Novel *Block* Diagonalization for Reducing Features and Computations in Medical Diagnosis

Tahira Ghani[1] and B. John Oommen[2] ⋆

School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6.
[1]tahira.ghani@carleton.ca,[2]oommen@scs.carleton.ca

Abstract. Diagonalization is an "age old" technique from Linear Algebra, and it has had significant applications in Pattern Recognition (PR) and data pre-processing. By using the eigenvectors of the covariance matrix of a single class as the basis vectors describing the feature space, the transformed data can be rendered to have a diagonal covariance matrix. If the covariance matrices of two classes are utilized, the covariance matrix of transformed data of the first class can be the made the Identity, while that of the second can be diagonal, implying independence in the case of Normally distributed data[1]. In all of the cases reported in the literature, the entire covarince matrix is diagonalized, which is, computationally, a very tedious and cumbersome process. In this paper, we propose a radically different paradigm where we opt to render the transformed data to be *block* diagonalized. In other words, the covariance of the transformed data is made up of a predetermined number of block matrices, implying that *these* corresponding features are assumed to be correlated, while the others are assumed independent. Regression is now done by getting the best value based on each of these sub-blocks and averaging between them. This is essentially an ensemble machine, where the sub-blocks lead to their own respective regression values, which are then averaged to obtain the overall solution. This technique has been used to analyze the survival rate of cancer patients depends on the type of cancer, the treatments that the patient has undergone, and the severity of the cancer when the treatment was initiated. In our *prima facie* study, we consider adenocarcinoma, a type of lung cancer detected in chest Computed Tomography (CT) scans on the entire lung, and images that are "sliced" versions of the scans as one progresses along the thoracic region. The results that we have obtained using such a *block* diagonalization are quite amazing. Indeed, they surpass the results obtained from some of the well-established feature selection/reduction strategies.

Keywords : *Diagonalization, Block Diagonalization, Medical Image Processing, Lung Cancer Treatment, Prediction of Survival Rates*

⋆ *Chancellor's Professor*; *Life Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway.

[1] Similar diagonalizing schemes form the basis of the Principal Component Analysis (PCA) and some feature selection/reduction etc. schemes.

1 Introduction

There numerous ways by which one can reduce the dimensionality of a problem when it concerns Pattern Recognition (PR) and classification. The question of determining which features contain the most discriminating power is not easy to resolve. The strategy that has been used for a few decades is to examine the data by considering the eigenvalues and eigenvectors of the feature space, and by projecting the data onto its most "prominent" eigenvectors. Similar principles are, typically, used in designing feature selection and feature reduction methods. Indeed, almost all of the acclaimed methods deal with the eigenvectors of the covariance matrices of the classes. Since this matrix is symmetric and positive-definite, the problem is less tedious, because the eigenvectors are orthogonal.

The problem, although easily stated, is rather complex because one has to compute the eigenvectors of the covariance matrix of the entire feature space. This is, usually, a very difficult problem, especially if the dimension of this space is large. In this paper we resolve the problem in a completely different manner. The paradigm that we advocate is the following: Rather than work with the entire set of features, we split them into small blocks, which are then *individually* diagonalized. The tacit assumption of invoking such a partitioning is that within each of these small blocks, the intra-block features are correlated, but also that the inter-block features are uncorrelated. Resolving the problem in this manner leads to a block diagonal matrix. Observe that with such a modelling philosophy, the PR-related processing is much easier because the inversion of these block diagonal matrices involves block diagonal operations, and to achieve this, we only have to work with matrices of much smallers sizes.

The specifics of the partitioning, the block diagonalization and the subsequent regression, are detailed in the body of the paper. As far as we know, such a block diagonalization paradigm has not been proposed or applied in the PR or regression literature. In particular, we have used these techniques in a regression analysis, by which we can predict the survival times of lung cancer patients based on various features of the tumor. The results that we have obtained surpass the results obtain by invoking the best-known feature selection and reduction techniques.

1.1 Contributions of this Paper

The contributions of this paper can be summarized as follows:

- The fundamental contribution of this paper is to demonstrate the advantages of utilizing a "block diagonalization" paradigm, instead of invoking a diagonalization process of the entire feature space;
- Rather than achieve a regression analysis based on the entire set of features, we have shown that we can perform such an analysis on the various subsets of the data (each obtained from the block diagonalized submatrices);
- The overall regression will then be obtained by combining the results of the regression analysis of the blocks. Observe that this is equivalent to merging the concept of "ensemble" machines with the latter "block diagonalization" phase;

- Although a lot of work has been done when it concerns the diagnosis of lung cancer, the work related to the survival times and their correlation to the size/shape of the tumor is relatively unexplored. In an earlier paper [10], we had shown that by a regression analysis, we can predict the survival times based on various features of the tumor. This paper builds on those results to use ensemble machines and block-diagonal phenomena.
- While these results have been proven to be relevant for our lung cancer scenario, we believe that these phenomena are also valid for other tumor-based cancers, and hope that other researchers can investigate the relevance of the same hypothesis for *their* application domains.

2 Diagonalization and Block Diagonalization

2.1 Diagonalization

To initiate discussions, we submit a few brief paragraphs about the phenomenon of diagonalization. A square matrix A is referred to as being "diagonalizable"if it is similar to a diagonal matrix. In other words, there exists an invertible matrix P and a diagonal matrix D, such that $P^{-1}AP = D$, which equivalently implies that $A = PDP^{-1}$. Diagonalization is the process of finding the above P and D.

Diagonalizable matrices are especially easy for computations. One can raise a diagonal matrix D to a power by simply raising the diagonal entries to that power, and the determinant of a diagonal matrix is simply the product of all its diagonal entries. Such computations generalize easily due to $A = PDP^{-1}$.

The process of diagonalization is implicitly related to the set of A's eigenvectors $\{\underline{e}_1, \underline{e}_2, \ldots \underline{e}_d\}$, and its respective eigenvalues, $\{\lambda_1, \lambda_2, \ldots \lambda_d\}$. Indeed, if A can be diagonalized, the diagonalizing matrix P contains the eigenvectors of A as columns, where the i^{th} column is the i^{th} eigenvector, $\underline{e}_i$, and the corresponding diagonal entry is the respective eigenvalue, λ_i. The invertibility of P also suggests that the eigenvectors of A are linearly independent and form a basis of A, which is the necessary and sufficient condition for diagonalizability. Thus,

$$P^{-1}AP = \Lambda, \quad \text{or} \quad AP = P\Lambda, \tag{1}$$

where, $P = [\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_d]$, and Λ is the $d \times d$ diagonal matrix, $\text{Diag}[\lambda_1, \lambda_2, \ldots, \lambda_d]$.

Diagonalization is a fundamental phenomenon in PR and classification, where A is the covariance matrix of the underlying distribution, and since this is always positive-definite and symmetric, the eigenvectors are orthogonal, implying that:

$$P^T AP = \Lambda. \tag{2}$$

Similar diagonalizing schemes form the basis of the Principal Component Analysis (PCA) and some feature selection/reduction etc. schemes.

If the covariance matrices of two classes are utilized, the covariance matrix of transformed data of the first class can be made Identity, while that of the second can be diagonal, implying independence in the case of Normally distributed data. Geometrically, a diagonalizable matrix is an inhomogeneous dilation (or anisotropic scaling) - it scales the space, as does a homogeneous dilation, but by

a different factor along each eigenvector axis, the factor given by the corresponding eigenvalue. For most applications, the matrices are diagonalized numerically using computer software, and numerous packages exist to accomplish this.

2.2 Invoking Block Diagonalization and Ensemble Regression

We shall now show how we can use a non-traditional eigenvector matrix to further enhance the accuracy of the scheme, and also simultaneously minimize the computations.

The monumental task associated with a covariance matrix of large dimensions is to obtain its eigenvalues and eigenvectors. While this is surely advantageous, the task is daunting especially when we deal with a feature space whose dimensionality is greater than 100. It would have been meaningful if we were working with simultaneous diagonalization where we tried to diagonalize or whiten data from *two* classes. In our case, however, since the problem we tackle is regression (rather than classification), we are dealing with only a single class which makes the problem less cumbersome.

In our application domain [9], [10], the dimensionality of the feature vector is 110. Computing the eigenvalues and eigenvectors of such a large matrix is, certainly, time consuming. The novel contribution in this section is that we advocate dividing the feature vector into multiple sub-vectors, for example, 5 sub-vectors. We are now faced with a problem of getting 5 sets of eigenvectors, each of dimension 22, which is a significantly smaller problem. Of course, this leads us to an approximated world in which the features within these subspaces are correlated, but it leads to a model in which the features outside of these blocks are assumed to be uncorrelated. This, in turn, leads to the concept of a block diagonal matrix, as displayed in Figure 1, where the blocks, B_n where $n = 1...5$, represents the subset of features chosen (i.e., 5 sets of 22-dimensional vectors), and all other elements outside of the blocks are set to zero.

$$\begin{bmatrix} \boxed{B_1} & 0 & 0 & 0 & 0 \\ 0 & \boxed{B_2} & 0 & 0 & 0 \\ 0 & 0 & \boxed{B_3} & 0 & 0 \\ 0 & 0 & 0 & \boxed{B_4} & 0 \\ 0 & 0 & 0 & 0 & \boxed{B_5} \end{bmatrix}$$

Fig. 1: Block diagonal matrix.

Our task now, is to diagonalize each of these blocks within the block diagonal approximation, and to choose a subset of their prominent eigen-directions, determined by the corresponding largest eigenvalues. For example, if we extracted the 5 principal eigen-directions in each of these blocks, the 110-vector space would reduce to 5 blocks of 5 features each, i.e., a feature vector of dimension 25.

Regression is now done by getting the best "regressed" value based on each of these sub-blocks and averaging between them. The reader will observe that this is essentially an ensemble machine, where the 5 blocks lead to their own

respective regression values, which are then averaged to obtain the overall solution. Although this involves computations that are significantly less than working with the 110-dimensional space, the results that we have obtained are actually marginally superior. This seems to be paradoxical but the reason for this is probably because the higher dimensional world tries to impose a dependence on the various variables when, in fact, there may not be such an explicit dependence.

3 Predicting Survival Times for Lung Cancer

Problem Domain: Research in computer vision in medicine has advanced to focus on automatic segmentations, feature extraction and classification for the presence of specific diseases or pathologies. CAD systems are divided into two sub-categories, computer-aided detection (CADe) and computer-aided diagnosis (CADx). Both branches are being actively researched, with CADe being more focused on computationally-efficient early detection with a higher sensitivity and low false-positive rate, and CADx being more focused on the lesions' characterization and classification. The most famous use-case for such an application is the detection (or classification) of a nodule being cancerous. However, if we can push this envelope one step further and are able to judge the *severity* of a nodule, the prognosis and determination of treatment plans can be adjusted to yield a greater chance of success. The results we have obtained [10] deal with the cancer treatments done on 60 patients[2] at varying levels of severity and with a spectrum of survival rates. For patients who survived up to 24 months, the average relative error is as low as 9%, which we believe is very significant.

This research is the continuation of previous work of the authors as part of a Masters program thesis study, whereby the foundation is the evaluation of a cancer nodule through computation of 2D features in the Chest Computed Tomography (CT) scan. In this work, we aim to evaluate the cancer nodule as a single entity in its 3D form rather than "slices" as 2D images. Furthermore, we evaluate the feature sets of the 3D cancer nodules through the lens of various feature reduction and feature elimination techniques, with the most extensive analysis on the process of diagonalization.

Aspects of Lung Cancer: Apart from folklore, the statistics about cancer are disheartening. The American Cancer Society (ACS) estimated their annual statistics for 2018, based on collected historical data [17]. Lung cancer is now the second leading type of cancer for newly diagnosed patients, behind breast cancer, and it has the highest mortality rate out of all cancer sites. The ACS projected 234,030 new cases of lung cancer. They also forecasted that 154,050 deaths would be caused by lung cancer. However, cancers diagnosed at an early phase, such as Stage 1, can be treated with surgery and radiation therapy with an 80% success rate. The low survival rate of lung cancer patients is primarily because of the late diagnosis, which results in ineffective treatment due to the growth and stage of the cancer.

[2] Understandably, it is extremely difficult to obtain training and testing data for this problem domain! Thus, both authors gratefully acknowledge the help given by Drs. Thornhill and Inacio, from the University of Ottawa, in providing us with the dataset.

The most common tests to detect lung cancer include (but are not limited to) sputum cytology, chest X-rays, computed tomography (CT) scans, and biopsies. The emerging domain of radiomics is the field of study that extracts quantitative measures of tumor phenotypes from medical images using data-characterization algorithms. These features are explored to uncover disease characteristics that are not visible to the naked eye, but which can then be used for further prognosis and treatment plans. Many researchers have begun focusing on the engineering of feature sets through implementation of radiomic analysis [8], [16]. Our goal, however, is that of predicting the survival rate of lung cancer patients *once they have been diagnosed*, and the result of this study is to demonstrate that a lot of this information resides in the *3D shape* of the tumor. Our hope is that our study can provide insight into the severity of the cancer, and can additionally, aid in formulating the treatment plans so as to increase the chances of survival.

3.1 *Brief* Literature Review

Apart from the mathematical foundations of diagonalization highlighted above, the main objective of the application domain of this research is to explore and investigate the existence of relations between statistical measures with regards to the texture and shape of a nodule classified as adenocarcinoma (a type of cancer) to the *survival time* of the patient post-diagnosis. The task at hand is a regression problem, instead of the more traditional classification problem.

When considering the applications of ML in healthcare, classification problems have been the dominant area of focus such that the presence, or lack thereof, of a specified anatomical structure can be stated. However, transforming the context of the application to a regression domain can enable a critical advancement in CAD systems. By suggesting a survival time for a given patient, the trajectory of the illness and treatment plan can be evaluated at a deeper level. Through a more extensive literature review, discussed in more detail in Chapter 2 of the thesis[3] [9], we have identified a significant gap in such regression analyses.

We aim to engineer a prognostic feature set based on quantitative measurements that are not visible with a simple glance or reading of the scan, and aspire to reduce the inherent subjectivity and variability when evaluating medical reports with such measures. It is important to note that our focus is heavily on the construction of the feature set, rather than the customization of regression models that have been used as testing thresholds. With this in mind, we also adjust the focus to form a valuable feature set through applying various block-diagonalization-based feature elimination and reduction techniques.

3.2 Data Source

We have used the publicly available data from The Cancer Imaging Archive[4] (TCIA), a service which hosts an archive of data for de-identified medical images of cancer. The dataset used for this work is the "LungCT-Diagnosis" data

[3] Unfortunately, due to space limitations, a comprehensive review is not possible. It is found in [9] and can be provided to interested readers if requested.

[4] More information can be found at `https://www.cancerimagingarchive.net/`.

[11] on TCIA, uploaded in 2014. The set consists of CT scans for 61 patients that have been diagnosed with adenocarcinoma, a type of lung cancer, with the number of images totalling up to about $4,600$ over all the scans. However, considering only images that have the presence of a cancer nodule, the count reduces to approximately 450 images. With healthcare, we are constrained to work with what we can obtain and what we are "provided" with, subject to privacy considerations. As we will see, it suffices for the purpose of regression analyses. Throughout the experiments and results explained in this paper, the condensed dataset of 450 images, or 61 scans, has been consistently split into training and testing data with a 70% and 30% split respectively. The dataset also includes the clinical metadata, where the survival time of the patient associated with each scan, is listed.

4 Fundamental Operations and Feature Sets

Computed Tomography (CT) Scans: The most common radiological imaging technique incorporates CT scans where X-ray beams are used to take measurements or images (i.e., "slices") from different angles, as shown in Figure 2, as the patient's body moves through the scanner. Depending on the section thickness and the associated reconstruction parameters, a scan can range anywhere from 100 to over 500 sections or images [1]. The scan records different levels of density and tissues which can be reconstructed to, non-invasively, create a 3-dimensional image of the human body.

(a) Axial Plane **(b)** Coronal Plane **(c)** Sagittal Plane

Fig. 2: Planes captured in a Computed Tomography (CT) scan.

High-resolution Computed Tomography (HRCT) is specifically used in detecting and diagnosing diffuse lung diseases [7] and cancerous nodules, due to its sensitivity and specificity. It enables the detection and analysis of feature aspects such as morphological lesion characterization, nodule size measurement and growth, as well as attenuation characteristics.

Nodule Segmentation: Rather than segmenting the entire lung region as our Region of Interest (ROI), in this research, we segment only the cancerous nodule in the "slices" where the presence of the tumor is observed. Similar to the topic of lung segmentation, there is an abundance of published work discussing the automation of so-called nodule segmentation and extraction [2], [15] and [19]. Since this is not the primary goal of our research, in our work, we opt to segment

the scans manually to obtain the nodules, as our focus is on the creation of an informative feature set.

We made masks of the tumors using the ImageJ software[5]. This was done by manually tracing a contour around the nodule on the images where it was present, filling the shape as "white", and clearing the background to "black". The images that did not contain the nodule were cleared to a "black" background.The CT scans were reviewed, and the segmentation of cancer tumors were validated by a clinical doctor from the Ottawa Heart Institute.

We achieved this by creating a mask for each scan in the dataset. This enabled us to obtain a simpler implementation of the algorithm to achieve the ROI extraction. For each scan, the respective segmentation mask was loaded and all images were iterated. If a connected component was found (signifying the presence of a nodule) in a mask image, the original image was multiplied against the mask to precisely extract only the nodule. In this manner, we were able to attain the 3D matrix of the nodule in a scan, which could be further used for visualization and feature extraction.

3D Feature Set Compilation: Proceeding now with the work that we did, the next stage included the exploration, compilation and modification of a 3D feature set, which essentially implies that we considered the entire lung as a single entity (or observation) as opposed to the images considered for the 2D feature sets alluded to in the previous sections. We used the `Pyradiomics` library[6], an open source Python package, for extracting radiomics[7].

The main emphasis for this feature set was to compute measures in a 3D consideration. Compared to the benchmark feature set, which contained quantification measures of texture analysis in the 2D domain, this feature set included recalculated texture analysis components in a 26-connectivity capacity, i.e., if the centre pixel shares a face, edge or corner with another pixel. The radiomics features extracted using the `Pyradiomics` library (defined in the `Pyradiomics` documentation) are defined in sub-categories known as feature classes as:

- **First Order Statistics:** This feature class [9], [10], focused on computing features that described the histogram of the nodule image, i.e., the grey-level intensity distribution etc., amounting to a total of 19 features.
- **3D Shape-based:** These 16 features [9], [10], unlike the rest of the feature classes, computed the values from the mask of the given nodule as they were independent of the grey level intensities. The `Pyradiomics` library built a triangle mesh from the provided mask using the marching cubes algorithm.
- **Grey Level Cooccurrence Matrix (GLCM):** This matrix was computed based on the probability function where the $(i, j)^{\text{th}}$ element represents the number of times grey levels i and j appeared next to each other in the image, and led to a total of 24 features [9], [10].

[5] The ImageJ Software is a Java-based program developed at the National Institutes of Health and the Laboratory for Optical and Computational Instrumentation.

[6] Documentation is available at `https://pyradiomics.readthedocs.io/en/latest/`.

[7] Radiomics is the field of study that extracts quantitative measures of tumor phenotypes from medical images using data-characterization algorithms. These features are explored to uncover disease characteristics that are not visible to the naked eye.

- **Grey Level Run Length Matrix (GLRLM):** This matrix captured grey level runs, in which, a *run* refers to the number of consecutive pixels that had the same grey level value. The GLRLM was quantified by individual measures such as short run and long run emphasis, non-uniformity values, run variance, etc., amounting to a total of 16 features [9], [10].
- **Grey Level Size Zone Matrix (GLSZM):** This matrix was similar to the GLRLM, however, voxels of the same grey level intensity were taken into consideration rather than just pixels. It was quantified by individual metrics such as area and zone emphasis etc., amounting to 16 features [9], [10].
- **Neighbouring Grey Tone Difference Matrix (NGTDM):** This matrix evaluated the difference between a pixel or voxel's grey value and the values of its neighbours. From it, we extracted to a total of 5 features [9], [10].
- **Grey Level Dependence Matrix (GLDM):** This matrix measured the grey level dependencies in the image. From it, we extracted a total of 14 features, as explained in [9], [10].

5 Implementation and Results

Model Evaluation: To evaluate the performance of the tested regression models, we utilized two measures, namely the Mean Absolute Error (MAE), measured in months, and the Mean Relative Error (MRE), both of which are defined in Eq. (3) and (4) respectively:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - z_i|, \qquad \text{and} \tag{3}$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - z_i|}{z_i}, \qquad \text{where:} \tag{4}$$

- n is the number of test-set data points,
- y_i is the predicted value (i.e., the expected survival time in months), and
- z_i is the true value (i.e., the survival time in months).

The MAE is the average difference between the true values and the predicted values. It provides an overall measure of the distance between the two values, but it does not indicate the direction of the data (i.e., whether the result is an under or over-prediction). Furthermore, this is also seen to be a scale-dependent measure, as the computed values are heavily dependent on the scale of the data, and can be influenced by outliers present in the data [5]. In order to circumvent the scale-dependency, we also computed the MRE which introduces a relativity factor by normalizing the absolute error by the magnitude of the true value. This means that the MRE should, generally, consist of values in the range $[0, 1]$.

As mentioned earlier, all regression tests were done on the data with a 70% to 30% split of the data for training and testing, respectively.

Regression Results: The diagonalization technique applied to the dataset, defined in Eq. (5), is a linear transformation that transforms a set of vectors,

$\{X\}$, with a given covariance matrix to a new set of vectors, $\{Y\}$, with a covariance that is the Identity matrix. This indicates uncorrelated features with a variance of 1 satisfying:

$$Y = W^T X. \tag{5}$$

The transformative factor here is defined by W, which is a $d \times d$ matrix such that:

$$W^T \Sigma_X W = \Lambda, \text{ and} \tag{6}$$

$$\Lambda = \begin{bmatrix} \lambda_1 \ldots 0 \\ \ddots \\ 0 \ \ldots \lambda_d \end{bmatrix}, \tag{7}$$

where $\lambda_1 \ldots \lambda_n$ are the eigenvalues of the covariance matrix. It is important to note that W also satisfies the following condition:

$$W^{-1} = W^T. \tag{8}$$

As mentioned, the context at hand is a regression problem and hence, we are dealing with data from only a *single* class. Therefore, we implemented a simple diagonalization rather than a simultaneous diagonalization of multiple classes.

We computed the covariance matrix of the scaled dataset, resulting in a 110×110 matrix. However, computing the eigenvalues and eigenvectors of the covariance matrix was computationally complex and inefficient, resulting in negligibly-small negative and complex eigenvalues[8]. To combat the inaccuracy of these results, we broke down the 110-dimensional covariance matrix into five 22-dimensional covariance matrices from which we could compute the eigenvalues. This led to the resemblance of an ensemble-based model, as we transformed the data five times with only the k^{th} group of features extracted from the covariance matrix. Running the regression models with selecting the corresponding transformed features of the first d significant eigenvalues (where d is the number of eigenvalues that have values above a threshold of 1.0), we took the average of the five predictions and attained our final regressed prediction.

Table 1: Performance of Regression Models with Complete Baseline and PCA-Reduced Feature Set.

	Baseline		**PCA**	
Model	**MAE**	**MRE**	**MAE**	**MRE**
Linear Regression	34.41	1.95	14.83	0.78
kNN Regression	16.73	0.97	17.15	0.90
Gradient Boosting	14.76	0.85	18.36	1.00

Table 2 shows the regression results on the diagonalized data. Compared to results from PCA (Table 1), there seems to be an overall improvement across all models with gradient boosting improving the most, with a 20% decrease in the MRE. Table 3 displays the regression results on the subset data where the survival time was less than or equal to 24 months. As expected, all the regression

[8] This was, of course, due to the computations involving very small quantities.

Table 2: Performance of Regression Models with Block Diagonalized Reduced Feature Set.

Model	MAE	MRE
Linear Regression	14.28	0.76
kNN Regression	15.41	0.83
Gradient Boosting	15.13	0.80

models improved greatly with an average decrease in the MAE of over 50%, and the MRE reaching as low as 41% with gradient boosting.

Table 3: Performance of Regression Models with Block Diagonalized Reduced Feature Set on Subset Data.

Model	MAE	MRE
Linear Regression	6.98	0.53
kNN Regression	6.10	0.42
Gradient Boosting	5.89	0.41

6 Conclusions and Future Work

In this paper, we discussed the domain of healthcare imaging for diagnostics and the implementation of radiomics on CT scans, in particular, to predict the survival rates of lung cancer patients. We explored the engineering of a feature set based on 3D analyses and the related computations associated with the tumor.

We used `Pyradiomics` for the computation of the relevant features to both shape and texture, in the 3D aspect. This resulted in a 110-dimensional feature vector, which was subjected to feature selection and dimensionality reduction using a novel block diagonalization scheme. We achieved an overall improvement in performance from the baseline of the 3D feature set, with the greatest difference in Linear Regression from an MRE of 1.95 to 0.76. Notably, all models achieved better in a short term prediction with the 3D feature sets, enforcing the results found in [10].

With regard to future work, the use of block diagonal strategies for other PR and regressions methods is a fertile field. Besides, the further extension of our research by computing actual 3D features which can be used to augment the feature vector to the previously-computed 3D features through aggregation of successive 2D slices, can surely be applied to other types of cancers.

References

1. Al Mohammad, B. and Brennan, P. C. and Mello-Thoms, C. A Review of Lung Cancer Screening and the Role of Computer-Aided Detection. 2017. *Clinical Radiology*, 72:433-442.
2. Armato III, S. G. and Giger, M. L. and MacMahon, H. Automated Detection of Lung Nodules in CT Scans: Preliminary Results. 2001. *Medical Physics*, 28:1552-1561.
3. Armato III, S. G. and Sensakovic, W. F. Automated Lung Segmentation for Thoracic CT: Impact on Computer-Aided Diagnosis. 2004. *Academic Radiology*, 11:1011-1021.

4. Chabat, F. and Yang, G. Z. and Hansell, D. M. Obstructive Lung Diseases: Texture Classification for Differentiation at CT. 2003. *Radiology* 228.
5. Chen, C. and Twycross, J. and Garibaldi, J. M. A New Accuracy Measure Based on Bounded Relative Error for Time Series Forecasting. 2017. *PloS one*, 12.
6. Demir, O. and Camurcu, A. Y. Computer-Aided Detection of Lung Nodules using Outer Surface Features. 2015. *Bio-Medical Materials and Engineering*, 26.
7. Elicker, B. M. and Webb, W. R. Fundamentals of High-Resolution Lung CT. 2013. Wolters Kluwer.
8. Fan, L. and Fang, M. and Li, Z. and Tu, W. and Wang, S. and Chen, W. and Tian, J. and Dong, D. and Liu, S. Radiomics Signature: A biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground glass nodule 2019 *European Radiology*, 29:889-897.
9. Ghani, T. MCS Thesis, Carleton University. 2019. *On* Forecasting *Lung Cancer Patients' Survival Rates Using 3D Feature Engineering.*
10. Ghani, T. and Oommen, B. J., Enhancing the Prediction of Lung Cancer Survival Rates using 2D Features from 3D Scans. *Proceedings of ICIAR'20, the 2020 International Conference on Image Analysis and Recognition.* Povoa de Varzim, Portugal (Virtual), June 2020.
11. Grove, O. and Berglund, A. E. and Schabath, M. B. and Aerts, H. JWL. and Dekker, A. and Wang, H. and Velazquez, E. R. and Lambin, P. and Gu, Y. and Balagurunathan, Y. Quantitative computed tomographic descriptors associate tumour shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. 2015. *PloS one*, 10.
12. Hall, E. L. and Kruger, R. P. and Dwyer, S. J. and Hall, D. L. and McLaren, R. W. and Lodwick, G. S. A Survey of Preprocessing and Feature Extraction Techniques for Radiographic Images. 1971. *IEEE Transactions on Computers*, 100.
13. Haralick, R. M. and Shanmugam, K. and Dinstein, H. Textural Features for Image Classification. 1973. *IEEE Transactions on systems, man, and cybernetics*, 6:610-621.
14. Kim, N. and Seo, J. B. and Lee, Y. and Lee, J. G. and Kim, S. S. and Kang, S. H. Development of an Automatic Classification System for Differentiation of Obstructive Lung Disease using HRCT. 2009. *Journal of Digital Imaging*, 22.
15. Messay, T. and Hardie, R. C. and Tuinstra, T. R. Segmentation of Pulmonary Nodules in Computed Tomography using a Regression Neural Network Approach and its Application to the Lung Image Database Consortium and Image Database Resource Initiative Dataset. 2015. *Medical Image Analysis*, 22:48-62.
16. Paul, R. and Hawkins, S. H. and Schabath, M. B. and Gillies, R. J. and Hall, L. O. and Goldgof, D. B. Predicting malignant nodules by fusing deep features with classical radiomics features. 2018. *Journal of Medical Imaging*, 5.
17. Siegel, R. L. and Miller, K. D. and Jemal, A. Cancer Statistics. 2018. *CA: A Cancer Journal for Clinicians*, 68:7-30.
18. Singadkar, G. and Mahajan, A. and Thakur, M. and Talbar, S. Automatic lung segmentation for the inclusion of juxtapleural nodules and pulmonary vessels using curvature based border correction. 2018. *Journal of King Saud University - Computer and Information Sciences.*
19. Zhao, B. and Gamsu, G. and Ginsberg, M. S. and Jiang, L. and Schwartz, L. H. Automatic Detection of Small Lung Nodules on CT Utilizing a Local Density Maximum Algorithm. 2003. *Journal of Applied Clinical Medical Physics*, 4:248-260.
20. Zhou, S. and Cheng, Y. and Tamura, S. Automated lung segmentation and smoothing techniques for inclusion of juxtapleural nodules and pulmonary vessels on chest CT images. 2014.]newblock *Biomedical Signal Processing and Control*, 13:62-70.