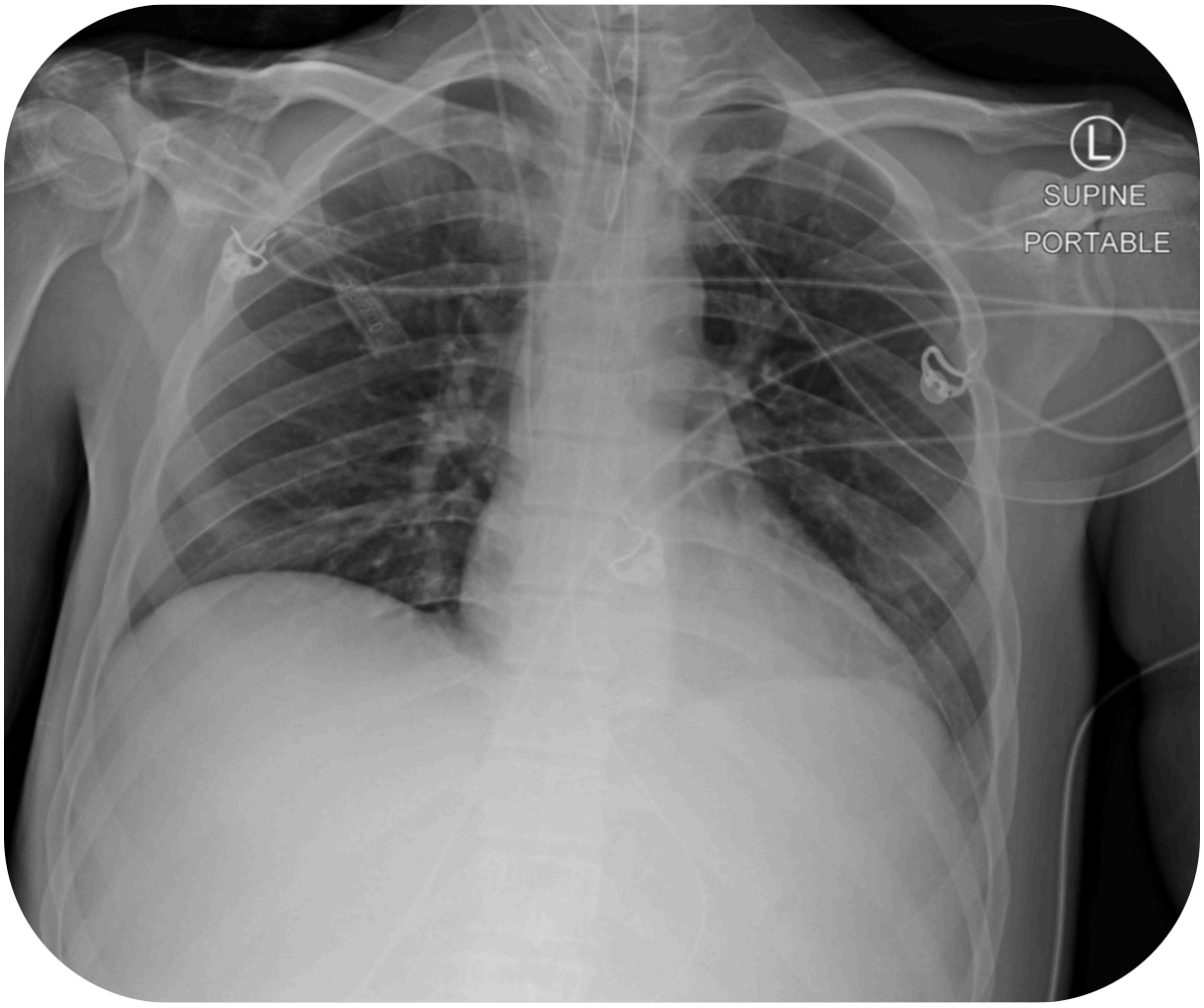


CATHETER PLACEMENT ASSESSMENT

Medical Image Classification

Adam Choong, Louise Childs, Isaac Wood, Fengze Yang

October 2023

Executive Summary

With a vast array of tasks to handle in a chaotic environment such as a hospital, medical professionals are required to use their time and resources effectively on patients that require urgent life saving intervention. In a system which is already overwhelmed with the volume of patient queries received, automation of diagnostic procedures can reliably and safely reduce the workload of stressful doctors and nurses. For hospitals with limited funding, the automation process can help optimise the use of their resources and for patients, results can be delivered reliably and quickly, which may lead to increased satisfaction.

In the creation of the automation process, image detection models were implemented to varying degrees of success. The success of each of these models was primarily measured by their ability to predict correct diagnoses. However, a range of other measures of success were taken into consideration. Each model's ability to detect false negatives, for example, was an especially vital qualifier of their feasibility in a hospital setting because false negatives are often very dangerous for patients. In this case, if a catheter's placement posed a serious threat to a patient's life, but was not flagged (false negative diagnosis), then the model would be liable for the potential death of said patient. Other metrics such as the model's ability to confidently predict outcomes correctly were measured to quantify the efficiency of the automation process and reduce patient anxiety. For hospitals, effective models are able to strike a "balance between recall and precision" (OpenAI, 2023) to manage the need for patient satisfaction, efficient resource management and safety. Furthermore, these measures will be used to imply whether patients' trust in the automation process will be sufficient for use in a practical setting.

Ethical considerations will have to be made regarding the need for patients to acknowledge and consent to the risks associated with the use of machine learning enabled processes in situations that can be life threatening. Additionally, image detection models must be carefully monitored during the process of development to ensure transparency of unexpected behaviours that can be dealt with promptly. To avoid further mistrust of AI enabled services among the public, these models must be carefully scrutinised before being tested in a focus group. Given the severity of the consequences associated with a model's failure to perform, extremely high standards must be set for a model to qualify to the next stage in implementation, which is testing on a focus group of patients in real time under a controlled setting. Furthermore, due to the high priority that doctors and nurses have, of maintaining patient safety, even after successful development and implementation, model activity will continuously have to be monitored for increased rates of misdiagnosis to mitigate the risk of severe outcomes for patients. Finally, as patient confidentiality is paramount in the medical field, the data provided was anonymised to the best extent possible by replacing patient information with unique arbitrary IDs to differentiate between them.

The models developed were utilised to handle the binary problem of differentiating between normally and problematically placed catheters as well as the multi-class classification problem of determining whether a catheter placement was Normal, Abnormal or Borderline. They performed reasonably well, with RNN achieving better results than kMeans. In particular, moderately high recall, meaning a low amount of false negatives, was achieved. This is significant as the models helped minimise human error. Moderate levels of precision was also achieved, thereby demonstrating that the models did not decrease the efficiency of hospitals in processing the needs of patients.

With the findings outlined in this report, beyond the assessment of the automation process implemented, burnout among healthcare workers can be prevented, given that the performance of the models created meets the standards of care and safety required for hospitals to operate effectively. Reducing burnout among healthcare workers is essential because it allows them to work efficiently

and effectively with a reduced likelihood of making mistakes in other procedures that may not yet have sufficient resources to be automated or tasks that cannot be performed by AI enabled services because of a lack of advancement.

- Adam Choong and Louise Childs October 2023

Introduction

The purpose of this project is to automate the assessment of catheter placement in patients. The improper use and incorrect placement of patient catheters pose a significant risk of serious complications. In the medical field, particularly, the risks of these complications are elevated when issues arise but are not diagnosed due to the occurrence of false negative results or type II errors during the prediction of catheter placement. Thus, while predicting catheter placement correctly is important in the process of automation, detecting these type II errors is also just as vital. Consequently, this project prioritises the assessment of two metrics: recall and accuracy. Precision will also be considered as a measure of how well the automation process can save resources and increase the efficiency of diagnostic procedures.

Three catheters' (the nasogastric tube (NGT), central venous catheter (CVC) and endotracheal tube (ETT)) placements will be assessed for multiple lung X rays and classified based on the quality of their placements. A CVC is connected from the heart to bloodstream and "helps patients receive drugs, fluids or blood for emergency or long-term treatment" (Cleveland Clinic, 2022). Having a mispositioned CVC also has "immediate complications includ[ing] cardiac arrhythmia [and] delayed complications includ[ing] infections" (Mon, 2016). Cardiac arrhythmia describes the "problem [that occurs] with the rate or rhythm of a patient's heartbeat" (NHLBI, NIH., 2022) and can lead the patient to have a heart attack. An ETT is placed through the trachea which is located near the throat and ensures patients with breathing difficulties can access oxygen through their throat. A misplaced ETT may result in a range of complications such as "hypoxemia, difficult or inadequate ventilation [and] atelectasis" (Miller, 2016). Atelectasis is the partial or complete collapsing of a lung while hypoxemia occurs when there is a lack of oxygen in the bloodstream and can greatly reduce brain and heart functionality. An NGT is placed through "the nose or mouth and slid into the stomach" (Nasogastric Tubes (Insertion and Feeding), n.d.) to deliver nutrients to patients who often have difficulties ingesting food due to throat and tracheal complications. Complications that can arise from having a misplaced NGT include "severe aspiration pneumonia" (Felipe, 2012) which is a life-threatening lung infection "caused by inhaling saliva, food, liquid, vomit" (Cleveland Clinic, 2021). It is placed in large central vein. (in the right side of the heart and the arteries leading to the lungs).

Another catheter which will not be assessed in terms of placement quality is the Swans-Ganz catheter which "monitors the heart's function and blood flow and pressures in and around the heart" (Swan-Ganz - Right Heart Catheterization, n.d.). It is placed in large central vein. (in the right side of the heart and the arteries leading to the lungs). The placement of this catheter will be dismissed in the analysis because it has only binary labels and not enough information on the quality of its placement.

The significance of investigating methods of automating catheter placement assessments relates to the need for hospitals, which are already under heavy time and resource constraints, to increase the efficiency at which patients are handled to alleviate the issues that arise from having an overwhelmed healthcare system. In a similar respect, the automation process aims to reduce the labour intensive procedures that doctors and nurses have to perform in diagnosing issues with catheter placements.

However, before implementing automated processes to assess catheter placements, patient safety must be prioritised to avoid the serious complications associated with misdiagnosing harmful

placements as malign. Yet, patients cannot be unnecessarily flagged for further inspection by doctors otherwise the model risks reducing the efficiency of an already fragile healthcare system. In addition to reducing efficiency, a model that mistakenly raises false alarms for patients may induce anxiety thereby worsening overall patient experience. Consequently, patients who have been subjected to an automation process that consistently misdirects healthcare staff, may be less inclined to visit the hospital and have increased mistrust of artificial intelligence enabled services.

On the other hand, an automated process which correctly identifies instances of misplacement can help identify errors that are missed by healthcare professionals, thereby maintaining the standards of vigilance required in patient care.

Some issues that could arise pertain to the lack of data surrounding malpositioned catheters, which is to be expected since hospitals have an obligation to prioritise patient safety and minimise to the best of its ability, instances of negligence. This issue could create potential biases in the development of models if it is not handled properly. To handle this issue, data balancing techniques need to be employed.

Upon the completion of our analysis of the X-rays, we expect to see a high accuracy and recall being achieved by our models but the results they predict must be meaningful. A common example of an instance whereby high accuracy may be achieved but the predictions of a model may not be entirely useful is if almost all predictions are the same. To clarify further, if an overwhelming majority of images were classified as belonging into the most prevalent class, this would defeat the purpose of the project to differentiate images based on catheter placement. Furthermore, we intend to achieve high recall to avoid instances of misdiagnosis because of the dangerous implications that unattended issues regarding placement can have for patients. A secondary goal of this project is to have the automation process be reliably efficient, meaning that ideally, models should be able to identify normal catheter placements reliably so that resources that are used to inspect catheter placement can be saved for those who are in urgent need of catheter replacement.

KMeans was selected as a model to explore due to its ability to ideally cluster the image data and find distinctive differences between catheter types. Additionally, it is relatively simple to implement. We can assess how accurately it predicts a given catheter type based upon the cluster that it assigns it to and whether it is associated with the matching catheter types.

For the deep learning component of the analysis, there will be a couple of main approaches being taken. The first approach taken will be based on coordinate mapping and sequencing. Through this approach, we intend to map the placement of a catheter in the form of a trail to compare with other graphed trails of similarly placed catheters to realise a trend and assess various metrics on that trend. The other approach will incorporate the use of images and pixel intensity statistics to highlight regions of interest that can be used to differentiate catheter placements. Given that the model interpretability of multilayered neural networks is ex, the primary assessors used from the results will be accuracy and recall and precision. The methods used to model our results range from being efficient to being computationally expensive. An important caveat in our findings regarding neural networks is that they will be stochastic and this inherent randomness can only guarantee, to some level of confidence, that our numerical results lie in an interval.

Data Quality and Exploratory Data Analysis

In the data set provided, 3255 patients' catheter placements were monitored through 300083 different images of chest X rays. Accompanying 9095 of these images were coordinates annotating positions of different catheters' placements along with their labels indicating the quality of their positioning.

Furthermore, patients' X rays were monitored for the presence of a Swans Ganz catheter. A summary of the relevant statistics for the annotated images is provided in figure_.

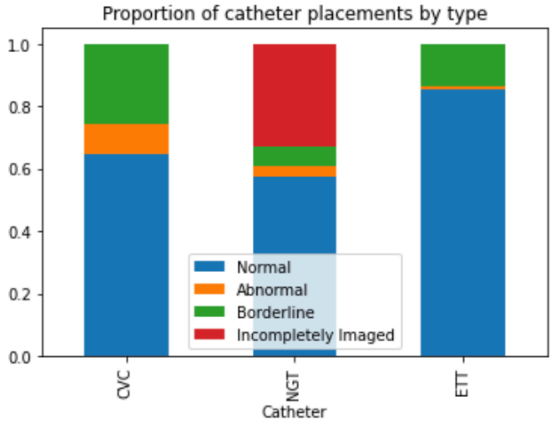
Figure 1: a brief overview of the quantities of different images available for analysis.

	CVC	NGT	ETT
Normal	7437	1870	2536
Abnormal	1206	111	30
Borderline	2986	219	428

In addition to the tabulated summary of relevant statistics, there were 1019 incompletely imaged X rays and 157 images containing a Swans Ganz catheter.

Only a small subset of images were chosen to train our models for practical reasons pertaining to the limited availability of computing resources. However, as shown in figure 2, there is a clear need for techniques, such as data balancing and masking, to be applied to deal with the proportionally insufficient information about abnormal and borderline placements of catheters. To avoid issues with bias, careful selection of images will be required. To further assess the imbalance between each class in the data, figure 2 demonstrates the degree to which certain classes are overrepresented or underrepresented.

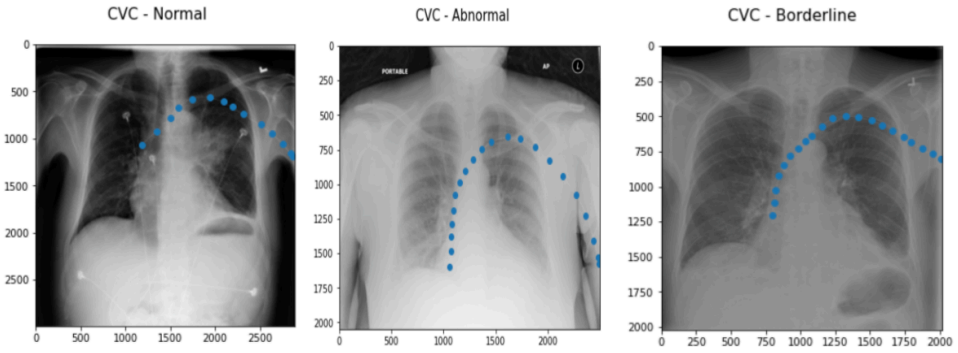
Figure 2: A visual description of the imbalance between classes present in the data provided



As observed, there is a clear majority of normal placements but what is most concerning is the substantial proportion of incomplete images. Furthermore, the comparative lack of malpositioned catheters, especially for NGTs and ETTs, creates a lot of challenges when identifying ways to produce predictions of non-normal catheters.

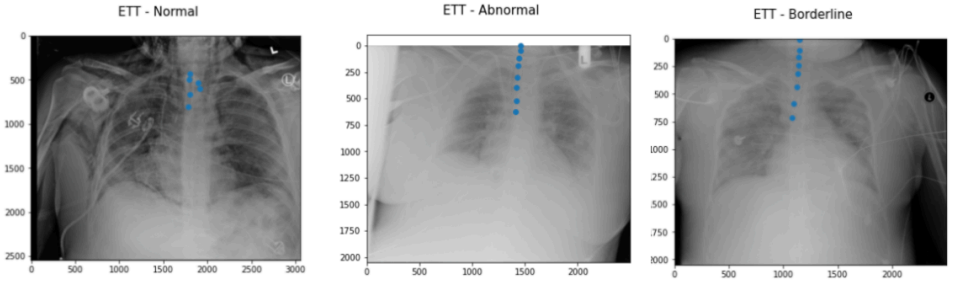
Analysing the images, distinct characteristics can be identified for each type of catheter's placement and status. The main characteristic that differentiates CVC catheter placements seems to be the depth at which the catheter is placed around the cardiac region. In the images shown at figure 3, borderline placements seem to describe catheters that have protruded the wrong area entirely. Abnormal placements seem to allude to partially correct positionings. This may mean that the catheter correctly covers the veins but around the heart, the catheter may be too deep. Normal placements have been correctly connected from the heart to the veins to the left arm.

Figure 3 ; Side by side comparisons of CVC catheter placements show distinct characteristics that differentiate placements.



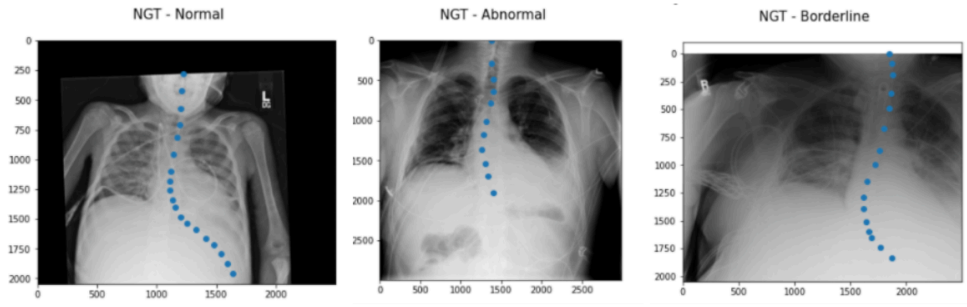
In figure 4, the side by side comparison of ETT placements show that a distinguishing characteristic between normally, abnormally and borderline placed ETT catheters is the depth at which they are placed. Normal catheter placements barely stick through the throat area into the lungs, while abnormal placements have slightly deeper protrusions and borderline placements have deeper protrusions as well as a more rigid placement. Since the neck is slightly curved, a rigid placement may be the differentiating feature between an abnormal and borderline placement.

Figure 4: Side by side comparison between different statuses of ETT placements



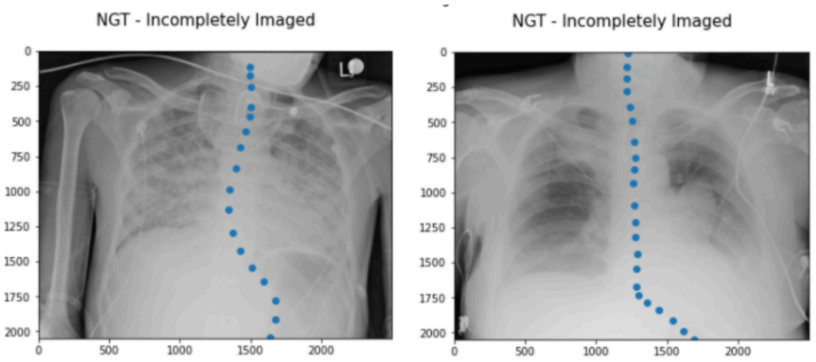
In figure 5, the stark difference between the normal catheter placement of the NGT and the borderline placement seems to be associated with how well the catheter placement is able to penetrate the stomach. If the placement is borderline, it would be safe to assume that it barely penetrates the stomach. If the catheter placement is abnormal, then the catheter does not even reach the stomach region. These differentiating characteristics are expected because our findings indicate that NGT catheters are supposed to deliver nutrients to the digestive system so they must reach the stomach and penetrate it enough for nutrients to be absorbed. However, the lack of depth shown in borderline and abnormal placements prevents the delivery of substances to the stomach.

Figure 5: A side by side comparison of the statuses of placements of the NGT catheter.



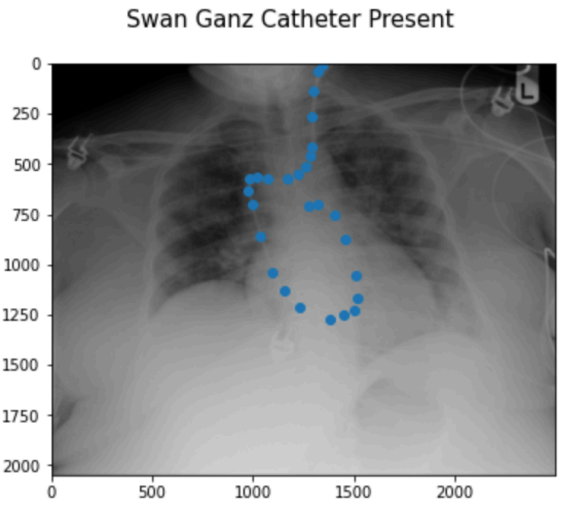
The common characteristic of incomplete images is that they fail to capture the depth at which the catheter penetrates the stomach, as shown in the images on figure 6.. The ambiguity that arises from the incompletely imaged placement is the main reason for exclusion of similar images.

Figure 6: A common trait among incomplete images is their failure to capture important regions in the stomach.



Regarding the remaining type of catheter, the Swans Ganz catheter, as mentioned earlier, there seems to be a lack of information surrounding the degree to which it has been correctly or incorrectly placed. Instead, in figure 7, there does not seem to be a well defined pattern other than that the catheter is connected to a similar region as the CVC catheter, which supports previous findings about the Swans Ganz catheter being connected to the central venous region.

Figure 7: Swans Ganz catheters are connected to similar areas as CVCs but in the data provided, lack a “degree of misplacement”.



A small sample of 750 images was selected to test the KMeans models on due to their dimensional complexity. Of this selection 731 contained a CVC catheter, 137 contained an NGT catheter and 190 contained an ETT catheter. These were selected due to their ease of implementation as an initial model test to expand our understanding of the large images dataset as well as our capability to work appropriately with it. An issue with this approach however is the imbalanced amount of data, with CVC catheters vastly outweighing the rest.

Given the limitations imposed with regards to available computing resources, only a small sample of images could be selected for training and testing models. For the neural network, images were selected at random because balancing using different techniques such as SMOTE, hindered the achievement of better results as indicated later on. Figure 8 outlines the quantities of random images selected for training the neural network models.

Figure 8: Summary of image sample used in testing neural networks.

Images used in RNN	CVC	NGT	ETT
Normal	800	500	800
Abnormal	800	111	30
Borderline	800	219	428

As observed, there is a distinct lack of availability of abnormal images, particularly for the NGT and ETT catheters, which may pose an issue during multi-classification tasks.Furthermore, the issue with random selection, particularly of the CVC images, is that they are not necessarily reflective of a properly balanced population of normal, abnormal and borderline images, which may increase models' vulnerability to disregarding non-normal images due to their lack representation. Consequently, although accuracy may be high, other metrics may suffer from this dilemma of underrepresentation.

Data Preprocessing and Wrangling

In order to be applicable to our KMeans model the data initially had to be imported in a grayscale format to remove the third dimension it defaulted to implementing. A new data frame was then created for each type of catheter which contained the type of catheters contained within the image as well as the feature array for the image. For the binary classification models only the 'Normal' column was kept whereas the multi-classification retained all categories. These feature arrays then had to be resized to the same shape in order for use within the KMeans model. The dimensions of the smallest image within the data frame were chosen to resize the other images to the same shape as it was being implemented on a relatively small test sample of the entire dataset. Finally, the feature arrays were flattened so as to be in the correct format for KMeans.

To prepare the image data for use in a neural network, specifically a convolutional neural network, image files had to be pre-split into train, validation and test directories and pre-classified within these directories. They also had to be resized to match the input layer dimensions. Unlike the KMeans approach, these images were significantly reduced in size to 224 x 224 images because of the high space complexity it would require for these images to be processed in a neural network. Additionally, input layers and hidden layers within the convolutional neural network would need to be enlarged significantly to account for the increased dimensions, increasing the number of operations needed to train the model, thus increasing time complexity to an unmanageable level. Furthermore, when resizing the images the coordinates given in the annotations dataset had to be repositioned to map correctly onto the newly sized images.

To obtain these resized coordinates the annotations dataset was merged with the image dataset based upon 'StudyInstanceUID' however not every image was accompanied by annotations so some images were lost. After this, each of the old x values were divided by the current related image width and then multiplied by the resized image width. For the new y values, the old were divided by the current image height before being multiplied by the desired image height. These new coordinate values would then correctly map onto the resized images. An example is shown in figure 9.

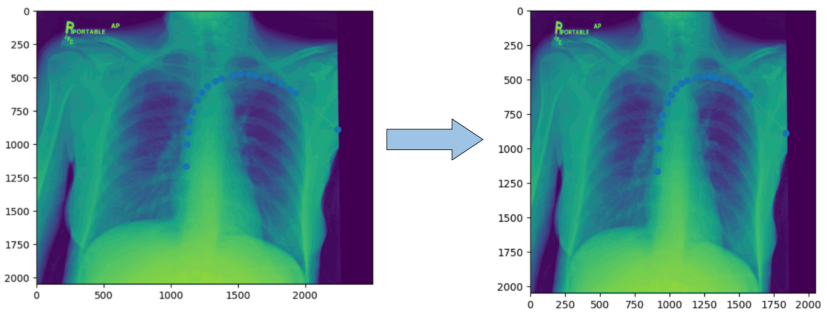


Figure 9: The width of the image has been reduced to a standardised size

Model Development

As seen during the exploratory data analysis in figure 2, the data is significantly unbalanced, both across Catheter types and within each catheter type, as our research question revolves around using a catheter type to predict Normal, Abnormal or Borderline it is not necessary to balance the three catheter types with each other, but only within each catheter type. Balancing can either be done by undersampling which removes valuable data points, or oversampling which creates replicate data points to balance the data, however neither of these are desirable techniques. Synthetic minority oversampling technique (SMOTE) is a balancing technique that uses oversampling techniques to create augmented data points to balance the data, this avoids excessive replicas while still reducing false negatives.

Initially the CVC dataset was balanced using 10% of the data and the kMeans model was rerun, this increased the number of false negatives from 5 to 7 as well as diminishing the accuracy, this is likely due to how severely the data is unbalanced as there isn't enough data points to successfully augment new data points. As this defeats the whole purpose of balancing the data it was decided to continue with the original data going forward.

Before (CVC)		After (CVC)	
True negatives:	2	True negatives:	3
False positives:	1	False positives:	0
False negatives:	5	False negatives:	7
True Positives:	2	True Positives:	0

Figure 10: False Negatives before and after SMOTE

To begin with, we decided to implement a KMeans model upon a binary version of the dataset, attempting to predict whether a catheter was placed normally or not. This model was chosen as it is relatively simple to implement, analysing the feature arrays purely from the images, providing a good basis for beginning our exploration of the data. However it came with it's associated challenges. KMeans assumes that clusters can be formed in circles which presents an issue due to the wide variety in the shaping and positions of the catheters, but it is hoped that the model will be able to identify consistencies between the normally placed catheters, using two clusters as the normal should be reasonably grouped. This may also be limited by the camera alignment when taking the X-ray as some of the images are angled differently resulting in different positions for the same type of catheter.

Building upon this, a multi-classification model was designed in order to predict whether a known type of catheter was normal, abnormal or borderline. Using three clusters, one for each category, this model adopts many of the same assumptions as the binary model however faces further limitations due to the potentially overlapping nature of the additional clusters required which will cause issues as KMeans aims to distinctly group the given data.

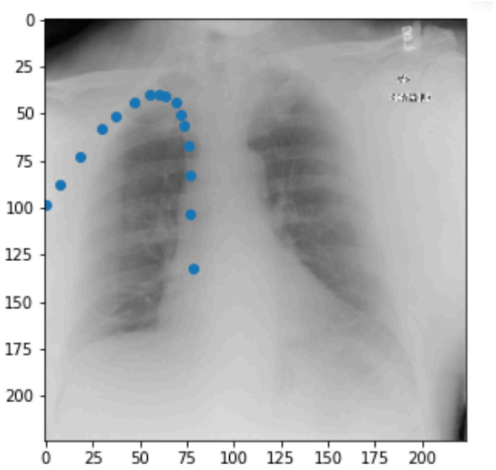
A further challenge to these KMeans models comes in the form of the immense amount of background noise contained within each X-ray image as they are very detailed illustrations. This has the effect of taking the cluster centroids away from their ideal position giving ineffective results due to modelling based upon the other features in the images and not correctly identifying the catheters themselves. This issue can be potentially addressed through the use of the coordinate annotation data, providing the exact catheter positions within the image, and will be further explored in the deep learning approaches.

The first deep learning approach taken was the use of recurrent neural networks (RNNs) to map sequenced coordinate data.Typically, RNNs are reserved for sentiment analysis and natural language processing but by exploiting the long-short-term memory (LSTM) property of such architecture, we can potentially predict catheter placement and assess its accuracy. In some respects, this method was efficient and computationally cheap, which was another main reason for its nomination. However, it had some flaws, regarding the fundamental purpose of its mechanisms. There have been some cases where RNNs can be applied to imaging data; for example, Jurkus (2023) demonstrated the use of LSTM to model “a different trajectory calculation strategy” for maritime trade vessels. In this LSTM approach, we would like to test the validity of the assumption that catheter placements can be treated as trajectories and be used to yield reliable predictions of other images’ catheter placements.

The main underlying assumption made, which may weaken this claim, include that each patient’s chest size, relative to the corresponding x-ray image’s size, is similar. Given that we are mapping sequential coordinates, the assumption made about having standardised chest sizes among patients being a requirement could safely be ignored since we are using RNN models that employ the LSTM property independently for each image. The outputs are “the relationships between values [from] the beginning [to] end of a sequence” (Zvornicanin, 2022) meaning that there is no absolute mapping system for the coordinates. This is a useful property because X rays are often taken at different angles and orientations, even for the same patient’s body parts, as is the case for these images. To quantify the validity of this method, as well as other methods, a confusion matrix was produced and the model’s accuracy, precision and recall were also calculated to compare the performance of the models. Given the stochastic nature of using neural networks, a Monte Carlo simulation was done to produce a confidence interval of the metrics predicted.

Regarding the architecture and features of the RNN model employed across all catheter types, a few dense layers with varying numbers of neurons (64 or 128 neurons) were used at each layer. The main feature of this model was the LSTM layer which was nominated for its ability to retain memory of sequential coordinates.The rationale behind retaining the memory of patterns in sequential coordinates was that similar patterns could also be classified as belonging in the same category. For example, in figure 11, a sequence of coordinates can be predicted based on the trail which the catheter tube follows over time.

Figure 11: A sequence of mapped coordinates on an example X ray for a borderline CVC placement



Hypothetically, the LSTM layer during training would be able to retain the sequence of the catheter's annotation as a borderline case then detect similar trajectories and reach similar conclusions but with the absence of a CNN element, the model may operate on the sequence of coordinates within an image rather than between images. Furthermore, issues may arise if the sequence of coordinates of normal, abnormal and borderline placements follow similar trajectories or patterns.

A model summary is provided below in figure 12. The activation functions used in the construction of the network were sigmoid and softmax for binary and multi-classification respectively as well as the ReLU activation function, mostly for its simple method of introducing non-linearity which is required to handle a sequence of coordinates.

```
Model: "sequential_89"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_89 (LSTM)               (None, 64)                17152
_________________________________________________________________
dense_267 (Dense)            (None, 128)               8320
_________________________________________________________________
dropout_89 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_268 (Dense)            (None, 64)                8256
_________________________________________________________________
dense_269 (Dense)            (None, 3)                 195
=================================================================
Total params: 33,923
Trainable params: 33,923
Non-trainable params: 0
_________________________________________________________________
```

Figure 12 : Model summary of multiclass RNN and CNN hybrid

The second deep learning approach involves using convolutional neural networks which have been trained on images with increased pixel intensity at the annotated areas. The rationale behind this process is that as the image is passed through the neural network, areas with stronger pixel intensity will be honed in on, therefore reducing the background noise of other pixels and producing more accurate results at a more efficient rate. The advantage of this method is that at the testing stage, the images that are being predicted on can be unannotated. This is because the model will be able to detect similar pixel patterns in the unseen data as the patterns highlighted during the training and validation stages. The disadvantages, compared to the previous method, are that this approach is computationally expensive. But with the ability to focus on the annotated areas with higher pixel intensities, the time taken to train models will be reduced and the accuracy will increase.

The architecture of the CNN ResNet50 model was altered so that it could accept greyscale images as inputs, meaning that the dimension of the colour channels was reduced, and its activation function at the output layer was altered according to the binary classification or multi-classification problem being tackled. Since, in this instance, we were utilising transfer learning of a pre-built model, there were approximately 50 layers, including max pooling layers, convolutional 2D layers, and padding layers.

The rationale behind max pooling is to “extract the most important features” (Kumar, 2021) in an image which, for the purposes of our project’s aim, can be very useful if the pixels containing the catheter are intensified. In this case, a function which increased pixel intensity based on annotations’ coordinates was implemented, making the max pooling layer more effective.

Results

K Means

Beginning with our binary identification upon a small sample of 750 images from the train dataset. The initial KMeans models resulted in accuracies of all approximately 0.7 with similarly high recalls as seen in figure 13 below. The ETT model performed the best of the three with CVC and NGT following respectively. While NGT had the lowest accuracy it actually had the highest precision indicating a low false normal identification rate.

Catheter Type	CVC	NGT	ETT
Accuracy	0.660	0.643	0.737
Recall	0.771	0.680	0.800
Precision	0.771	0.895	0.857

Figure 13: Binary KMeans Classification Results

The confusion matrices created for these models as seen in figure 14 also reflect the general high performance of the models in predicting truly normal catheters. In this case ‘False’ resembles a normally placed catheter and ‘True’ is the opposite. It is evident that the next most common prediction seems to be the false abnormal catheters however this isn’t the most undesirable outcome in our medical context as further discussed later on.

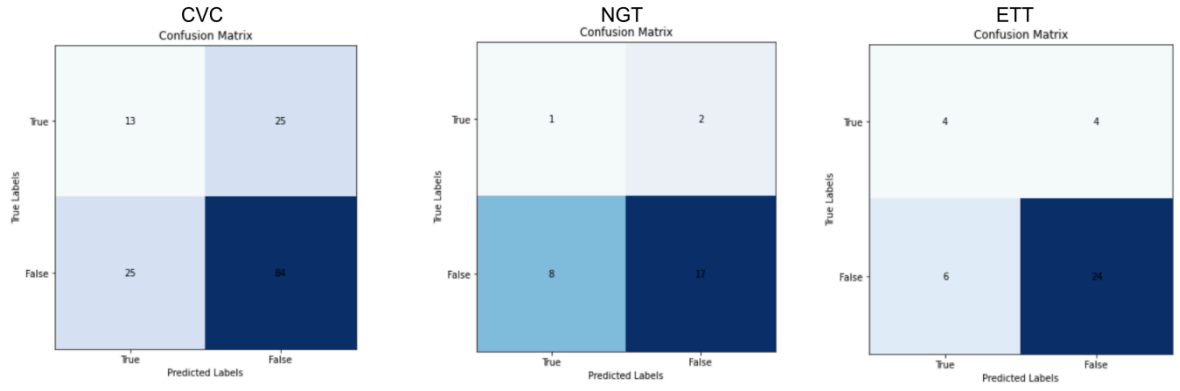


Figure 14: Confusion Matrices For Binary KMeans Classification

Following on, a KMeans multi-classification model was implemented with the same size of sample which determined whether a catheter was normal, abnormal or borderline based upon it’s given type. This resulted in far lower accuracies of 0.361 for CVC, 0.357 for NGT and 0.474 for ETT. While these accuracies are much lower than the binary classification model, they rank in the same order as the binary model with ETT being the most accurate followed by CVC and NGT. A further metric of

KMeans performance is the silhouette score. While the accuracy is easily understandable the silhouette score defines how well the model clusters the information which is vital to a KMeans model. These scores are depicted in figure 15 and are further discussed later on. Additionally, NGt once again had the highest precision followed by ETT and CVC respectively.

Catheter Type	CVC	NGT	ETT
Accuracy	0.361	0.357	0.474
Silhouette Score	0.105	0.193	0.151
Precision	0.697	0.889	0.857

Figure 15: Table of Results for Multi-Class KMeans Classification

Confusion matrices were also plotted for these multi-classification models which from it can be discerned that the ETT model had the most success in predicting the normal catheters, likely due to their abundance. The CVC model correctly predicted more borderline catheters than the other two however also had the highest proportion of incorrect abnormal and borderline predictions however this is preferential to false normal predictions. The NGT model similarly had these predictions although with the correct borderline predictions.

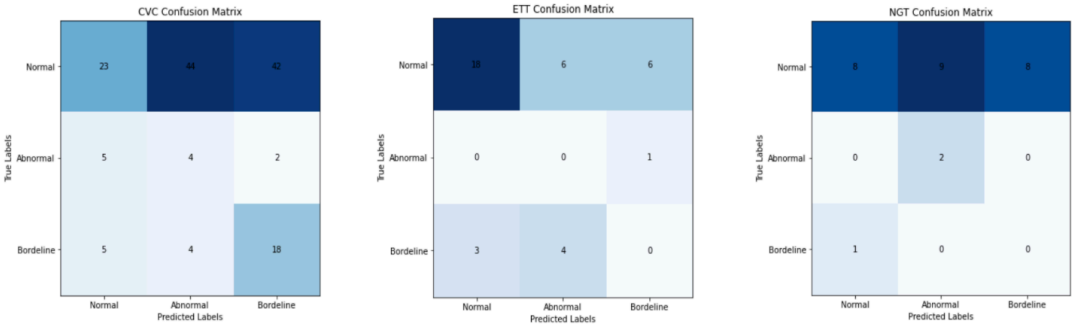


Figure 16: Confusion Matrices For Multi-Class KMeans Classification

Recurrent Neural Networks (RNNs)

The conditions maintained throughout this analysis were to have a consistent random state, a training size of 0.8, a validation size of 0.1 and a testing size of 0.1. The training and validation accuracies of the recurrent neural networks varied and were eventually assigned as the final accuracies achieved by the model with the best weights during the training and validation stages through callback functions. Confusion matrices were produced for both the binary and multi-class approaches for the recurrent neural networks to determine testing accuracy. For the CVC catheter, the results produced for identifying the difference between normal and not normal placements were promising. To reinforce this, a confusion matrix, demonstrating the quality of these differentiations is shown in figure 17.

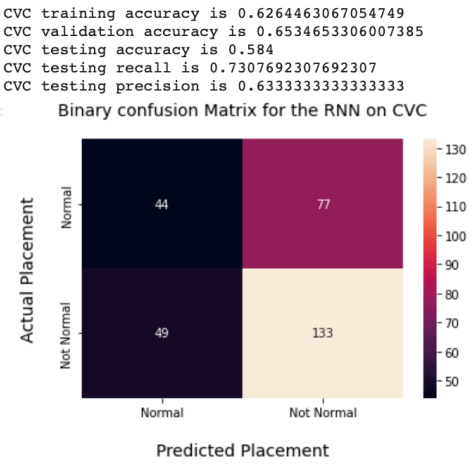


Figure 17: testing on CVC promising start towards identifying false negatives

In this confusion matrix, we can see that although the test accuracy is just slightly better than guessing catheter placements, the recall is reasonably high. In this case, the 'positives' would refer to the catheters which are not normal. The precision is average, showing that the model is neither efficient nor inefficient. To assess the model itself for evidence of underfitting or overfitting, a loss curve for both training and validation loss was plotted on figure 18. In a loss plot, if the training loss (labelled as just "loss" in the legend) is consistently higher than the validation loss then this suggests that underfitting is occurring.

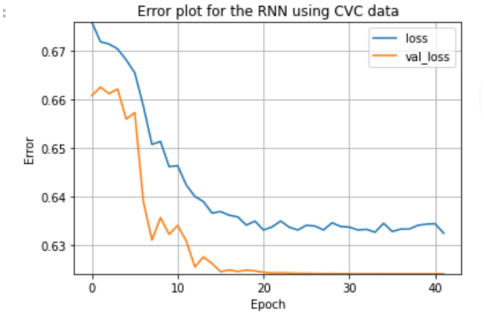


Figure 18: Since the validation loss is consistently below the training loss, there are strong signs of underfitting when testing CVC placement

The presence of underfitting in this neural network suggests that the model is not yet complex enough to detect patterns in the data. The implications of having such a severe degree of underfitting include the production of unreliable predictions. This may reduce credibility of the conclusions made from the models used because the primary purpose of these models was to predict false negatives reliably. It is important to note that this model was trained on an unbalanced selection of data as well which may have hindered its ability to predict results accurately.

When testing on the next catheter, NGT, there was a visible improvement in results, especially accuracy. For example, the training and validation accuracies both saw an increase compared to the other catheter's data. Perhaps what may be concerning is that the validation accuracy is marginally higher than the training accuracy which may suggest that either the model is able to generalise well on unseen data or that there are issues pertaining to the architecture of the neural network, which will be discussed further later on.

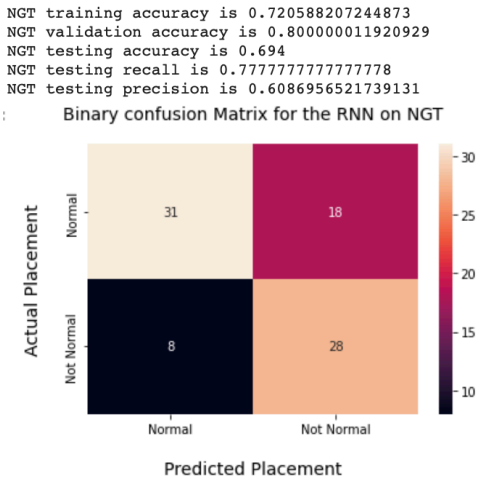


Figure 19: A confusion matrix showing increased accuracy in predictions of NGT catheters and a reduced recall.

From figure 19, it can be seen that the accuracy is greatly improved from the previous catheter but the recall is reduced marginally which cites a decline in model quality. The precision cites that model performs with the same efficiency on this data as before. The loss plot in figure 20 shows evidence of underfitting as well. This may mean that the model's architecture is too simple. Perhaps more properties other than the LSTM layer need to be exploited.

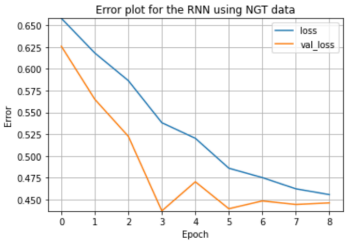


Figure 20: Loss plot shows sign of underfitting as well when using the NGT catheter data.

Finally, for our binary approach to the ETT data, the training and validation accuracies were slightly worse in magnitude. The validation accuracy was slightly higher than the training accuracy, indicating the presence of similar issues as before.. These metrics were calculated from the confusion matrix in figure 21. The precision was mildly increased but at the cost of a significant reduction in recall which is extremely concerning.

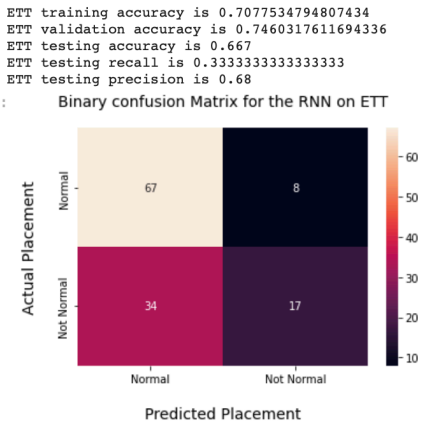


Figure 21: The confusion matrix indicates a somewhat improved accuracy compared to previous types of catheter data and an improvement in recall.

Similarly to before, we can assess the model for evidence of overfitting or underfitting through the corresponding loss plot in figure 22.

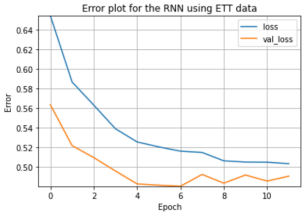


Figure 22: Again, there is clear evidence of underfitting.

Underfitting has occurred, which is likely to affect the reliability of the predictions for this catheter as well.

Upscaling to a multi classification problem, the intention was to maximise the recall of all classes if possible with the priority of maximising the recall of the “abnormal” and “borderline” classes to reduce instances of false negatives among these classes. Having false negatives among the “normal” class was not as much of a priority because in context, although having high recall among the “normal” class can increase the efficiency of automating diagnosis of catheter placement, the implications of identifying false negatives among normal placements are less severe

For the CVC, the training and validation accuracies were very low, only slightly performing better than guessing classes. The results produced in the corresponding confusion matrix show that the testing accuracy for this run was higher than both the training and validating accuracy scores which is cause for concern. When viewing the confusion matrix in figure 23 as well, it can be seen that there are extremely poor predictions being made about borderline placements of catheters which already indicates poor model performance. Regarding the efficiency of this model, which is quantified by its ability to avoid instances of false negatives among the normal class, this model had an average efficiency, which is consistent with the findings of the binary classification approach.

```
CVC training accuracy is 0.3987603187561035
CVC validation accuracy is 0.3861386179924011
CVC testing accuracy is 0.406
CVC testing recall for normal is 0.9834710743801653
CVC testing recall for abnormal is 0.011764705882352941
CVC testing recall for borderline is 0.030927835051546393
CVC testing precision for normal is 0.40476190476190477
CVC testing precision for abnormal is 0.2
CVC testing precision for borderline is 0.75
```

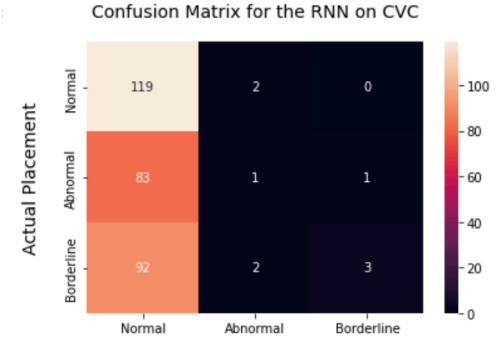


Figure 23: Almost no borderline or abnormal placements were predicted despite the dataset compromising roughly one-third of each type of placement which may indicate underlying issues with neural network architecture or random seeding.

The loss plot for this model interestingly shows moderate signs of overfitting as the validation loss is consistently y larger than the training loss of the model, indicating that the model was trained too well on the training data and consequently failed to perform well on unseen test data. This is unusual given that the accuracy for validation is higher than training.

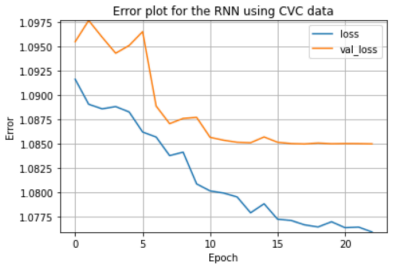


Figure 24: The loss plot for this model shows levels of overfitting unlike previous models, which may explain the poor predictions. Perhaps, the random split did not contain enough abnormal instances in its validation and training sets, leading to this issue.

Proceeding to the next catheter, the NGT, almost all metrics were greatly improved compared to the CVC data. The recall for identifying normal instances of placement was extremely high which is promising, in terms of its efficiency.

```
NGT training accuracy is 0.7308823466300964
NGT validation accuracy is 0.7764706015586853
NGT testing accuracy is 0.682
NGT testing recall for normal is 0.9387755102040817
NGT testing recall for abnormal is 0.5454545454545454
NGT testing recall for borderline is 0.24
NGT testing precision for normal is 0.7301587301587301
NGT testing precision for abnormal is 0.5
NGT testing precision for borderline is 0.6
```

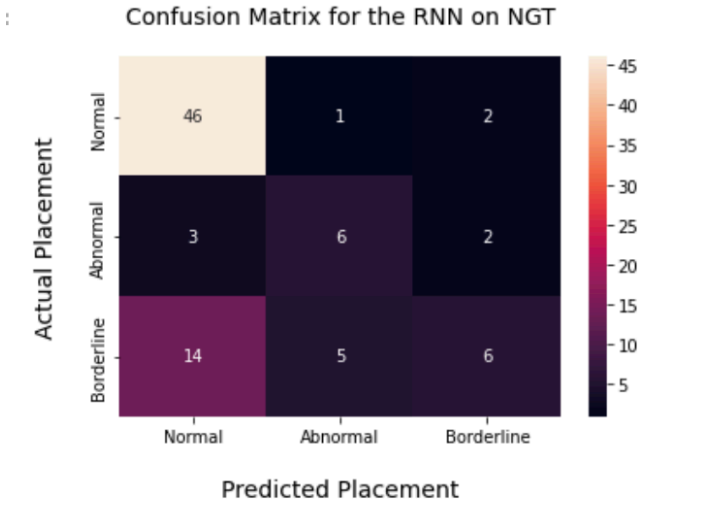


Figure 25 : Majority of predictions are correct and as expected, they are in the most prevalent class

To analyse further, the loss plot at figure 26 for this run shows that the training loss is marginally higher than the validation loss as expected, based on the results from the accuracy scores but once again, this is indicative of underfitting, providing further evidence of the modell used being too shallow and lacking the required complexity to properly predict instances of catheter placement.

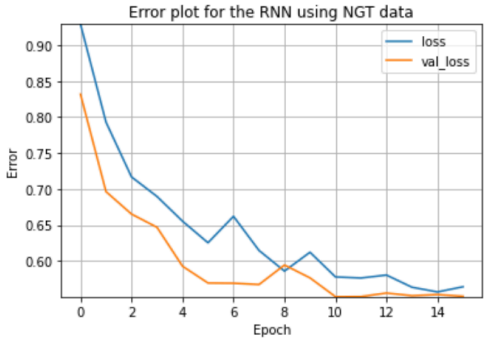


Figure 26:: This follows the previous findings about signs of underfitting.

Lastly, for the ETT catheter, similar conclusions were reached as the training accuracy was lower than the validation accuracy. The testing results in this case were similar with respect to the magnitude of their metrics. However, the model failed to pick up recall for abnormal placements due to a lack of data. The efficiency of this model was also similar for this data. The model did not perform too well when trying to reduce instances of dangerous false negatives though.

```
ETT training accuracy is 0.7077534794807434
ETT validation accuracy is 0.7301587462425232
ETT testing accuracy is 0.69
ETT testing recall for normal is 0.8933333333333333
ETT testing recall for abnormal is 0.0
ETT testing recall for borderline is 0.43478260869565216
ETT testing precision for normal is 0.7052631578947368
ETT testing precision for abnormal is 0.0
ETT testing precision for borderline is 0.6451612903225806
```

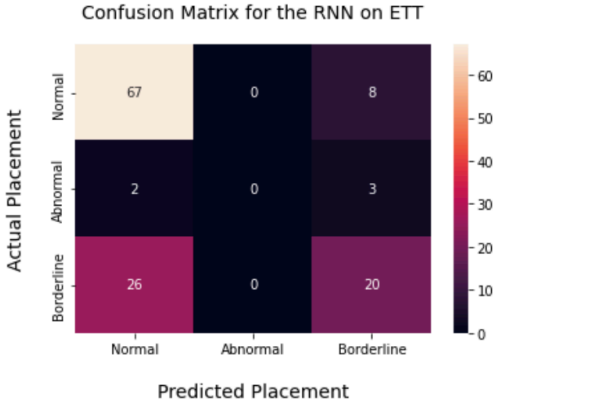


Figure 27 : No abnormal placements were predicted. They were extremely scarce in the testing set as well which may have explained this occurrence.

Once again, to assess the model for overfitting or underfitting, a loss plot was used. Figure 28 reinforces that the model is consistently underfitting the data it is predicting. This reaffirms the argument that more complex methods or architectures are needed for the neural network to operate effectively on the data.

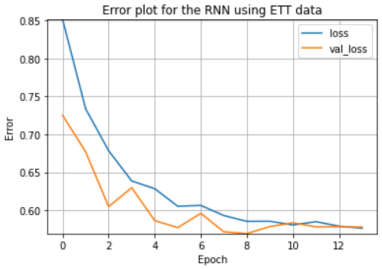


Figure 28:: On the final configuration of the neural network to the last type of catheter, clear signs of underfitting can be seen once again.

A summary of all results, including the previously mentioned results for binary classification using the recurrent neural network modelling is provided below in figure 29. The results for each catheter in the multi-classification process is provided in figure 30..

Figure 29: A tabulated summary of binary classification results for randomised runs of each catheter type

Binary Classification	CVC	NGT	ETT
Training Accuracy	0.63	0.73	0.71
Validation Accuracy	0.65	0.8	0.75
Testing Accuracy	0.58	0.69	0.67
Testing recall	0.73	0.78	0.33
Testing precision	0.63	0.61	0.68

Figure 30: A tabulated summary of multi-classification results for randomised runs of catheters

Multi-classification	CVC	NGT	ETT
Training Accuracy	0.40	0.73	0.71
Validation Accuracy	0.39	0.78	0.73
Testing Accuracy	0.41	0.68	0.69
Normal Recall	0.98	0.94	0.89
Abnormal Recall	0.01	0.55	0
Borderline Recall	0.03	0.24	0.43
Normal Precision	0.40	0.73	0.71
Abnormal Precision	0.20	0.75	0
Borderline Precision	0.75	0.6	0.65

As mentioned earlier in the model development section, training neural networks is a stochastic process with respect to the initialisation and convergence to their final weights and configurations. Therefore, after each training, validating and testing session, different results will be outputted, even for models with the same architecture and random split of data. Therefore, to account for this randomness, a confidence interval was generated for each of the relevant metrics using a Monte Carlo simulation which approximated to a normal distribution through the use of the central limit theorem. More specifically, 30 runs of the stochastic training and validating process for the RNN were done on identically distributed data, (the training, validation and testing sets remained consistent throughout) whereby in each run, an identical but new model was made with weights that were initialised randomly and independently from other runs then trained, validated and tested on with the identically distributed data. After each run, the accuracy and recall were recorded then interval estimation was performed on these data. A 95% confidence interval was chosen because in context, assessment of the model's metrics need to be as certain as possible but not at the expense of

increasing the interval estimation too much until the findings are meaningless. Figures 31 and 32 show the 95% confidence interval for each metric..

Figure 31: A tabulated summary of the 95% confidence intervals for the results of each metric in binary classification

Binary Classification	CVC	NGT	ETT
Training Accuracy	(0.607,0.616)	(0.758,0.780)	(0.700,0.710)
Validation Accuracy	(0.628,0.637)	(0.798,0.807)	(0.747,0.755)
Testing Accuracy	(0.595,0.606)	(0.730,0.753)	(0.678,0.706)
Testing Recall	(0.90,0.982)	(0.606,0.698)	(0.495,0.595)
Testing Precision	(0.602,0.619)	(0.702,0.745)	(0.612,0.668)

Figure 32: A tabulated summary of the 95% confidence intervals for the results of each metric in the multi classification

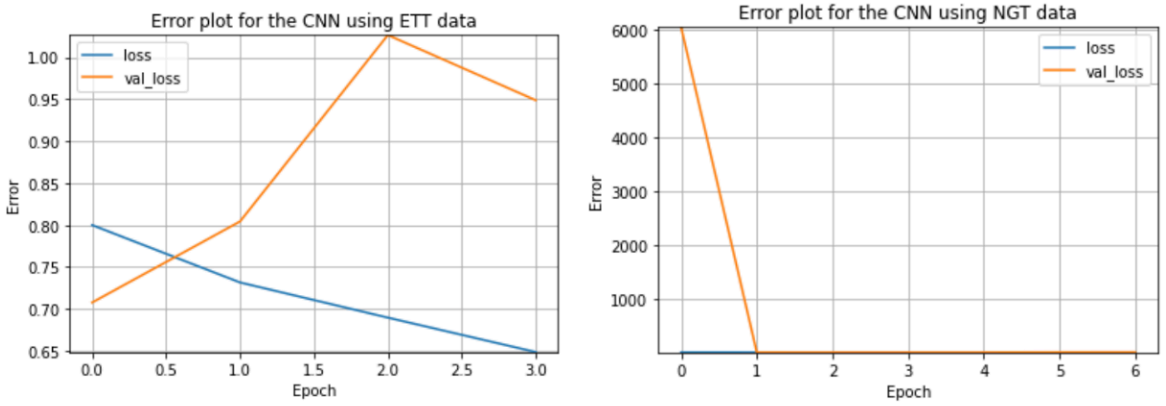
Multi-classification	CVC	NGT	ETT
Training Accuracy	(0.407, 0.431)	(0.729, 0.749)	(0.681, 0.699)
Validation Accuracy	(0.394, 0.426)	(0.761, 0.772)	(0.732, 0.739)
Testing Accuracy	(0.410, 0.433)	(0.674, 0.690)	(0.667, 0.701)
Normal Recall	(0.842202, 0.940167)	(0.874, 0.915)	(0.847, 0.903)
Abnormal Recall	(0.106792, 0.313600)	(0.621, 0.652)	(0,0)
Borderline Recall	(0.011, 0.031)	(0.242, 0.328)	(0.365, 0.529)
Normal Precision	(0.417, 0.450)	(0.748, 0.773)	(0.688, 0.732)
Abnormal Precision	(0.129, 0.268)	(0.494, 0.546)	(0,0)
Borderline Precision	(0.236, 0.510)	(0.488, 0.607)	(0.508, 0.668)

Convolutional Neural Networks (CNNs)

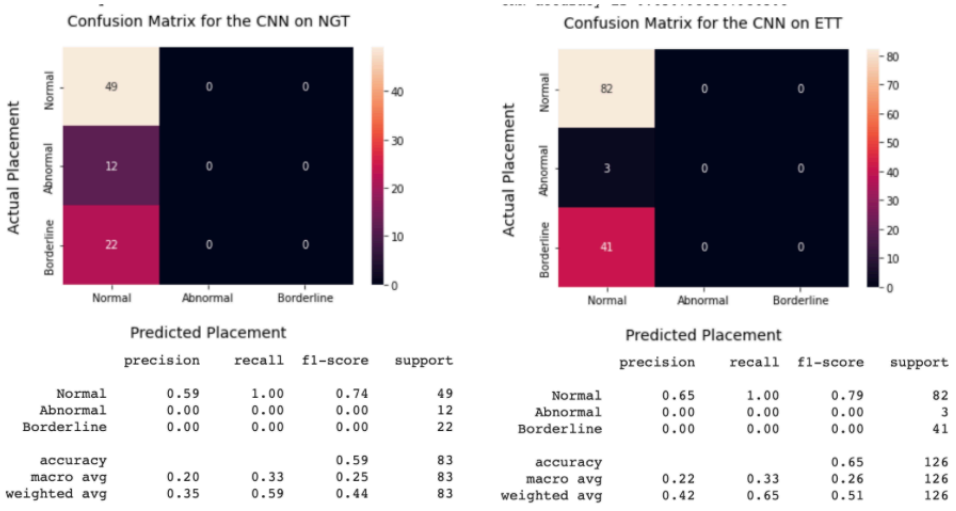
The ResNet50 model was computationally expensive to operate and its results, especially the error plots, demonstrated extreme levels of overfitting when the method of increasing pixel statistics around

the regions of interest was employed. Figures 33 and 34 show extreme levels of overfitting when ResNet50 was fitted to the image data of the NGT and ETT catheters respectively.

Figures 33 and 34 : the validation loss is extremely high then declines suddenly. This may be evidence of severe overfitting (left). The validation loss continues to increase while the training loss declines, citing clear overfitting issues



Subsequently, as seen in the corresponding confusion matrices at figures 35 and 36, the predictions made were not meaningful and this process was abandoned.



	precision	recall	f1-score	support
Normal	0.59	1.00	0.74	49
Abnormal	0.00	0.00	0.00	12
Borderline	0.00	0.00	0.00	22
accuracy			0.59	83
macro avg	0.20	0.33	0.25	83
weighted avg	0.35	0.59	0.44	83

	precision	recall	f1-score	support
Normal	0.65	1.00	0.79	82
Abnormal	0.00	0.00	0.00	3
Borderline	0.00	0.00	0.00	41
accuracy			0.65	126
macro avg	0.22	0.33	0.26	126
weighted avg	0.42	0.65	0.51	126

Figures 35 and 36: These predictions are meaningless as all images are being classified as normal placements. Therefore, the method was excluded for its lack of contributions to a meaningful conclusion and computational expenses.

As mentioned earlier, due to the high amount of computational resources used to complete one epoch of ResNet50, Monte Carlo simulation methods were abandoned.

Further Analysis of Results

As described earlier, it can be seen that reasonably accurate results were obtained for the binary KMeans classification models. The relatively high recall scores also meant that false negatives were being reduced, however these could be further improved. In the associated correlation tables, falsely abnormal predictions were second most common which is preferred to falsely normal predictions as

these could lead to extensive complications if someone was misdiagnosed with a normally placed catheter. The amount of false normal predictions was also seen to be minimised the most in the NGT model as it had the highest precision rating. This implied that the procedure was relatively efficient.

The multi-classification KMeans model had much poorer accuracies than the binary model however the same order of improving accuracies, reflected across both tests with ETT as the most accurate suggests more definition between catheters within the ETT images allowing for better classification by the KMeans modelling. The continued higher precision of the NGT model demonstrates it's stronger ability to reduce falsely normal predictions which is very beneficial with regards to increasing the efficiency of identifying misplaced catheters. The silhouette scores described previously range from -1 to 1 with higher scores meaning well defined points. The silhouette scores associated with our models are all positive which suggests there is slight definition of the clusters however all scores are still quite close to 0 meaning that there are many points on the borders of clusters. This makes it difficult to obtain the overwhelmingly accurate model desired for medical purposes.

In figure 31, it can be seen that the binary RNN model's efficiency in handling misplaced catheters of all types is slightly above average. This suggests that the RNN model does not really save healthcare workers'' time or a hospital's resources. The RNN model was able to avoid false negatives effectively when handling CVC placements but struggled to reduce false negatives when handling other catheters. This suggests that the effectiveness of the RNN in achieving its purpose is limited. Despite the fact that the model managed to reduce instances of false negatives among CVC placements, the accuracy of its predictions was significantly lower than other catheter types which discredits its reliability. Both the NGT and ETT catheters had similar, above average levels of accuracy when differentiating between normally placed and misplaced catheters. However, none of the model's performances exceeded the expected levels of accuracy, recall and precision required to be applied to a hospital. (typically, hospitals would expect accuracy to exceed 0.95 at least to be safe to use)

As shown in figure 32, the RNN model used on multi classification performed similarly in terms of accuracy which was expected, given that minimal changes were made to the structure of the RNN used for binary classification. The model's consistent performance in maximising recall of normal catheter placements demonstrates that the model is able to determine correctly whether a patient has a normally placed catheter which contributes strongly towards reducing the wastage of resources on manually inspecting normally placed catheters. The relatively low recall among abnormal and borderline predictions implies that the model was not able to proficiently reduce instances of false negative cases. Perhaps the only exception to this could be that the model's ability to detect false negatives among NGT placements was average, particularly when predicting instances of abnormal placement. However, the consequences of abnormal placements are not as severe as borderline cases so the usefulness of this finding may be reduced. The lack of precision among both abnormal and borderline cases shows a lack of the model's confidence in its predictions which further undermines its credibility for use in a hospital.

Evaluation

KMeans was easy to implement as an initial model, benefitting our understanding of the data, however it lacked complexity in it's definition of the catheter location. The use of circular clusters resulted in decreased accuracy when identifying the complex shapes of various catheters. While it's performance as a binary classifier was reasonable this overlapping of clusters was keenly seen in the multi-classification models and their associated low silhouette scores. Furthermore, KMeans is easily susceptible to 'The Curse of Dimensionality' in which the model degrades as the number of features increases, the computational complexity increases, and it can begin to model based upon a lot of the noise within the images. To improve upon these models in future the images could be further reduced

allowing for even more images to be analysed. Other more complex clustering models such as DBSCAN, which better defines outliers, could also be implemented in order to obtain an improved result.

The RNN model was not tested on properly balanced data. Instead, image data was selected randomly. This may have led to a reduced quality in results as bias and variance among data was not properly managed.

Although the RNN method by itself was computationally cheap, to run a Monte Carlo simulation on a sample size of 30 to approximate the results to a normal distribution, the computing resources and time taken were greatly increased. To save computational resources and time, the Monte Carlo simulation could have had a reduced sample size and assumed a t-distribution instead to predict the confidence intervals of different metrics. This could have reduced the time taken to handle the stochastic elements of the neural network. The normal approximation assumption also had considerable shortcomings in that it assumed that the variance of the distribution of the predictions was consistent which is likely untrue.

As outlined consistently throughout the results, the construction of the RNN model was too simplistic to handle patterns in the sequential coordinate data. To improve on the RNN approach, more modifications needed to be made at deeper layers to reduce underfitting so that patterns in the data could be easier to detect. A major issue that undermined the results of this model was that coordinate data may have been treated as time series frames in a moving sequence of images which is problematic given that the images do not have a time series dependency element at all. If LSTM were to still be employed to map out the trail of catheter placements, modifications would need to be made regarding the implementation of multilayered independence to treat each image as an independent instance and a convolutional layer would at least need to be implemented. As acknowledged earlier in this report, the approach taken to use LSTM on a sequence of coordinates that were not connected by time series dependencies had fundamental shortcomings. A possible alteration that could be made to the data included providing a timestamp for the time that each image was taken along with the patient's ID to add a time series element but convolutional layers would still need to be implemented to analyse the sequence of coordinates as a whole. If the task given was to predict placements of catheters as doctors inserted them in real time then this current approach would have been very effective. However in practice this is not realistic and this is not how x-rays work.

Furthermore, as suggested before, the CNN approach chosen had an extreme degree of overfitting, producing unmeaningful results which suggests that the architecture of the ResNet50 implementation needed to be modified further, with regards to the deeper layers to reduce the model complexity. The approach taken may not have been correct as well. Increasing pixel intensity may not have been a sufficient technique to highlight the region of interest for the kernel to focus on in the neural network. Hyperparameter adjustments were likely required at more complex levels beyond the scope of what was explored.

To enhance the current approaches, a fusion of convolutional and recurrent layers could have been implemented to determine if metrics would improve or not. Another potential approach that could have enhanced the results significantly was to use transformers to increase the attention of highlighted areas in the image instead of using pixel intensity statistics.

Conclusion

The primary research question was 'Is a type of Catheter able to be classified as either Normal, Abnormal or Borderline', this is relevant as serious complications due to catheter misplacement can

occur, the protocol for a qualified individual to manually sort through x rays to determine misplacement is time consuming and also prone to human error, early recognition of malpositioned tubes is also critical and prevents complications. A kMeans and RNN model were developed to achieve this.

The main measurements of model performance were accuracy, recall and precision. Accuracy provides an overall statistic on how well the model predicts whether a catheter is normal or not and recall is a measure of the amount of false negatives in the model prediction, a higher recall indicates less false negatives, and precision is a secondary measure of efficiency to demonstrate if the methods implemented help hospitals improve or not. In context, recall is particularly important as a false negative means the model has taken an abnormal or borderline catheter, and reported that it is normal, this risks medical complications and prevents early detection, both of which are the main motivation behind the research question. Overall the RNN model performed better than the kMeans model, in particular it achieved much higher recall indicating that it would be the preferred model for the task at hand. When considering the implementation of either of these models in a practical sense, the recall would need to be improved as any false negatives are problematic when diagnosing real-life situations, and thus further model development would be required.

The models developed can be further refined using the methods outlined above to assist with performance and solidify their purpose in a practical sense, in theory the neural network model can also be improved by adding the element of time, for example mapping the coordinates of the catheter as it enters the body as time series data significantly helps RNN, however in practice this is not realistic and this is not how x-rays work. To assist with data management hospitals can organise their X-ray images into the three Catheter types to begin with or further models can be developed to first recognise whether each type of catheter is present before determining if it has normal, abnormal or borderline placement, attempting to minimise the amount of incomplete images by taking full X-rays of every patient will also avoid data being lost when modelling.

References

What Is A Central Venous Catheter? (2022, July 28). Cleveland Clinic. https://my.clevelandclinic.org/health/treatments/23927-central-venous-catheter

Felipe-Silva, A., & Campos, F. P. F. de. (2012). Nutrothorax complicating a misplaced nasogastric feeding tube in a severely ill patient. Autopsy and Case Reports, 2(1), 19–24. https://doi.org/10.4322/acr.2012.003

Hosseini, B., Montagne, R., & Hammer, B. (2020). Deep-Aligned Convolutional Neural Network for Skeleton-Based Action Recognition and Segmentation. *Data Science and Engineering*, *5*(2), 126–139. https://doi.org/10.1007/s41019-020-00123-3

Jurkus, R., Treigys, P. & Venskus, J. (2021). Investigation of Recurrent Neural Network Architectures for Prediction of Vessel Trajectory. *Communications in Computer and Information Science*, 194–208. https://doi.org/10.1007/978-3-030-88304-1_16

Kumar, P. (2021, August 25). *Max Pooling, Why use it and its advantages.* Geek Culture. https://medium.com/geekculture/max-pooling-why-use-it-and-its-advantages-5807a0190459#:~:text=Max%20Pooling%20is%20an%20operation

OpenAI (2023), ChatGPT (September 2021 version), [Large Language Model] https://chat.openai.com/c/5b7d6410-15d3-4607-8a70-9cfa61ad1035

Miller, K. A., Kimia, A., Monuteaux, M. C., & Nagler, J. (2016). Factors Associated with Misplaced Endotracheal Tubes During Intubation in Pediatric Patients. *The Journal of Emergency Medicine*, *51*(1), 9–18. https://doi.org/10.1016/j.jemermed.2016.04.007

Mon, W., Boyle, H., & Thoburn, C. (2016). Misplacement of central venous catheter into the left internal mammary vein. *Anaesthesia Cases*, *4*(2), 102–105. https://doi.org/10.21466/ac.mocvcit.2016

Nasogastric Tubes (Insertion and Feeding). (n.d.). Www.nationwidechildrens.org. https://www.nationwidechildrens.org/family-resources-education/health-wellness-and-safety-resources/helping-hands/nasogastric-tubes-insertion-feeding#:~:text=The%20NG%20tube%20is%20placed

Swan-Ganz - Right Heart Catheterization. (n.d.). Ucsfhealth.org. Retrieved October 10, 2023, from https://www.ucsfhealth.org/medical-tests/swan-ganz---right-heart-catheterization#:~:text=Swan%2DGanz%20catheterization%20is%20the

Arrhythmias - What Is an Arrhythmia? | *NHLBI, NIH*. (2022, March 24). www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/arrhythmias#:~:text=An%20arrhythmia%2C%20or%20irregular%20heartbeat

Zvornicanin, E. (2022, January 25). *Differences Between Bidirectional and Unidirectional LSTM* | *Baeldung on Computer Science*. Www.baeldung.com. https://www.baeldung.com/cs/bidirectional-vs-unidirectional-lstm