so that all classes have same number of samples in the training set
  - *Unclassified classes were neglected because 1. No specific pattern observed, 2. small size of sample*
- All 4 classes now have 31727 samples
- Data was split into 2 sets: 70% training (56% train, 14% validate), 30% test sets.

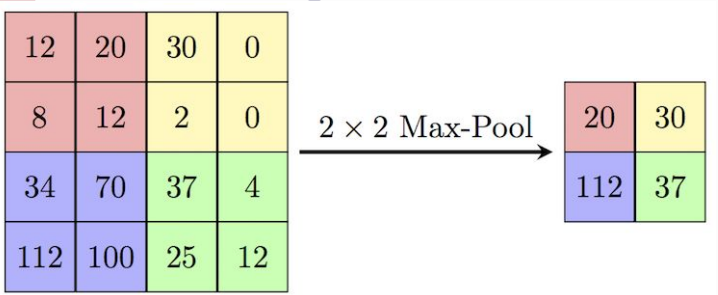
The Task

- The goal is to classify ECG beats with one of four labels
  - Normal beats, fusion beats, supraventricular ectopic beats, ventricular ectopic beats
  - As there is a large variation in unclassified beats, we decided to exclude them from the classification task

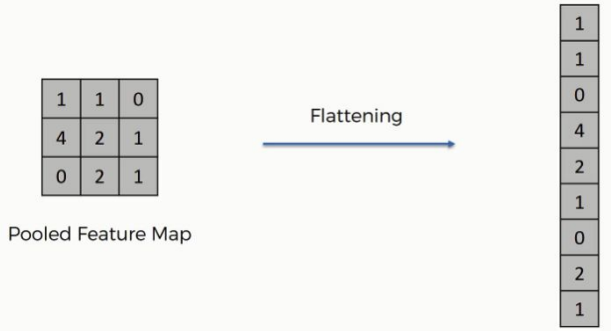
Step 3. Classification: CNN



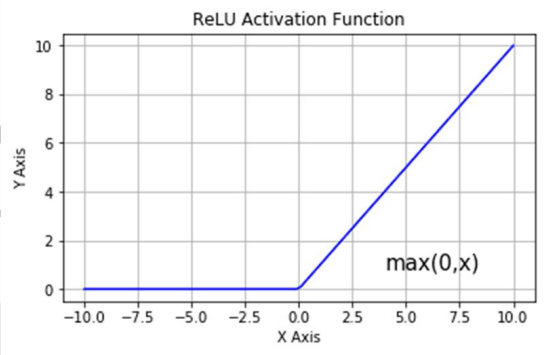
- We used two methods to classify the data, convolutional neural network (CNN) and support vector machine (SVM)
- Convolutional neural network:
  - A classification algorithm with layers made of functions that take multiple inputs and produce a single output
  - Each function has different weights that can extract features that distinguish different classes
  - Here, we applied a 1-dimensional convolution operation to the input data by sliding a small window (known as a filter) over the input data and performing a dot product between the filter and the data within the window.



Max Pooling: a downsampling technique that reduces the spatial dimensionality of the input data. This reduces the size of the data, while preserving the most important features.

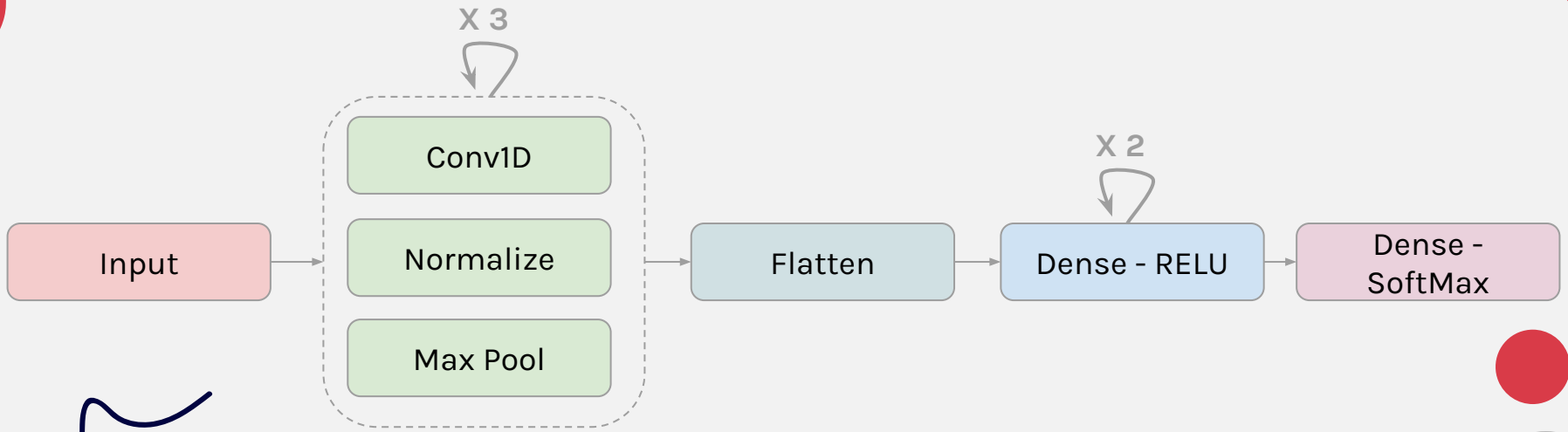


Flatten: Flatten is a layer that simply reshapes the input data into a 1-dimensional vector.



ReLu: allows the decision boundary to be non-linear. (It applies the function f(x) = max(0,x))
Sigmoid: It applies the function f(x) = 1 / (1 + exp(-x)) to each element of the input tensor, squashing the output values between 0 and 1.
SoftMax: Exponential functions are used to make sure outputs sum to 1. (Applies f(x) = exp(x_i) / sum(exp(x_j)))

CNN1

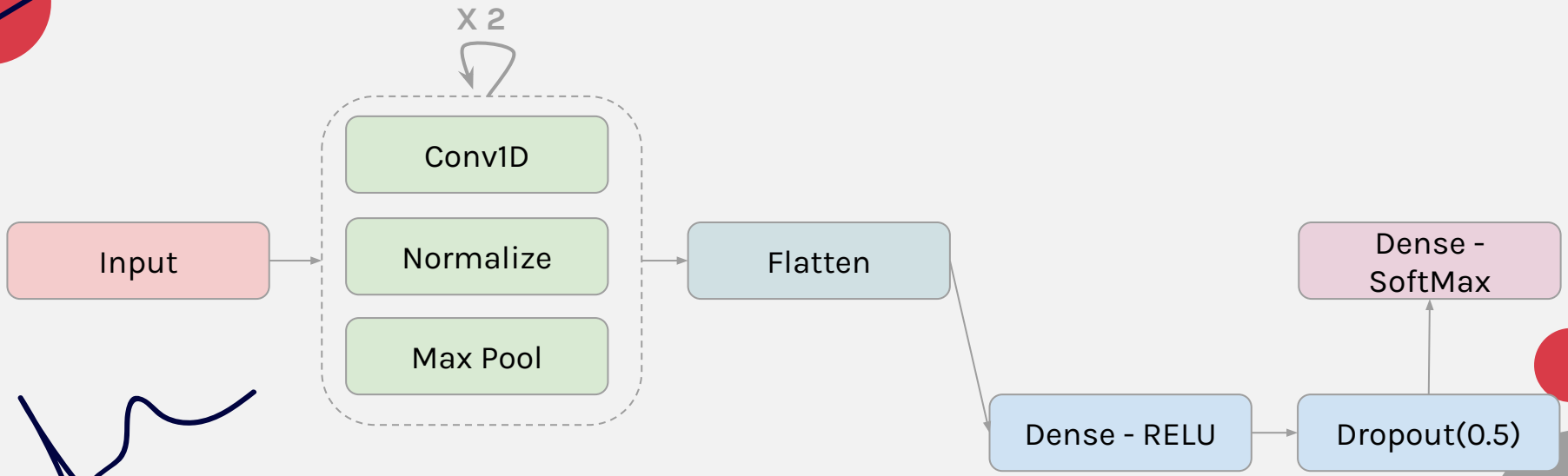


Conv1D
* 64 kernels of size 6,
* ReLU activation

Max Pool
*Pool Size of 3

Multiple Dense Layers:
helps in changing the dimensionality of the output from the preceding layer so that the model can easily define the relationship between the values of the data in which the model is working

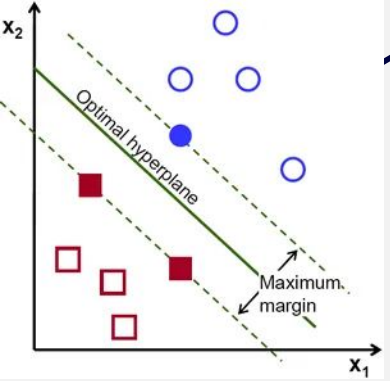
CNN2



Conv1D
* 32 kernels of size 3,
* Sigmoid activation

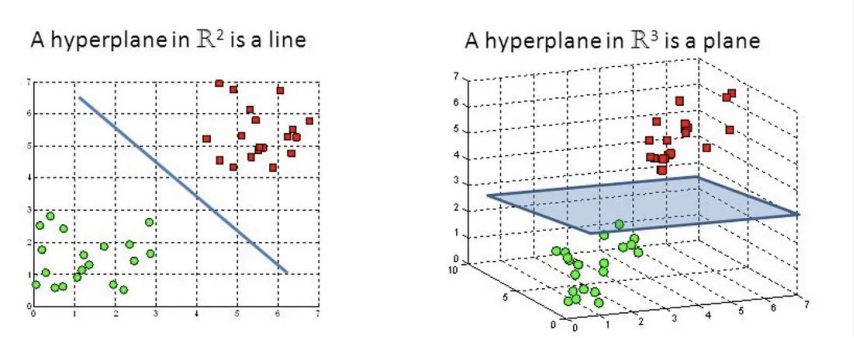
Max Pool
*Pool Size of 2

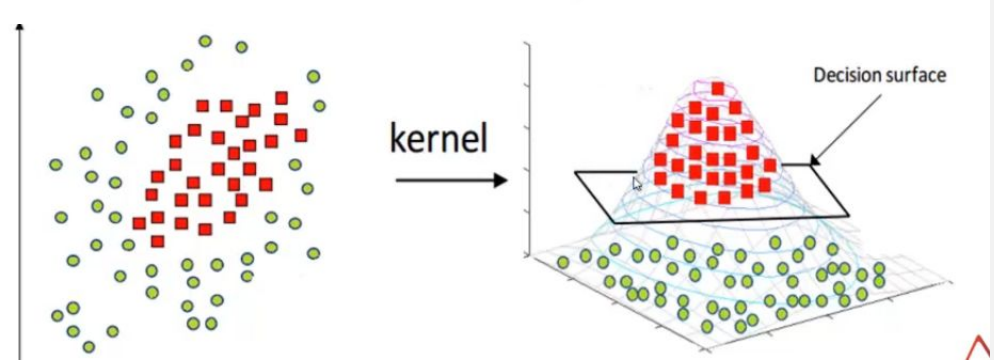
Drop Out:
a mask that nullifies the contribution of some neurons towards the next layer and leaves unmodified all others. It prevents overfitting.



Step 3. Classification: SVM

- SVM: support vector machine
  - An algorithm that finds a linear boundary between two classes
    - In 2D a line, in 3D a plane, in nD a n-1 D hyperplane (hard to imagine)
  - The boundary location is found by maximizing the distance between the points close to the boundary and the boundary
  - A kernel function is used to transform data that is not linearly separable to a higher dimension where it is
  - We used an rbf kernel (projects into a Gaussian distribution)





Findings: Criteria

- Four metrics: accuracy, precision, sensitivity, specificity
  - **Accuracy**: the percentage of correct classifications
  - **Precision**: out of all the classifications of a certain class, what proportion were correct
  - **Sensitivity (recall), specificity:** true positive and true negative rate
  - **Confusion matrix**: shows the number of each class correctly and incorrectly classified

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Findings: Accuracy

Both our CNN and SVM model scored excellently, with testing and training scores ≥ 0.95.

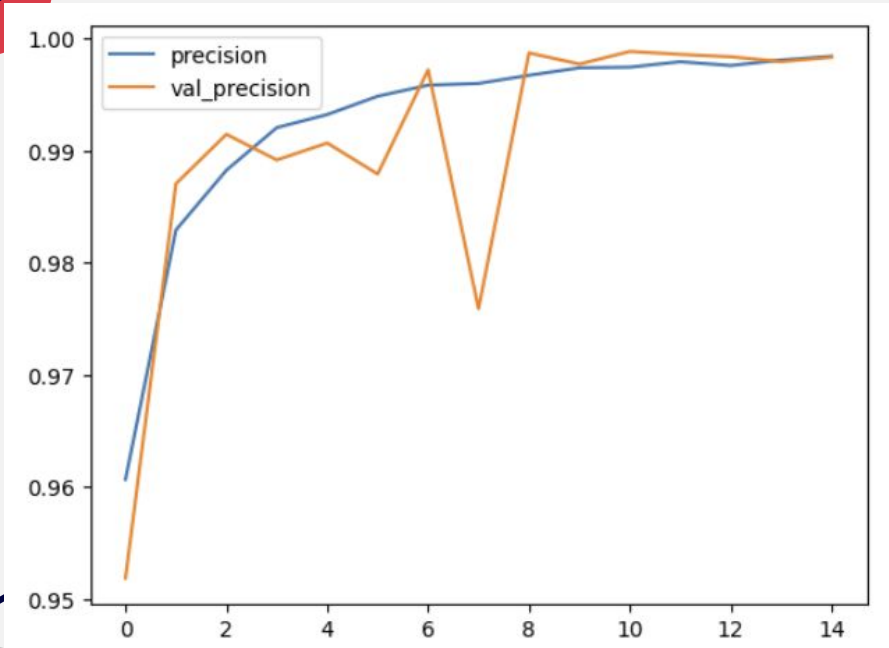
- **CNN1** Overall accuracy scores: 0.9987
- CNN1 Accuracy scores by classes:
  - N: 0.991
  - S: 0.823
  - V: 0.968
  - F: 0.798

- **CNN2** Overall accuracy scores: 0.9886
- CNN2 Accuracy scores by classes:
  - N: 0.991
  - S: 0.809
  - V: 0.960
  - F: 0.855

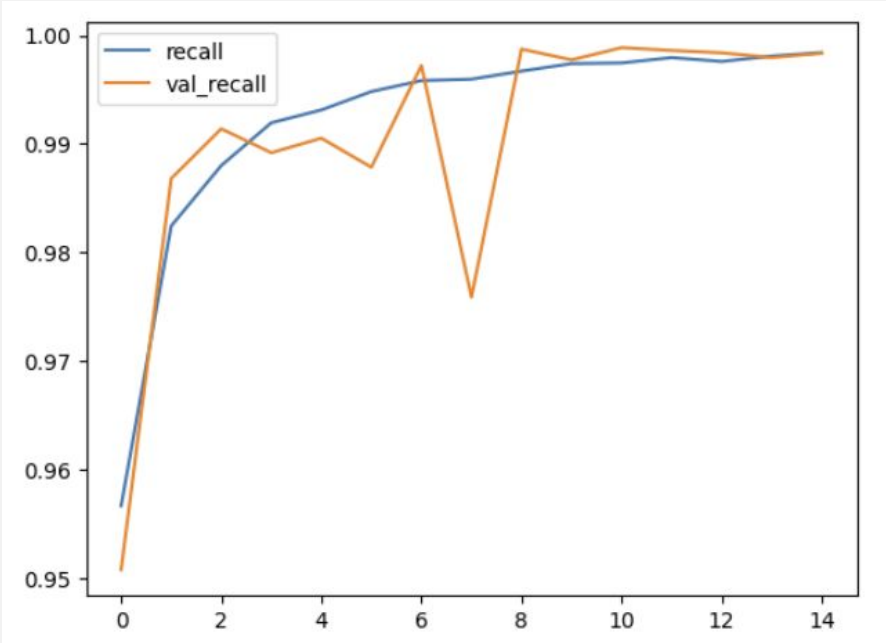
- **SVM** Overall accuracy scores: 0.9485
- SVM Accuracy scores by classes:
  - N: 0.950
  - S: 0.904
  - V: 0.954
  - F: 0.957

The CNN scored slightly better than the SVM, and both had high testing scores, meaning the models generalize well to new data.

Precision, Recall- CNN1

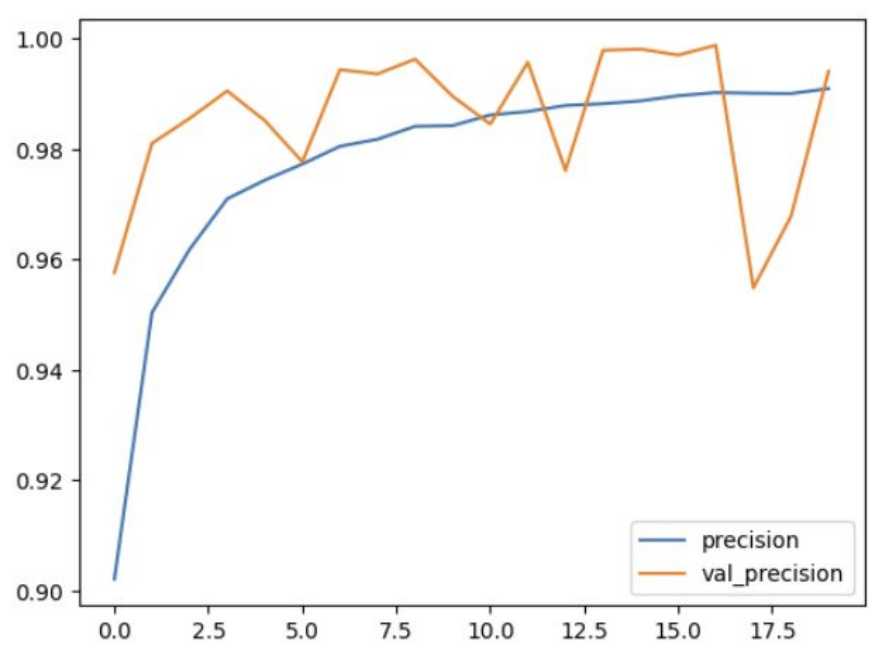


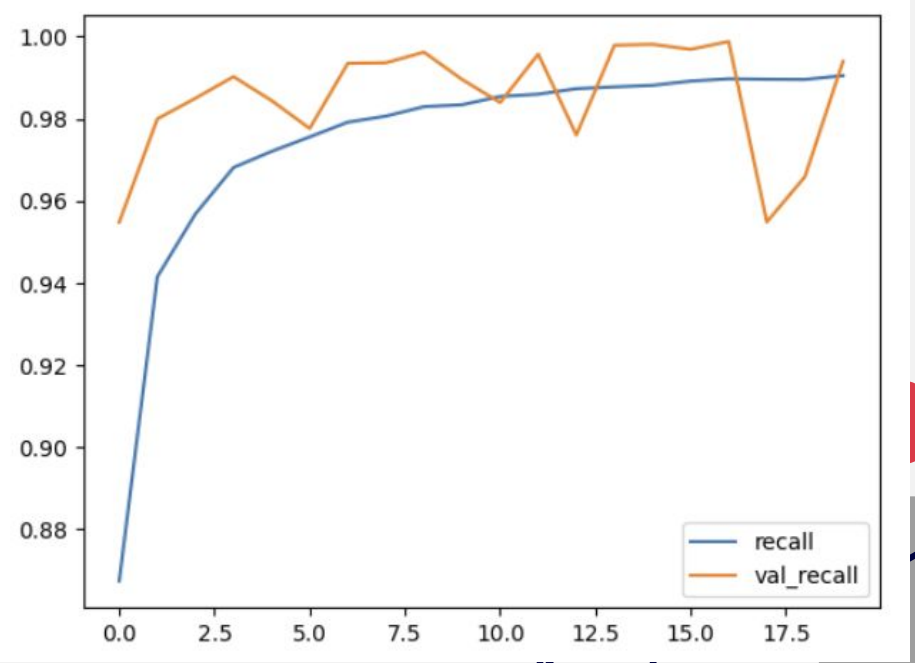
Precision Graph: 0.9987 at Epoch 15



Recall Graph: 0.9987 at Epoch 15

Precision, Recall- CNN2





Precision Graph: 0.9890 at Epoch 20

Recall Graph: 0.9884 at Epoch 20

Precision, Recall — SVM

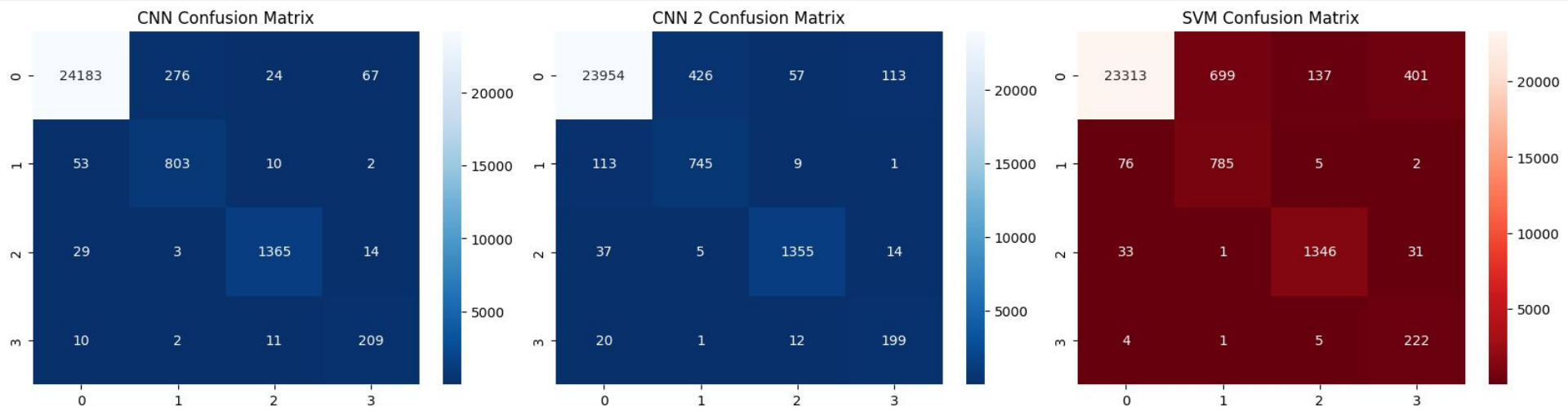
- Normal beats (N)
- Supraventricular ectopic beats (S)
- Ventricular ectopic beats (V)
- Fusion beats (F)

	precision	recall	f1-score
0	1.00	0.95	0.97
1	0.53	0.90	0.67
2	0.90	0.95	0.93
3	0.34	0.96	0.50
accuracy			0.95
macro avg	0.69	0.94	0.77
weighted avg	0.97	0.95	0.96

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Confusion Matrix

- Shows for each class the amount of correct and incorrect classifications
- x axis: actual class, y axis: predicted class



Sensitivity/Specificity

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

- **Sensitivity (recall)**: Sensitivity or recall refers to a test's ability to designate an individual with disease as positive. *(Known as True Positive Rate)*
- **Specificity**: The specificity of a test is its ability to designate an individual who does not have a disease as negative. *(Known as True Negative Rate)*

Findings: Sensitivity/Specificity

CNN1

class	sensitivity	specificity
0	0.953007	0.989124
1	0.991181	0.900922
2	0.998441	0.970234
3	0.998845	0.892241
Avg:	**0.985**	**0.938**

CNN2

class	sensitivity	specificity
0	0.952210	0.984114
1	0.989310	0.898618
2	0.997778	0.964564
3	0.996683	0.922414
Avg:	**0.984**	**0.942**

SVM

class	sensitivity	specificity
0	0.954998	0.949613
1	0.973237	0.904378
2	0.994269	0.953933
3	0.983823	0.956897
Avg:	**0.976**	**0.941**

Summary

	Accuracy	Precision	Recall	Specificity	Time
CNN1	0.9987	0.9984	0.9984	0.938	10 mins
CNN2	0.9818	0.9819	0.9818	0.942	9 mins
SVM	0.9485	0.97	0.95	0.941	29 mins

Pros / Cons

CNN

- Takes less time to train
- Slightly more accurate
- Is more precise across all classes
- Is less accurate for some classes
- Can create non-linear decision boundaries

SVM

- Takes longer to train
- Still very accurate (0.95)
- Is not precise across all classes
- Accuracy is consistent across all classes
- Can only transformed using the kernel function

Conclusions / Recommendations

- After using both CNN and SVM models to classify arrhythmia data, we conclude that the CNN is a more effective classification model than the SVM. This is because the CNN takes less time to run than the SVM while also having scoring higher in measures of model accuracy.
- Therefore, for classification of different types of heartbeats, we would recommend using a convolutional neural network.

References

Darmawahyuni, A., Nurmaini, S., Rachmatullah, M. N., Tutuko, B., Sapitri, A. I., Firdaus, F., Fansyuri, A., & Predyansyah, A. (2022). Deep learning-based electrocardiogram rhythm and beat features for heart abnormality classification. *PeerJ Computer Science*, *8*, e825. https://doi.org/10.7717/peerj-cs.825

Kachuee, M., Fazeli, S., & Sarrafzadeh, M. (2018). ECG Heartbeat Classification: A Deep Transferable Representation. *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 443–444. https://doi.org/10.1109/ICHI.2018.00092

Moody, G., & Mark, R. (2001). The impact of the MIT-BiH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45–50. https://doi.org/10.1109/51.932724

Santosa, F., Ardekani, S. (2023). *Gateway Data Science Lectures* [PowerPoint slides]. Johns Hopkins University Gateway Data Science Canvas: https://canvas.jhu.edu/

Saxena, S. (2021, March 12). Beginner's Guide to support vector machine (svm). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/beginners-guide-to-support-vector-machine-svm/

Shung, K. P. (2018, March 15). Accuracy, precision, recall or F1?. Medium. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Zhang, Z., Dong, J., Luo, X., Choi, K.-S., & Wu, X. (2014). Heartbeat classification using disease-specific feature selection. *Computers in Biology and Medicine*, *46*, 79–89. https://doi.org/10.1016/j.compbiomed.2013.11.019

Appendix

	label_store	symbol	description
0	0		Not an actual annotation
1	1	N	Normal beat
2	2	L	Left bundle branch block beat
3	3	R	Right bundle branch block beat
4	4	a	Aberrated atrial premature beat
5	5	V	Premature ventricular contraction
6	6	F	Fusion of ventricular and normal beat
7	7	J	Nodal (junctional) premature beat
8	8	A	Atrial premature contraction
9	9	S	Premature or ectopic supraventricular beat
10	10	E	Ventricular escape beat
11	11	j	Nodal (junctional) escape beat
12	12	/	Paced beat
13	13	Q	Unclassifiable beat
14	14	~	Signal quality change
16	16	|	Isolated QRS-like artifact
18	18	s	ST change
19	19	T	T-wave change
20	20	*	Systole
21	21	D	Diastole
22	22	"	Comment annotation
23	23	=	Measurement annotation
24	24	p	P-wave peak
25	25	B	Left or right bundle branch block
26	26	^	Non-conducted pacer spike
27	27	t	T-wave peak
28	28	+	Rhythm change
29	29	u	U-wave peak
30	30	?	Learning
31	31	!	Ventricular flutter wave
32	32	[	Start of ventricular flutter/fibrillation
33	33	]	End of ventricular flutter/fibrillation
34	34	e	Atrial escape beat
35	35	n	Supraventricular escape beat
36	36	@	Link to external data (aux_note contains URL)
37	37	x	Non-conducted P-wave (blocked APB)
38	38	f	Fusion of paced and normal beat
39	39	(	Waveform onset
40	40	)	Waveform end
41	41	r	R-on-T premature ventricular contraction

Appendix 1. all the labels present in the dataset