School of Electronic Engineering
and Computer Science

Interim Report

Programme of study:
BSc Computer Science

Project Title:
Constructing a Machine Learning algorithm used in the detection of pancreatic cancer and converting that into TinyML

Supervisor:
Dr Joseph Doyle

Student Name:
Gargi Agrey

Final Year
Undergraduate Project 2022/23

Queen Mary
University of London

Date: 27/11/2022

Abstract

Artificial Intelligence is being rapidly used in almost every industry however the applications in the field of healthcare and medicine is still an emerging area but promises great potential. By enabling these intelligent systems to intelligently assist not replace doctor's tasks can be a true breakthrough.

This report discusses how AI is being used currently in healthcare, current research and practises and further development.

Contents

Chapter 1: **Introduction**

1.1 Background

Pancreatic cancer is known to be the deadliest cancer with only a survival rate of 11%. The reason for such little survival rate is because it is only detected at a much later stage once the cancer has become a lot worse. Early detection for this cancer still remains a pretty big challenge.

The standard approach to detect pancreatic cancer is by CT scans however it misses about 40% of tumours which are less than 2cm hence why once the tumour has grown significantly is the CT scan able to detect it.

This remains a huge problem and seeing how the remarkable accuracy and precision Artificial Intelligence delivers, applying it into the field of medicine would be a tremendous breakthrough.

The main problem with detecting a pancreatic tumour is how it can change rapidly in shape, size and locality around the pancreas. Due to the small size of the pancreas it only occupies 1.3% of a CT scan. Another problem is also the lack of specific pancreatic tumour image datasets.

1.2 Problem Statement

Despite there being existing ML algorithms used in the detection of pancreatic cancer, my project attempts to improve the accuracy of the existing neural networks already created by being able to detect tumours at less than 1cm, increase validation with more datasets and correctly distinguish a pancreatic tumour since it can simulate other related diseases such as pancreatis. Furthermore, to improve accuracy I will be using only one imaging plane and conducting a detailed automated preprocessing of the original data as this has tampered with the correctness of the model in the past. By converting this into a Tiny ML application additionally the performance and acceleration of the ML will improve significantly.

1.3 Aim

The objective is to create an efficient Machine Learning algorithm and model used in the detection of pancreatic cancer and be able to convert that into Tiny ML. The algorithm will be passed datasets from two categories: diagnosed and undiagnosed patients which directly impact the certainty of knowing if a patient does or doesn't carry the disease. The aim is to then be able to deploy that algorithm to Tiny ML. This seeks to consume less power and memory and still provide effective and immediate results.

1.4 Objectives

The high-level objectives of this project can be summarised in these three key components:

- Fully functional Machine Learning algorithm and model for detecting pancreatic cancer

- Fully functional deployed Tiny ML applications that carry the ML models for pancreatic cancer and heart disease

- A comparison of the TinyML model and the ML model in terms of latency, memory usage and accuracy

1.5 Research Questions

The questions this project aims to answer is stated as follows:

- How can tumours less than 1cm be detected?

- How can you minimise memory usage with these algorithms?

- Will the model(s) still run accurately on TinyML?

1.6 Report Structure

This is an example. [delete as appropriate]

Chapter 2: **Literature Review**

2.1 Machine Learning

2.1.1 What is Machine Learning?

Machine learning is a branch from Artificial Intelligence that has a key focus on learning from data to improve its performance on a series of tasks by attempting to imitate human behaviour. Through thorough statistical analysis, machine learning algorithms can be used for a variety of tasks. The 3 main types of Machine Learning are Supervised, Unsupervised and Reinforcement learning. Supervised learning is performed by using labeled datasets to train and feed the algorithm to perform tasks such as regression or classification. Regression is used to predict outcomes and classification is used to categorize the data into classes.

Unsupervised learning on the other hand uses unlabeled datasets to perform tasks such as clustering and association. The whole purpose of unsupervised learning is to discover hidden patterns and grouping of data without human interaction. Clustering is similar to classification however in clustering the goal is to group unlabeled objects into groups. In association the algorithm is trying to learn from the unlabeled dataset and then trying to discover relationships behind the datasets.

In reinforcement learning, the whole purpose is to optimise decision-making, this is done via an intelligent agent.;. who attempts to make the best decision given the data around its environment. Reinforcement learning can be used for classification as it goes through all the possible classes to find the best one however the only issue with this is that since it is unlabeled data the convergence rate slows done a lot in comparison to supervised learning.

2.1.2 Machine learning in the field of healthcare and medicine

Given the uncertainty and unpredictability in the healthcare field, artificial intelligence has huge potential to make a significant impact. The main applications AI has in healthcare include patient diagnosis, treatment recommendations, supporting clinical decisions, robotic surgeries and virtual assistance. The whole purpose of AI being used in this sensitive field is to aid physicians not replace them due to ethical concerns.

Due to the accuracy and precision of AI algorithms and models there have been cases in which they have outperformed medical professionals in diagnosing diseases by being able to detect various segments of a disease and being able to predict patients who are at a greater risk of getting diagnosed. This hence gives us hope that serious changes can be made here that can lower the rate of deaths and change and save many lives.

There are a series of different AI technologies that can be utilised to revolutionise the field of healthcare. These include machine learning, natural language processing and robotics.

Machine learning can particularly be used for the prediction and detection of diseases by using supervised learning. The reason for using this type of ML is because labeled datasets improve the efficiency of classification since the neural network is flexible in adjusting its weight even if it selects the wrong class and given that we are working in an incalculable environment precision and rapidity are highly crucial which unsupervised learning wouldn't be able to meet.

NLP or natural language processing can be used in areas like speech recognition, entity and character recognition and text analysis and there has been an emerging ability for robots to be used in surgeries and automating administrative tasks.

2.1.3 Established applications

One solution that was developed by IBM is Watson Health which aimed to deliver various solutions such as providing insights to oncologists, matching patients to clinical trials and more however this down fell significantly due to firstly rushing to complete their goal and solution and incorrect data. The algorithm was unable to learn the different types of cancer and provide suitable treatment recommendations due to the fact that humans kept on feeding Watson information on how they think the treatment recommendations should be rather than it attempting to do that from the original data and give its own insights. The key intention of Machine Learning is to learn from data it is given and by tampering with the data it can certainly cause inaccuracy and crash.

Another example is the Streams application developed by Google in attempt to integrate AI prediction models into that app which contained a detail dataset of patient records. This however collapsed due to DeepMind not following the privacy regulations and not having implemented their goal of fusing AI into the app as promised.

The key takeaway from the above two endeavours is that sustaining the correctness of data whilst still protecting patient's privacy still remains a challenge in this emerging field.

As stated in a study (Ke Si, 2021), it was mentioned that the key to successful clinical application of a deep-learning framework is detailed automated preprocessing of the original data. These steps would include data cleaning, integration, reduction and transformation.

Studies have opted for targeted screening and assessing a sub-population at a higher risk (Ananya Malhotra, 2021) rather than using a general population due to a higher cost and time. While this isn't an unreasonable choice since higher-risk patients are given a bigger priority, it isn't necessary that only people in that group are prone to getting diagnosed with pancreatic cancer. Even if a patient does not inherit the genes that can put one at risk of getting the disease or have

chronic inflammation, the sure cause is still unknown. These two factors simply put an individual at a greater risk.

After assessing the existing systems and algorithms I have decided that I am going to take time to find an appropriate dataset, thoroughly perform preprocessing and attempt to use several ML techniques to train my model to better the validity.

2.1.4 What is TinyML and how can it help?

TinyML is a field in Machine Learning that aims to combine embedded systems with Artificial Intelligence by integrating sensors to allow ML algorithms to run by constraining resource and power usage but still running efficiently and effectively. One of the reasons Artificial Intelligence hasn't been 100% materialised is due to the fact that it has AI/ML applications require a substantial amount of resources such as large datasets which is resource restricted. Since there is a tremendous amount of data in the healthcare and medicine field we can understand why AI hasn't fully been transpired in this discipline. As a result, we can fix this by using TinyML which is a low power, low cost and low resource application that will run just as successfully and hope that medical professionals will be more favoured to then utilising Artificial Intelligence in their area. TinyML also promises data privacy on their devices (Adam Zewe, 2022) which is essential as that was one of the disaster of the DeepMind's project.

2.2 A deep-learning model used for pancreatic cancer detection by National Taiwan University

I have chosen the study conducted by the National Taiwan University as my starting point, the following describes thoroughly what was the main objective, the method, results, limitations and the conclusion.

The main goal of their project was to construct a Deep Learning model to detect pancreatic cancer at CT. CT scans are medical imaging techniques used to display internal parts of the body. The way they proceeded to develop this model was by creating a segmentation convolutional neural network which is when a visual input is divided into segments to make the analysis easier. I believe adopting this strategy was wise as the main problem with diagnosing pancreatic cancer is its difficulty to detect in the CT scan, by using segmentation the DL model will be able to extract the object (in our case the tumour) by pixels which can improve the accuracy for classification. The study also used a classifier ensembling five CNNs. The results were as follows: within 546 patients diagnosed and 733 undiagnosed the model achieved 89.9% (with pancreatic cancer) and 95.9% specificity (without pancreatic cancer) and an

area under the receiver operating characteristic curve of 0.96. While this result is highly precise, the 6% difference does support my statement in that it isn't necessary to be still have pancreatic cancer despite all the symptoms and factors. The AUC result shows that the overall diagnostic performance of the test was indeed successful. The model was also able to distinguish between malignant tumours achieving a sensitivity of 89.7% and specifity of 92.8% and being 74.7% being able to detect tumours less than 2cm. This promising result is really a huge breakthrough given that it was able to detect ones less than 2cm which is a huge step to early diagnosis in this devasting disease.

Overall I believe this project did exceptionally well since they were able to externally validate the data from public datasets from Cancer Imaging Archive and Memorial Sloan Kettering Cancer Centre (Linda C Chu & Elliot K Fishman, 2020). While this is a risky move since using data from other instituitions can encounter technical differences, the DL model was able to detect differences in imaging features rather than technical. To minimise risk however I believe gathering data from one source is the best way to move forward since the downfall of IBM Watson Health came from too many external sources (humans) feeding different data to the model (Casey Ross, 2017). I also believe perchance this may have improved this projects accuracy slightly.

2.3 Critical review of existing studies and developments

Study conducted by:	Po-Ting Chen & Others, National Taiwan University	Ke-Si & Others Zhejiang University	Jayashree Chakraborty & Others, Memorial	Yoshiki Naito, Kurume University Hospital	Gregory R. Hart & Yale University	My Project
1. More than 1 dataset used	Yes	Yes	Yes	No	Yes	Yes
2. CNN used	Yes	No – ResNet was used	No – Random Forest Algorithm & Support Vector Machine Algorithm	Yes	No - ANN as used	Yes, will also try other ML methods
3. More than 70% accuracy for sensitivity and specifity	Yes, (85.9 % & 95.9%)	N/A	N/A	Yes (93.02% & 97.06%)	Yes (80.7% & 80.7%))	Yes
4. More than 0.7 AUC	Yes (0.96)	Yes (0.871)	Yes (0.81)	Yes (0.9836)	Yes (0.86)	Yes

2.4 Conclusion

After researching the established projects conducted at the following university using a CNN seems like the best path as it gives a higher mark for sensitivity, specificity and AUC which are all markers that show that the model's performance is well. I can conclude this as the projects that did not use a CNN scored slightly lower than the ones that did and it was also challenging for them to get figures for specifity and sensitivity and these markers are highly crucial for the motive of this project as it will be able to tell us whether a patient does or doesn't have pancreatic cancer. I believe better accuracy can be obtained by using more data as I can do a better analysis therefore, I will be aiming to use as many relevant datasets as possible.

2.5 Risk Analysis

Risk	Likelihood	Severity	Preventative Actions
Weak time management	Medium	Very high	Start early and have realistic daily goals to ensure work is completed and also chart up a realistic plan for completion
Illness	Low	Medium	Ensure to work out, eat healthy, take regular breaks and make sure I manage to find sometime for hobbies
Unexpected additional work	Medium	High	Review my study plan every 2 weeks and schedule in time for the additional work
Unable to acquire relevant datasets	Low	High	Evaluate each dataset to check if it is appropriate
Difficult of learning new content	Low	Very high	Enrol into good-quality courses

References

https://news.mit.edu/2022/machine-learning-edge-microcontroller-1004

http://web.mit.edu/spotlight/learning-edge/

https://www.allaboutcircuits.com/podcast/

https://www.allaboutcircuits.com/technical-articles/what-is-tinyml/

https://www.tinyml.org/

https://tinyml.mit.edu/

https://medium.com/tech-cult-heartbeat/what-is-tinyml-and-why-does-it-matter-f5b164766876

https://www.artiba.org/blog/tinyml-the-future-of-machine-learning

https://pancan.org/facing-pancreatic-cancer/about-pancreatic-cancer/survival-rate/

https://www.cancer.net/cancer-types/pancreatic-cancer/statistics#:~:text=The%20general%205%2Dyear%20survival,disease%20when%20it%20is%20diagnosed.

https://www.nature.com/articles/s41598-021-87748-0#Sec5

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6374001/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050835/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778580/#B7

https://en.wikipedia.org/wiki/Convolutional_neural_network

https://pubs.rsna.org/doi/full/10.1148/radiol.220152

https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30105-9/fulltext

https://www.statisticshowto.com/mcnemar-test/

https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

https://machinelearningmastery.com/difference-test-validation-datasets/

https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets

https://medium.com/@sadievrenseker/preprocessing-end-to-end-data-preprocessing-1b0672087977

https://www.mobihealthnews.com/news/google-pulls-plug-streams-amid-larger-google-health-shutdown

https://techcrunch.com/2021/08/26/google-confirms-its-pulling-the-plug-on-streams-its-uk-clinician-support-app/

https://www.deepmind.com/blog/using-ai-to-give-doctors-a-48-hour-head-start-on-life-threatening-illness

https://www.nature.com/articles/s41586-019-1390-1

https://ai.googleblog.com/2018/05/deep-learning-for-electronic-health.html

https://artificialintelligence.oodles.io/blogs/healthcare-chatbot-development-with-dialogflow/

https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health

https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html#:~:text=A%20doctor%20involved%20said%20that,was%20later%20audited%20and%20shelved.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/#CIT0009

https://scholar.google.com/scholar_lookup?title=Rule-based+expert+systems:+The+MYCIN+experiments+of+the+Stanford+heuristic+programming+%C2%ADproject&author=BG+Buchanan&author=EH+Shortliffe&publication_year=1984&

https://www.ibm.com/uk-en/watson-health#2887129

https://www.ibm.com/uk-en/watson-health/solutions/healthcare-provider

https://www.ibm.com/industries/healthcare/services

https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html

https://global.hitachi-solutions.com/blog/nlp-in-healthcare/

https://marutitech.com/use-cases-of-natural-language-processing-in-healthcare/

https://pythonistaplanet.com/pros-and-cons-of-supervised-machine-learning/

https://www.foreseemed.com/blog/machine-learning-in-healthcare

https://www.sciencedirect.com/topics/computer-science/supervised-learning

https://builtin.com/artificial-intelligence/machine-learning-healthcare

https://www.analyticssteps.com/blogs/artificial-intelligence-healthcare-applications-and-threats

https://stackoverflow.com/questions/44594007/using-reinforcement-learning-for-classfication-problems

https://developer.ibm.com/learningpaths/get-started-automated-ai-for-decision-making-api/what-is-automated-ai-for-decision-making/

https://www.ibm.com/cloud/learn/unsupervised-learning#:~:text=Unsupervised%20learning%2C%20also%20known%20as,the%20need%20for%20human%20intervention.

https://www.coursera.org/articles/types-of-machine-learning

https://en.wikipedia.org/wiki/Machine_learning

https://www.eetasia.com/efinix-fpgas-integrate-tinyml-platform-for-ai-acceleration/

https://www.hackster.io/news/the-minun-tinyml-framework-squeezes-machine-learning-models-onto-resource-light-microcontrollers-07d581406e4c

https://www.diagnosticimaging.com/view/could-a-new-deep-learning-tool-enhance-ct-detection-of-pancreatic-cancer-

https://www.lshtm.ac.uk/newsevents/news/2021/machine-learning-brings-early-diagnostic-pancreatic-cancer-step-closer-reality

https://www.diagnosticimaging.com/view/new-computed-tomography-study-shows-high-20-year-survival-rates-for-early-stage-lung-cancer

https://www.diagnosticimaging.com/view/could-a-new-deep-learning-tool-enhance-ct-detection-of-pancreatic-cancer-

https://www.frontiersin.org/articles/10.3389/fonc.2022.973652/full

https://www.frontiersin.org/articles/10.3389/frai.2019.00002/full

https://www.sciencedaily.com/releases/2022/09/220913110443.htm

https://journals.lww.com/pancreasjournal/fulltext/2021/03000/artificial_intelligence_and_early_detection_of.1.aspx

https://inside-machinelearning.com/en/encoder-decoder-what-and-why-simple-explanation/

https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

https://en.wikipedia.org/wiki/Support_vector_machine

https://www.sciencedirect.com/science/article/abs/pii/S1570870516301718

https://edrn.nci.nih.gov/data-and-resources/data/

https://www.ibm.com/cloud/learn/convolutional-neural-networks

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/#:~:text=The%20DBN%20is%20a%20multilayer,abstract%20representations%20of%20this%20input.

https://datagen.tech/guides/image-annotation/image-labeling/

https://towardsdatascience.com/e2e-the-every-purpose-ml-method-5d4f20dafee4

https://www.scientificamerican.com/article/pancreatic-cancer-type-jobs/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4924574/#:~:text=Jobs%20was%20diagnosed%20with%20a,often%20rapidly%20fatal%20pancreatic%20adenocarcinoma.

https://www.rsna.org/news/2022/september/Pancreatic-Cancer-Detection#:~:text=Researchers%20in%20Taiwan%20have%20been,pancreatic%20cancer%20from%20noncancerous%20pancreas.

https://www.inverse.com/innovation/pancreatic-cancer-ai-diagnosis

https://pancan.org/news/5-things-know-brca-mutations-pancreatic-cancer/

https://cdas.cancer.gov/datasets/plco/10/#:~:text=The%20Pancreas%20dataset%20is%20a,participants%20in%20the%20PLCO%20trial.

https://arup.utah.edu/database/BRCA/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8975122/

https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

https://www.nih.gov/news-events/news-releases/brca-exchange-aggregates-data-thousands-brca-variants-inform-understanding-cancer-risk

https://fifarma.org/en/this-is-the-cancer-gene-that-all-humans-have/#:~:text=Each%20person%20has%20two%20copies,cells%20to%20not%20function%20properly.

https://www.pancreaticcancer.org.uk/research-projects/early-detection-innovation-projects/

https://www.zendesk.co.uk/blog/machine-learning-and-deep-learning/

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251876#sec005

https://highdemandskills.com/tinyml/#h2-1

https://pll.harvard.edu/course/future-ml-tiny-and-bright?delta=0

https://research.aimultiple.com/tinyml/

https://www.oreilly.com/library/view/tinyml/9781492052036/ch04.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4624268/

https://www.kaggle.com/datasets/samdemharter/brca-multiomics-tcga

https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-16

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301740/

https://link.springer.com/article/10.1007/s10586-022-03707-y#:~:text=Structured%20data%20algorithms%20include%20Artificial,that%20ordinary%20DNN%20cannot%20solve.

https://www.researchgate.net/publication/333134584_A_survey_on_various_machine_learning_algorithms_for_disease_prediction