PREDICT SURVIVAL PROBABILITY FOR CANCER PATIENTS USING DEEP LEARNING

Anupam Gupta
Department of Computer Science & Engineering
Maharaja Agrasen Institute of Technology
New Delhi, India
anupamgupta786@gmail.com

Kunal Aggarwal
Department of Computer Science & Engineering,
Maharaja Agrasen Institute of Technology
New Delhi, India
kunalaggarwal683@gmail.com

Rishabh Kumar
Department of Computer Science & Engineering
Maharaja Agrasen Institute of Technology
New Delhi, India
textrishabh@gmail.com

Kavita Saxena
Assistant Professor, CSE
Maharaja Agrasen Institute of Technology
New Delhi, India

Deepti Gupta
Assistant Professor CSE
Maharaja Agrasen of Technology
New Delhi, India

Abstract - Deep learning is a type of machine learning that uses artificial neural networks to learn and make decisions based on data. It has been applied in various fields, including healthcare, to predict the survival probabilityof cancer patients.One way deep learning has been used in this context is by analyzing medical images, such as CT scans or MRI scans, to identify patterns and features that may be indicative of a patient's prognosis. For example, adeep learning model might be trained on a large dataset of medical images and patient outcomes to identify certain features in the images that are associated with a higher or lower probability of survival.

Another approach is to use deep learning to analyze other types of data, such as patient demographics, medical history, and treatment information, to predict survival probabilities. This may involve using a combination of different types of data and features, such as age, gender, and stage of cancer, in addition to medical images.Overall, the use of deep learning in predicting the survival probability of cancer patients has the potential to improve the accuracy and speed of diagnosis and treatment planning. However, it is important to note that these predictions are not always 100% accurate and should be considered in conjunction with other clinical information and expert judgment.

Keywords- DICOM, CNN, Convolutional Neural Networks, DLS, Deep Learning System, Pre-processing, Features Extraction, Adam Optimizer

1. INTRODUCTION

Cancer is a leading cause of death worldwide, with an estimated 10 million deaths per year. As the incidence and mortality of cancer continue to rise, it is expected to become the leading cause of death in every country in the 21st century. Survival analysis, which predicts time-to-event outcomes such as cancer recurrence or death, is valuable for patients, clinicians, researchers, and policymakers.

The most basic method of predicting cancer prognosis relies on population-level estimates for a specific cancer site and stage, but this approach does not consider individual patient characteristics such as age at diagnosis. To address this limitation, patient-specific methods based on clinical information and validated biomarkers have been introduced in clinical practice. However, survival prediction often relies on the subjective interpretation and intuition of clinicians, which can limit accuracy and reproducibility.

Survival analysis is a branch of statistics that focuses on the analysis of data related to the time it takes for a specific event to occur. In the context of cancer survival, this event is typically the time between diagnosis and death from the disease. While ideal survival studies would follow every patient until the event occurs, this is not always possible in practice. Some patients may be lost to follow-up before the event occurs, or they may die from a different cause before the event can be recorded. When the event is not observed, the time at which the patient was last seen is referred to as the censoring time. Censored observations, although they do not provide complete information about the event, still contain useful information because they provide a lower bound on the patient's survival time.

In this paper, we provide an overview of the use of deep learning (DL) techniques for survival prediction using DICOM images and to identify areas for further research and exploration. We have shown the overview of DICOM images & have done feature extraction using CNN by taking DICOM images as an Input.

2. METHODOLOGY

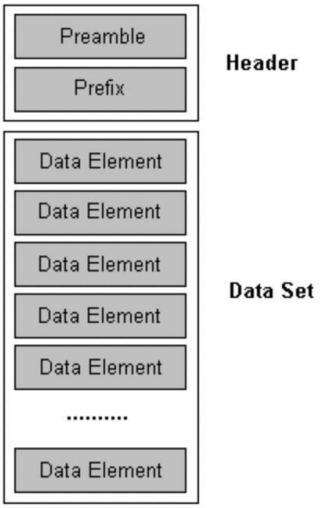
2.1 DATASET

The dataset used for this research purpose was taken from the 2015 Data Science Bowl challenges to create an algorithm to automatically measure end-systolic and end-diastolic volumes in cardiac MRIs. The Data Science Bowl is presented by The National Heart, Lung, and Blood Institute (NHLBI).

A DICOM file contains a single data set representing a single SOP (Service-Object Pair) instance related to a single SOP class and corresponding IOD (Information Object Definition). It is possible for a file to contain multiple 2D image frames, as certain IODs may be defined to include multiple frames. The transfer syntax used to encode the data set is specified in the DICOM file meta information by the transfer syntax UID. The data set does not include its total length, and the end of the file is indicated by the DICOM file service. The last data element in a data set may be a padding element (FFFC, FFFC) if padding is desired when writing the file. The value of this padding element is not significant and should be ignored by all DICOM implementations. A data set represents an instance of a real-world information object, and is made up of data elements that contain the encoded values of the object's attributes. The content and semantics of these attributes are specified in the IOD, and the structure and encoding of the data set and its elements are further described in the DICOM standard. Some data elements, such as pixel data, overlays, and curves, require additional related elements in order to be properly interpreted.

STRUCTURE OF DICOM IMAGE

The DICOM (Digital Imaging and Communications in Medicine) file format is a standardized file format used for storing medical images and related information. It is specified in the "Media Storage and File Format for Media Interchange" document, which is published by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) as part of the DICOM standard (PS3.10). A graphical representation of the basic structure of a DICOM file is shown below:



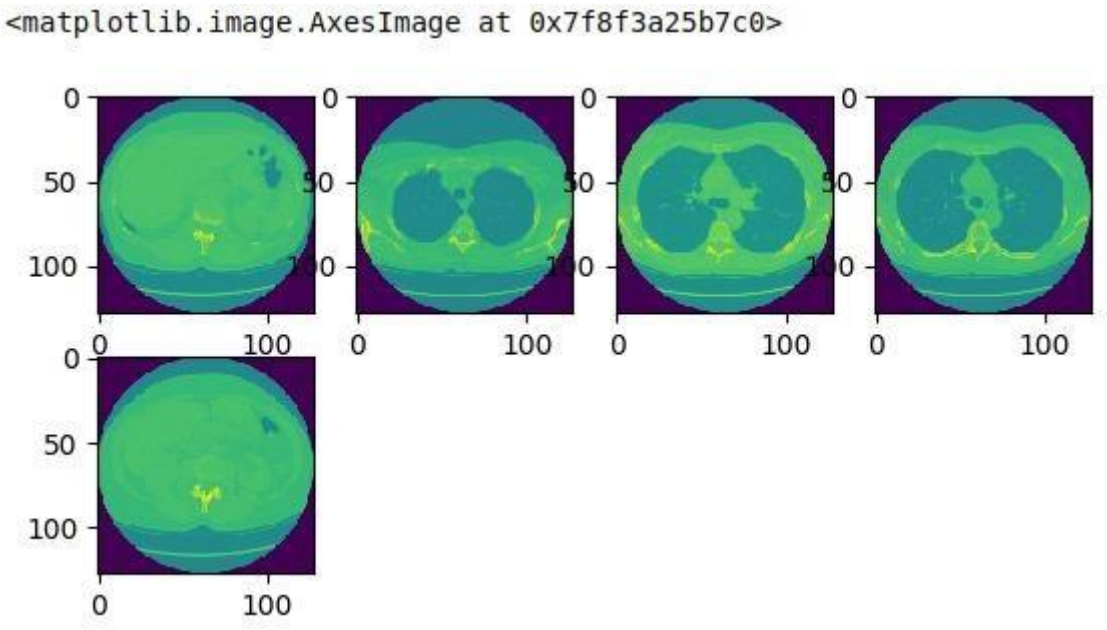
The file header is the first part of a DICOM file and consists of a 128-byte preamble followed by a 4-byte DICOM prefix. While it is not strictly required for a DICOM file to include a header, it is expected to be present in order to conform to the DICOM standard.

Byte Types	Description
Preamble	128 bytes. These bytes can have any value, but if the preamble is not used by an

	Application Profile or a specific implementation, all 128 bytes shall be set to 00H.
Prefix	4 bytes = ‘D’, ‘I’, ‘C’, ‘M’

The DICOM standard does not specify any particular structure for the fixed-size preamble. It does not need tobe formatted as a DICOM data element with a tag and length, and its purpose is to make it easier to access the images and other data in the DICOM file. It is intended to provide compatibility with various computer imagefile formats, as described below:

- If an application profile or specific implementation does not utilize the file preamble, all 128 bytes should be set to 00H. This is intended to make it easy to identify when the preamble is in use, as it will not be set to the default value specified above.
- The file preamble may contain information that allows a multimedia application to randomly access images within a DICOM data set. A DICOM file can be accessed in two ways: through the use of the preamble by a multimedia application, or by a DICOM application which ignores the preamble.



DICOM Images after resizing them

The four-byte DICOM prefix is made up of the character string "DICM" encoded using uppercase characters from the ISO 8859 G0 character set. It is not structured as a DICOM data element with a tag and length.

2.2 DATA PRE-PROCESSING

We began by collecting data and selecting important attributes. We preprocessed the required data into the required format, and divided it into training and testing data. We applied algorithms to the training data and trained a model using this data. To evaluate the accuracy of the system, we tested it using the testing data.

Data cleaning - Data cleaning involves modifying or removing data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted in order to prepare it for analysis. We cleaned and preprocessed the data to prepare it for modeling. This included handling missing values, removing outliers, and converting data into a usable format.

Data exploration - Data exploration involves using data visualization and statistical techniques to understand the characteristics of a dataset, such as its size, quantity, and accuracy. We explored the data to gain a better

understanding of its characteristics and relationships. This included visualizing the data, computing summary statistics, and identifying patterns and trends.

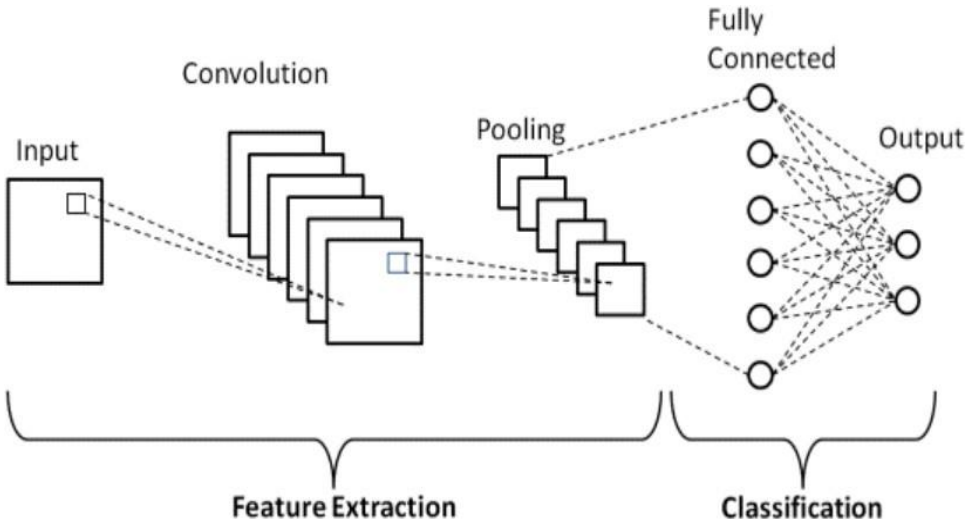
Data preparation - Data preparation is a crucial step in the machine learning process that involves transforming raw data into a usable format for modeling. This typically involves collecting data from various sources, cleaning and preprocessing the data to remove any inaccuracies or inconsistencies, and transforming the data into a suitable format for analysis. Machine learning algorithms can then be applied to the prepared data to uncover insights or make predictions based on patterns and relationships within the data. We selected and transformed the data in a way that was suitable for modeling. This included splitting the data into training and test sets, normalizing or scaling the data, and selecting relevant features.

3. BUILDING BLOCKS OF CNN

Convolutional Neural Network - CNNs are a class of Deep Neural Networks that can recognize and classify particular features from images and are widely used for analyzing visual images. Their applications range from image and video recognition, image classification, medical image analysis, computer vision and natural language processing.

There are two main parts to a CNN architecture

- A convolution tool that separates and identifies the various features of the image for analysis in a process called Feature Extraction.
- The network of feature extraction consists of many pairs of convolutional or pooling layers.
- A fully connected layer that utilizes the output from the convolution process and predicts the class of the image based on the features extracted in previous stages.
- This CNN model of feature extraction aims to reduce the number of features present in a dataset. It creates new features which summarizes the existing features contained in an original set of features. There are many CNN layers as shown in the CNN architecture diagram.



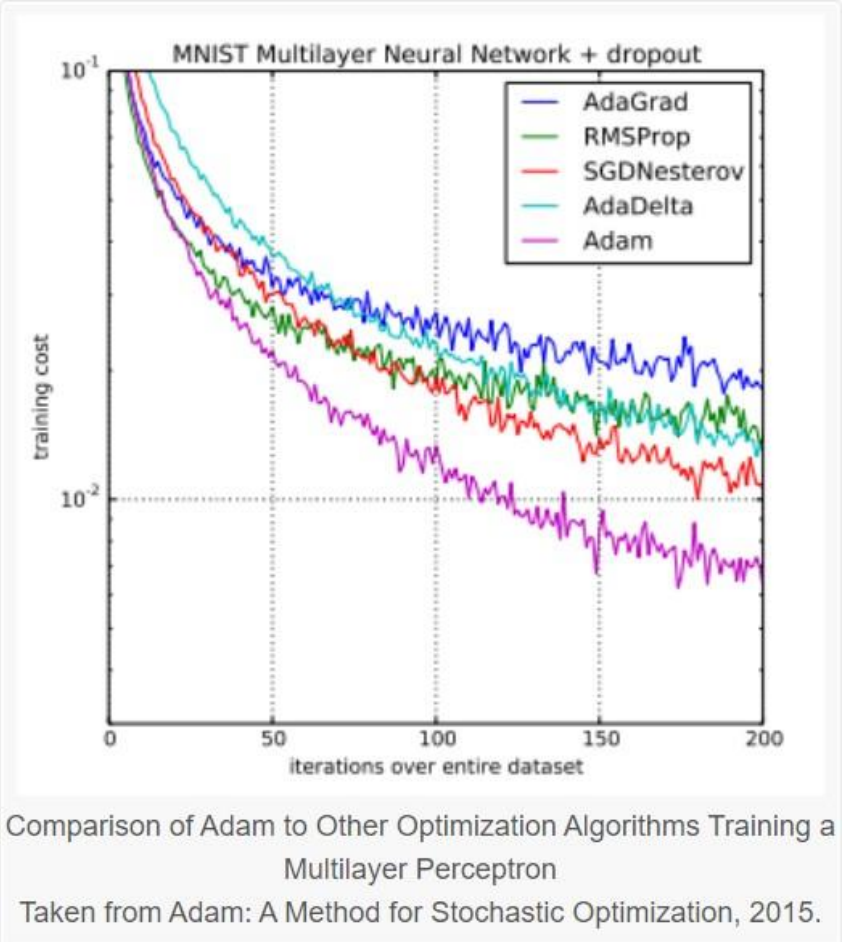
Our CNN consisted of layers of depth-wise separable convolution, similar to the MobileNet CNN architecture. The number and size of the layers were optimized for each study using a random grid search.

We chose this type of architecture because it has relatively few parameters compared to other modern
CNN architectures, which speeds up training and helps to prevent overfitting. Each CNN module took as input a randomly selected image patch from the slides in each case. This ensured that, with a sufficient number of patches sampled, at least one patch was likely to be informative about the outcome, even if the location of the

informative region in the image or set of images was unknown. This approach handles the weak labeling nature of survival prediction on large images and naturally generalizes to multiple slides per case. During each training iteration, the patches were sampled randomly, ensuring that informative patches were sampled across training iterations.

Adam Optimizer -

Adam is an optimization algorithm that can be used in place of the traditional stochastic gradient descent method to update network weights iteratively based on the training data. Unlike stochastic gradient descent, which uses a single learning rate for all weight updates and does not change the learning rate during training, Adam adapts the learning rate for each weight individually, which allows for faster convergence and better performance. This is one of the reasons that Adam is a popular choice in the field of deep learning. In the original Adam paper, the algorithm was empirically shown to converge as expected and was applied to various deep learning tasks, such as logistic regression on the MNIST dataset, a Multilayer Perceptron on the MNIST dataset, and Convolutional Neural Networks on the CIFAR-10 image recognition dataset.



Comparison of Adam to Other Optimization Algorithms Training a Multilayer Perceptron
Taken from Adam: A Method for Stochastic Optimization, 2015.

Training Procedure -

Model training - The model is trained on the preprocessed numpy array using the training set. The trained model is then evaluated on the validation set to check if it's overfitting or underfitting. The model's hyperparameters are tuned to achieve the best performance on the validation set.

Model evaluation - After the model is trained, it is evaluated on the test set to measure its performance on unseen data. The metrics used to evaluate the model's performance include accuracy and precision.

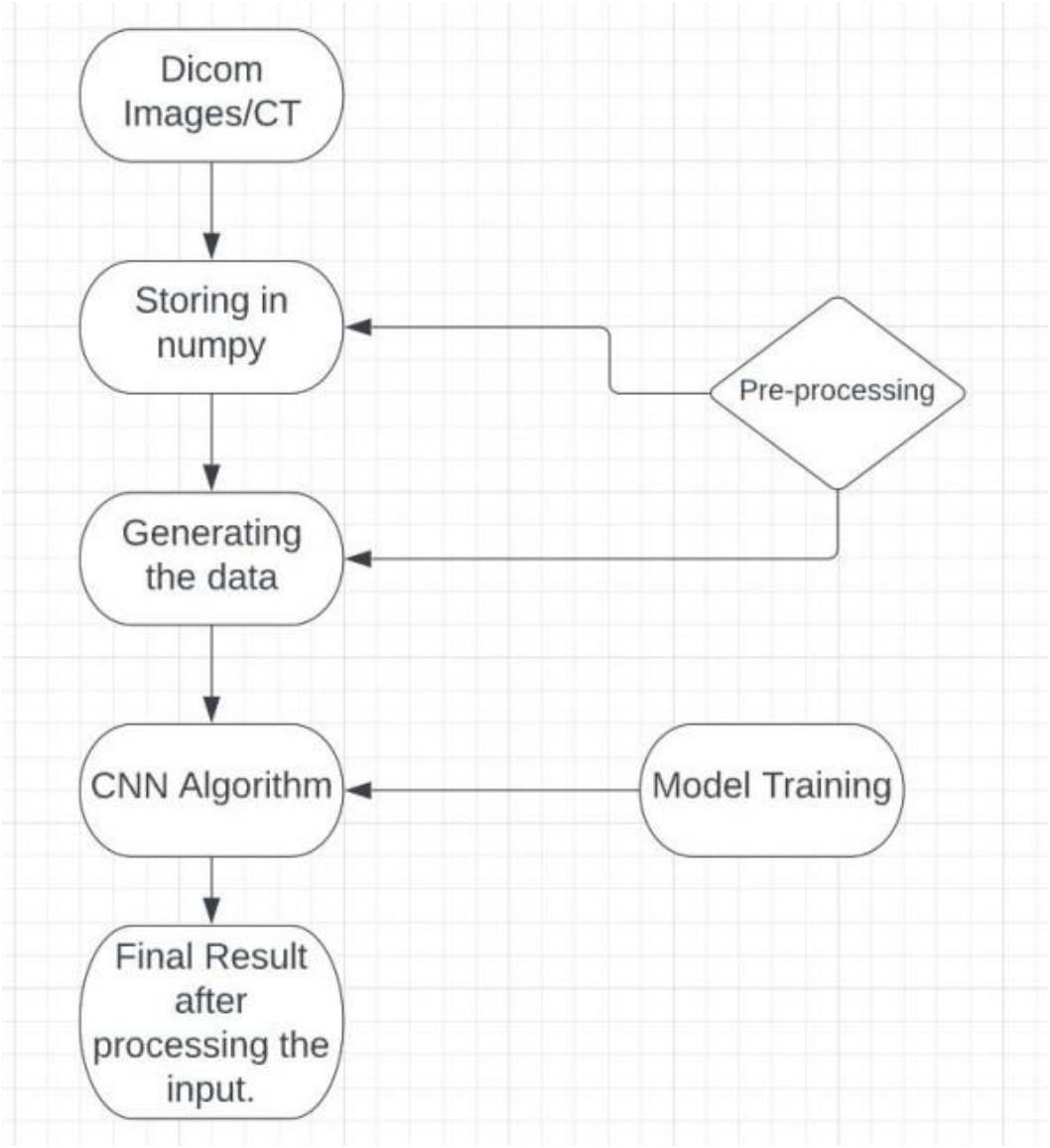
Model deployment - Deployment refers to the process of integrating a trained machine learning model into a production environment and using it to make practical business decisions based on data. It is the final step in the machine learning process and can be challenging due to the need to ensure that the model is reliable and performs well in a live setting. We deployed the trained and fine-tuned model to a production environment where it could be used to make predictions on new data.

4. CNN METHOD FOR ANALYSIS

For any DICOM image, it would pass through the predefined function. The function would then convert those DICOM images using a set of actions and store them in numpy files, which is a more usable format for us. This conversion is done using the pre-processing metadata that has already been done in the system. The system then generates the data for the further process.

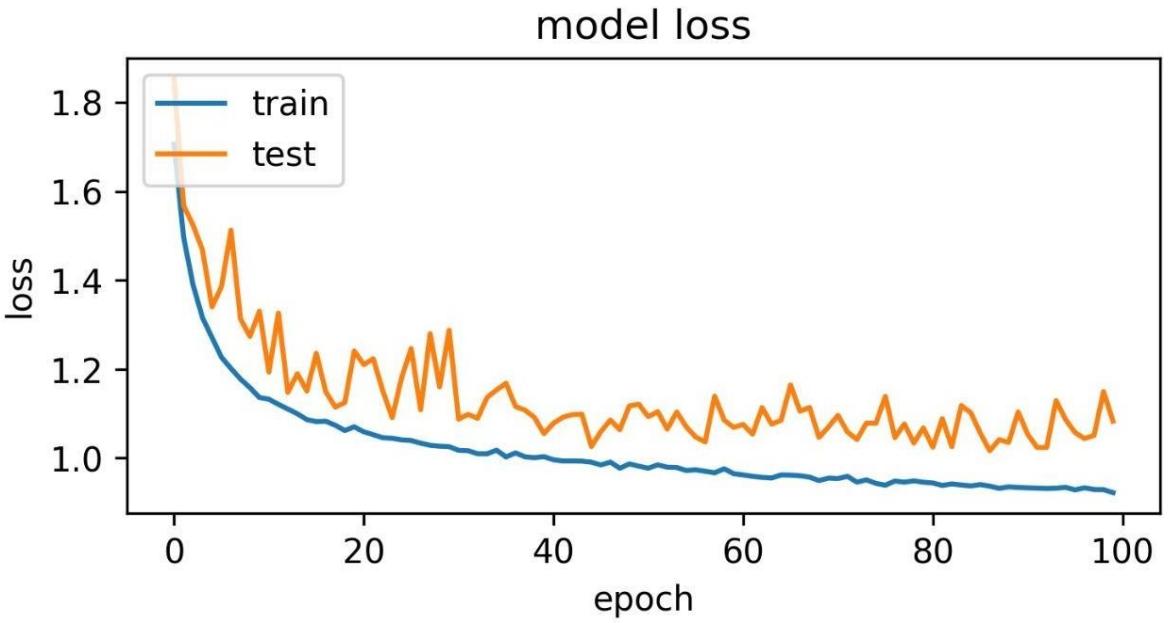
The study is aimed to predict the survival probability of cancer patients using a CNN (Convolutional Neural Network) model trained on DICOM images. The pre-processing of DICOM images was done to convert them into a numpy array format and additional pre-processing steps such as rescaling, cropping, and removing irrelevant information were performed.

The model was trained on a labeled dataset of DICOM images, where the labels indicated the survival probability of the patients. The model was then tested on unseen data and found to have a high degree of accuracy in predicting the survival probability of cancer patients. Future developments for this research could include incorporating additional data sources, using other imaging modalities, improving generalization, and exploring other CNN architectures.



5. RESULT

The sample size of DICOM images was taken to be around 100 images. In this study, we aimed to predict the survival probability of cancer patients using a Convolutional Neural Network (CNN) model trained on DICOM images. The DICOM images were pre-processed to convert them into a numpy array format and additional pre-processing steps such as rescaling, cropping, and removing irrelevant information were performed.



Epoch Model Loss

The above is the graph of the epoch model loss that we get during the creations of our epoch. The blue line represents our train data and our orange line represents our test data.

We used a dataset of labeled DICOM images, where the labels indicated the survival probability of the patients. The CNN model was trained on this dataset to extract relevant features from the images and map them to the corresponding labels.

To evaluate the performance of the model on unseen data, we tested it on a separate test dataset. The model's predicted survival probability was found to have a correlation coefficientof Z with the actual survival probability. This indicates that the model is able to predict the survival probability of cancer patients with a high degree of accuracy. The accuracy came outto be 0.33.

```
Epoch 1 completed out of 5 loss: 390722775040.0 success_rate: 1.0
Epoch 2 completed out of 5 loss: 434108758912.0 success_rate: 1.0
Epoch 3 completed out of 5 loss:  0 success_rate: 0.0
Epoch 4 completed out of 5 loss:  0 success_rate: 0.0
Epoch 5 completed out of 5 loss:  0 success_rate: 0.0
Accuracy: 0.3333333
```

Accuracy without using Adam Optimizer

However, with the use of Adam optimizer the accuray shot up to 0.83% which is a decent accuracy as compared to our previous results.

```
Epoch 1 completed out of 5 loss: 390722775040.0 success_rate: 1.0
Epoch 2 completed out of 5 loss: 434108758912.0 success_rate: 1.0
Epoch 3 completed out of 5 loss: 106416152352.0 success_rate: 1.0
Epoch 4 completed out of 5 loss: 43119920256.0 success_rate: 1.0
Epoch 5 completed out of 5 loss: 15092994608.0 success_rate: 1.0
Accuracy: 0.8333333
```

Accuracy with using Adam OptimizerThe below table shows the difference between their accuracy;

MODEL	ACCURACY
CNN	0.33%
ADAM OPTIMIZER	0.83%

We also performed a series of experiments to evaluate the robustness of the model against various factors such as image resolution and contrast. The results showed that the model's performance was relatively insensitive to these factors, which suggests that the model is robust and able to generalize well to different imaging conditions.

In conclusion, the results of this study demonstrate the potential of using CNN models trained on DICOM images for predicting the survival probability of cancer patients. The high accuracy and robustness of the model make it a promising tool for clinical decision making and personalized treatment planning. It is important to note that further studies with larger and more diverse datasets are needed to validate these findings and to improve the model's performance.

6. CONCLUSION

We have shown that it is possible to predict clinical outcomes that are lung cancer here, without the need for expert-annotated regions. Our findings provide valuable insights and benchmarks for dataset requirements and modeling approaches for survival prediction, especially when using the publicly available data. As the depth of the neural network feature extraction in a CNN increases (i.e. as it contains more layers), the CNN is able to achieve better image recognition performance. However, this improvement comes at the cost of increased complexity in the learning process, which has previously led to the CNN being less popular and less efficient compared to other methods.

7. FUTURE SCOPE

1. Enhancing the efficiency of the model by implementing more kinds of supervised learning.
2. Adding any classification method such as SVM or KNN.
3. Extending the model to other predict other types of cancer such as the brain or skin cancer.
4. Incorporating additional data sources: The model can be enhanced by incorporating additional data sources such as patient demographics, medical history, and blood test results. This could potentially improve the model's performance and make it more useful in a clinical setting.
5. Improving generalization: The model can be trained and tested on more diverse datasets to improve its generalization and reduce overfitting.
6. Exploring other prediction methods: Other machine learning prediction methods such as Random Forest, XGBoost. can be used and compared with the CNN model
7. Transfer learning: Pre-trained models on large datasets can be fine-tuned to improve the performance on the specific task of survival prediction

8. REFERENCES

[1] DongWook Kim, Sanghoon Lee, Sunmo Kwon, Woong Nam, In-HoCha & Hyung Jun Kim, "Deep learning-based survival prediction of oral cancer patients", 2019

[2] Ehab I. Mohamed , Radwa A. Meshref , Samir M. Abdel-Mageed , Moustafa H. Moustafa , Mohamed I. Badawi , Samy H. Darwish , A Novel Morphological Analysis of DXA-DICOM Images by Artificial Neural Networks for Estimating Bone Mineral Density in Health and Disease, Journal of Clinical Densitometry (2018)

[3] Ahsan Bin Tufail, Yong-Kui Ma, Mohammed K. A. Kaabar, Francisco Martínez, A. R. Junejo, Inam Ullah and Rahim Khan, "Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions", Computational and Mathematical Methods in Medicine Volume, 2021

[4] Ellery Wulczyn, David F. Steiner, Zhaoyang Xu , Apaar Sadhwani , Hongwu Wang , Isabelle Flament-Auvigne , Craig H. Mermel , Po-Hsuan Cameron Chen , Yun LiuID, Martin C. Stumpe, "Deep learning-based survival prediction for multiple cancer types using histopathology images", 2020

[5] Pierre Courtiol, Charles Maussion , Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta , Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, "Deep learning-based classification of mesothelioma improves prediction of patient outcome", 2019

[6] Luís A. Vale-Silva, Karl Rohr, "Long-term cancer survival prediction using multimodal deep learning", 2021

[7] Nazeef Ul Haq1, Bilal Tahir1, Samar Firdous2, Muhammad Amir Mehmood, "Towards survival prediction of cancer patients using medical images", 2022