Computing and Artificial Intelligence 2024, 2(1), 450.
https://doi.org/10.59400/cai.v2i1.450

Article

Medical image classification using a quantified hazard ratio and a multilayer fuzzy approach

Kishore Kumar Akula[1,*], Monica Akula[2], Alexander Gegov[3,4]

[1] Statistics eTeachers Group, Royal Statistical Society, North York, Ontario M3C 2Z1, Canada
[2] Department of Neuroscience, McMaster University, Hamilton, Ontario L8S 4L8, Canada
[3] School of Computing, University of Portsmouth, Portsmouth PO1 3HE, United Kingdom
[4] English Faculty of Engineering, Technical University of Sofia, 1756 Sofia, Bulgaria
*** Corresponding author:** Kishore Kumar Akula, kishorekumarakula@gmail.com

CITATION

Akula KK, Akula M, Gegov A. Medical image classification using a quantified hazard ratio and a multilayer fuzzy approach. Computing and Artificial Intelligence. 2024; 2(1): 450. https://doi.org/10.59400/cai.v2i1.450

ARTICLE INFO

Received: 2 January 2024
Accepted: 29 January 2024
Available online: 18 February 2024

COPYRIGHT

Copyright © 2024 by author(s). *Computing and Artificial Intelligence* is published by Academic Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: We previously developed two AI-based medical automatic image classification tools using a multi-layer fuzzy approach (MFA and MCM) to convert image-based abnormality into a quantity. However, there is currently limited research on using diagnostic image assessment tools to statistically predict the hazard due to the disease. The present study introduces a novel approach that addresses a substantial research gap in the identification of hazard or risk associated with a disease using an automatically quantified image-based abnormality. The method employed to ascertain hazard in an image-based quantified abnormality was the cox proportional hazard (PH) model, a unique tool in medical research for identifying hazard related to covariates. MFA was first used to quantify the abnormality in CT scan images, and hazard plots were utilized to visually represent the hazard risk over time. Hazards corresponding to image-based abnormality were then computed for the variables, 'gender,' 'age,' and 'smoking-status'. This integrated framework potentially minimizes false negatives, identifies patients with the highest mortality risk and facilitates timely initiation of treatment. By utilizing pre-existing patient images, this method could reduce the considerable costs associated with public health research and clinical trials. Furthermore, understanding the hazard posed by widespread global diseases like COVID-19 aids medical researchers in prompt decision-making regarding treatment and preventive measures.

Keywords: cox proportional hazards model; hazard ratio; CT scans; fuzzy system; survival analysis

1. Introduction

Physicians rely heavily on computed tomography (CT) scans to detect diseases like cancers. They spend much of their time observing one image to decide if the patient has cancer, and if so, the stage of the cancer. Qualitative analysis of an abnormality in a diagnostic image is not always a robust method of analyzing an abnormality and may be contribute to reduced inter-rater reliability. However, when the abnormality can be quantified as a percentage, this both improves consistency in the assessments performed by different healthcare practitioners, while also providing a more accurate method of evaluating the aberration.

Our previous studies described methods for quantification of the abnormality in images [1]. A novel method, an AI-based medical image classification tool using a multi-layer fuzzy approach (MFA), was introduced in the first study. The second study focused on developing a more robust version of the MFA method using the many-to-many relation [2] (MCM) (manuscript under review) to find the abnormality in the

images, which is the disease present in the objects in the images. In the first study, a simple process from cognitive science known as assessment was used, which involves comparison of a normal image with an abnormal image. In this process of comparison, the structural similarity index (SSI) between two images is the similarity among the images [3]. When this similarity in percentage form is subtracted from 100, the calculation quantifies the abnormality in the abnormal image. In our second study using MCM, the comparison was made more robust, such that instead of considering one image, multiple normal images were compared with a single abnormal image, and all the similarity scores were averaged to obtain the abnormality in the image. This process was continued for all of the abnormal images, which was a more robust and accurate method of analyzing the images. The images were then classified, which was performed using multilayer fuzzy systems, computational intelligence rules, computer vision from AI and manual testing tools.

A literature search was conducted using the key words for any studies like MFA or MCM that is to convert the image based abnormality to quantity or finding the hazard ratio corresponding to the image based abnormality, but no similar studies were found [4,5].

1.1. The rationale of the study

Currently, there is limited research on using diagnostic image assessment tools, particularly methods that can be used for small and large data sets like the MFA and MCM methods, for prognostic applications, such as prediction of mortality risk. Moreover, following the inspection of patient images, such as CT scans, the images are not typically used for subsequent research endeavours. In the present study, a novel method was developed using conversion of image-based disease severity quantity into the hazard ratio through the use of an Artificial Intelligence (AI) medical image classification-based multilayer fuzzy approach (MHM). In this method, the hazard ratio corresponding to an image-based disease severity quantity was found using an AI medical image classification-based multilayer fuzzy approach, and survival analysis, a domain of biostatistics, as well as concepts from MFA [1], and MCM, a modified version of MFA [2], were used. This study posits that quantifying abnormalities in these images and determining the associated hazard within a group of individuals could substantially contribute to public health efforts and research initiatives to optimize resource allocation, substantially improve treatment outcomes that increase survival and reduce the expenses incurred during the conduct of clinical trials involving patient studies. Furthermore, the simplicity of incorporating basic software code in this study renders it easily applicable with existing technologies. Ultimately, the utilization of images acquired for the individualized treatment of patients can extend the scope of the study to encompass the health of larger populations, countries, or even the entire global patient demographic. This extension arises from the study's classification of patients based on the hazard or risk associated with abnormalities resulting from various diseases.

1.2. Aims of the study

1.2.1. Primary aim

The primary aim of the study is the following:

a) To quantify the abnormality in the form of images present in the objects of images.
b) Finding the hazard ratio corresponding to this quantified abnormalities in a set of images as well as to classify the data into some sub-groups and to check the effect of the hazard due to the abnormality.
c) To classify the abnormality based on the hazard or risk due to the abnormality.

1.2.2. Secondary aim

The secondary aim of the study is to apply the concepts of the primary aim to a medical CT scan image data set taken to diagnose the lung cancer, as described in section 1.1. The specific components of the second aim are as follows:

a) To quantify the cancer or abnormality present in image form in a CT scans data set.
b) To find the hazard ratio corresponding to the above-quantified abnormality of the data set, as well as to classify the data into some sub-groups and to check the effect of the hazard present in the images due to abnormality.
c) To classify the abnormality based on the hazard ratio due to the abnormality present in the CT scans of the data.

2. Materials and methods

2.1. Materials

2.1.1. Participants

In the dataset [6] we considered, nothing was known about the age, gender, participants and ethnicity of the patient. The only information available was a confirmed lung cancer, and did not include the time of event, status or smoking status. To find the hazard ratios using the Cox PH model, few more variables were needed. That is in order to make the dataset a survival data set, some variables such as ‘time’, ‘status’ were simulated and for the subgroup analysis ‘smoking-status’, ‘gender’, and ‘age’ were simulated.

2.1.2. Data set and data dictionary used in the current MHM study

The number of CT scans in the data set used in this study was 67 [6]. Among the CT scans, images of the right lungs were extracted from the CT scan. The right lung was chosen at random for study due to the noise in the images. Additionally, the normal image data set was a small data set with 20 images, similar to the MCM study [2].

2.1.3. Variable description

First, after finding the abnormality in the different images, they were classified as per the stages of abnormality (**Table 1**). Stage 1 has less abnormality than the other stages and more normality, and stage 4 has greater abnormality and less normality, such that from stages 1 to 4, the normality decreases or the abnormality decreases. Furthermore, the elements at each stage represent the normality percentage in the images. In addition, the variable ‘time’ is the time to event, ‘status’ is the occurrence of disease or death, ‘gender’ is whether the patient is male or female, and ‘smoking’

is the smoking-status.

2.1.4. Data types

The variables ‘stage 1 to 4’ and time were continuous variables. Stages were in percentages and time was in time units, which can be days, months or years, while the rest were categorical variables. ‘Status’ can be categorized as live or dead, ‘gender’ was female (1) or male (0), and similarly for non-smoking or smoking- status.

2.1.5. Survival analysis time data

The survival analysis [7,8] data only has information on abnormalities in the form of CT scans. However, to find the hazard due to the cancer in that particular study, area or group, we need information on a few more variables, so these variables were simulated, because in order to find the hazard ratios, these variables are needed. Here, we included gender, age and smoking-status, and by using these variables the method was developed.

Survival time is defined as time until retrieval or time until an end point whether it is medical or industrial [9], and can also be called the ‘lifetime’. In the current data set, it is the time to disease (cancer), disorder occurrence or death. The units of the time in this study can be days, months or years.

The variable, ‘status,’ is the status of the patient, which is the death of the patient. It is a binary variable, such that ‘0’ denotes being alive or not affected, and ‘1’ denotes being affected by abnormality or cancer or death. The next variable added was ‘gender’, which is male or female, a binary variable. Lastly another variable ‘smoking-status’ was added, which was also a binary variable.

2.2. Methods

2.2.1. Test statistic

The test statistic used for the Cox PH model is the chi-square test for the *p*-value assessment of each variable or covariate used in the model.

2.2.2. Finding the risk or hazard due to the quantified abnormality or disease

A hazard represents any factor with the capability to cause damage, harm, or negative health consequences to an object or individual. The hazard ratio is also known as the relative risk [10]. The ratio compares the risk of disease or death between the exposed and unexposed populations, like, for example, the ratio of the number of patients in a country with lung cancer and the population of that country. The baseline hazard is the hazard when all the covariates are zero or at their reference levels. In medical sciences and epidemiology, the hazard ratio plays a very important role in helping assess potential patient outcomes and classify patients on the basis of the hazard. It explains how much risk is associated with a certain disease, drug or a habit like smoking. To find the hazard or risk due to the quantified abnormality or disease, the cox PH was used, which falls under the domain of survival analysis, a branch of statistics. The primary emphasis of survival analysis is on the time until an event, which is death, or until the disease occurs. This could be the time until a patient experiences a relapse or the time until a machine fails, depending on the application of the study.

2.2.3. Cox proportional hazard (PH) model to find the hazard ratios of variables

The cox PH semi-parametric regression model has been widely used in the medical and clinical fields, as well as the industry [11,12]. Since the Cox PH model is a statistical model, certain procedures used with statistical models, like hypothesis testing, need to be implemented when calculating the coefficients obtained when Cox PH model is used.

Normally, in clinical/medical, epidemiology, and industry studies, the Cox PH model is used in the context of treatment or intervention, age, gender, disease stage or severity, socioeconomic status, smoking-status, and other important factors. In addition to the mandatory variables to use or to calculate hazard ratios are time to event, and status of the disease (please clarify this sentence). In the current study, the covariates of the Cox PH model are the abnormalities in the images converted into percentages using MCM, a novel approach to find the hazard ratio, that provides information on mortality risk, corresponding to the abnormality obtained using MFA and MCM.

2.2.4. Steps in using Cox PH model

The steps in using Cox PH model to find the hazard due to image based quantified abnormality are the following:

Step 1. The data was prepared to use with the Cox PH model, using the CRAN-R software and with the proper libraries, which yielded the hazard ratios, *p*-values, and confidence intervals for hazard ratios.

Step 2. Next, the PH model was tested for accuracy, which was done by checking the statistical significance of the coefficients. The *p*-value associated with each beta coefficient helps assess the statistical significance of the corresponding covariate. A small *p*-value (typically less than a chosen significance level, e.g., 0.05) indicates that the covariate is statistically significant in predicting the hazard, suggesting that it is likely not due to random chance.

Step 3. The proportional hazards assumptions were checked. The Cox PH model was based on some assumptions [10–14], and when the model is used, the assumptions are to be checked. If the assumptions are not satisfied, then it means that the mode was not fit properly. At the same time, for all data sets, there is no need that all assumptions are to be checked. In this study, the only key assumptions that were checked were:

1) The hazard ratio is presumed to be constant over time, which means that the ratio of the hazards for any two persons or patients or commodities is constant over time.

2) The independence assumption, which states that the observations in the dataset are expected to be independent. This means that the incidence or non-incidence of an event for one patient, person or commodity does not provide information about the occurrence or non-occurrence of an event for any other subject.

2.3. Statistical procedures used in MHM

2.3.1. The software used

The software used in the current study were Python with Anaconda as the backend and spider as the frontend. Python was utilized to compare images and to acquire the similarity indices as described in our previous MCM study [2] using

OpenCV from AI. Python was also used to plot some graphs. Next, CRAN-R was used to calculate the hazard ratios using Cox PH model and to plot some additional graphs.

2.3.2. Data cleaning

The most important step in the MHM method was ensuring that when two images were compared, only the object, lungs, were compared. Parts of the image other than lungs were avoided, so that only the normal lung was compared with the abnormal lung in order to avoid any noise when obtaining the similarity index.

2.3.3. Statistical tests and tools used

As the secondary aim involves the application of the current study MHM to CT scans of lung cancer, in order to find the hazard due to the abnormality in CT scans, the Cox PH model was used. The hazard ratios were obtained by using the COX PH model on the abnormality data. First, the global model was tested for statistical consistency followed by the coefficients. Lastly, the confidence intervals were examined to check the model for goodness of fit or not.

2.3.4. Method to establish the primary aim

All the steps in the primary aim are shown in the schema in **Figure 1**. These steps are the same as the MFA or MCM method [1,2] for obtaining the fuzzy set. The general form of the fuzzy set obtained is {Patients' ID, abnormality score}. Next, the Cox PH model was used together with some basic statistics to get hazard ratios and the survival probabilities on graphs for the variables. The rest of the method from the fuzzy logic steps until the end of the method was similar to the MCM method.

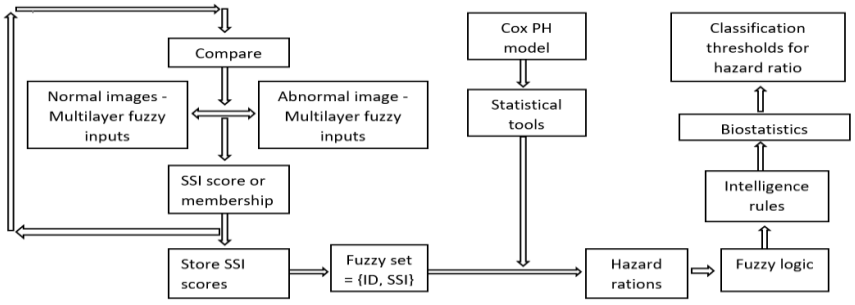

Figure 1. The schema for the study MHM.

2.3.5. Method to establish the primary aim 1.2.1 (a)

To quantify the abnormality in the form of images, MCM [2] was used. That is, a few normal images were compared with the abnormal image to get the SS followed by taking the mean of all the SSs to get the final SS of the abnormal image. The abnormality was obtained by subtracting the normality in percentage form from 100, which was the process used to quantify the abnormality in the form of an image.

2.3.6. Method to establish the primary aim 1.2.1 (b)

The data classified in the above section was used with the Cox PH model to calculate the hazard ratios as discussed in section 1.4.4, and three steps were used with the Cox PH model.

Step 1: The model is $h(t/X) = h_0(t)\exp(\beta_1 \text{Stage1} + \beta_2 \text{Stage2} + \ldots + \beta_k \text{Stage}k)$, where the required hazard ratios are $\exp(\beta_k)$. To get the specific β_k the data has to be used, which was carried out in the subsequent sections using CRAN-R software.

Step 2: The global statistical tests, the likelihood ratio test, Wald test, and logrank tests were used to check for the significant overall association between the variables. The *p*-value for each variable was used to check if the coefficients occurred by chance or if there is statistical consistency depending on the obtained *p*-value being less than or greater than a standard value (0.05).

Step 3: The Cox PH model's assumptions were checked as mentioned in section 1.4.7 (b) above. The first assumption is the constant hazard ratio (proportionality), which states that the hazard ratio should remain constant over time. A rule of thumb to prove the above is that the hazard curves for the groups should be proportional and cannot cross each other [9]. The second assumption is that the observations in the dataset are expected to be independent, and in the currently used data set, neither the patients nor their data are related to each other. Confidence intervals are discussed in detail numerically while discussing the secondary aim.

2.3.7. Methods to establish the primary aim 1.2.1 (c)

In the previous section, the hazard ratio was obtained as $\exp(\beta_i)$, the hazard ratio or the relative risk based classification is as follows [8–11]:

If $\exp(\beta_i) < 1$, then the hazard or risk is reduced.

If $\exp(\beta_i) = 1$, then there is no effect on hazard or risk.

If $\exp(\beta_i) > 1$, then there is increase in hazard, where 'exp' stands for the exponentiation of β_i.

3. Results

3.1. Methods to establish the secondary aim 1.2.2 (a)

The secondary aim is the application of the primary aim to a data set of CT scans taken for the diagnosis of the lung cancer. To quantify the abnormality in the form of CT scans or images of the data set, that is, to quantify the abnormality or cancer in the form of images present in the CT scans, the abnormality present in the lung in the CT scan images of the data set was quantified by using the MCM method [2] and the method presented in the primary aim 1.2.1 (a).

In addition, a few more variables were simulated for the full application of the primary aims. The new sample data set for stage 4 lung cancer is show in **Table 1**. The only variable that was not simulated was 'stages'. This variable consisted of the abnormalities obtained by using MCM to get SSI and subtracting the SSI from 100. The Cox PH model can be used on the data set to find the hazard ratios (**Table 1**).

Table 1. A sample of the data with simulation for variables and abnormality classified as stage 4.

Patient	Time	Status	Stage 4	Age	Gender	Smoking-status
1	6.6	0	48.80	50	Female	Smoker
2	6.05	1	46.69	50.05	Female	Non-Smoker
3	2.75	0	52.40	50.12	Male	Non-Smoker
4	49.67	1	36.76	50.04	Female	Non-Smoker

Table 2. The *p*-values for global statistical tests for stage 3.

Test	Calculated value	Degrees of freedom	*p*-value
Likelihood ratio test	13.98	4	0.0007
Wald test	10.7	4	0.03
Score (Logrank)	13.36	4	0.01

3.2. Methods to establish the secondary aim 1.2.2 (b)

The Cox PH model has to be used in three steps to find the hazard ratio corresponding to the above-quantified abnormality of the dataset, which is described below for the current data set:

Step 1. The hazard ratios, and *p*-values attached with the covariates 'stage1–4', 'gender', and 'smoking-status' were calculated by applying the Cox PH model using the CRAN-R packages. Since this model is a statistical model, basic statistical concepts like global study's *p*-values, and the *p*-values for the covariates were also obtained to check the significance of the model fit to the data. The results were tabulated below. In these tables all the hazard ratios were related to the covariates or variables.

Step 2. The statistical significance of each of the coefficients in the table and whether they formed due to chance or not was checked. The global statistical significance was checked in **Table 2**. Subsequently, the statistical significance of the variables, and the confidence intervals for the good fit of the parameters were also checked.

Firstly, the tables for all stages and the statistical analysis of the global model with *p*-values, for example, are given in **Table 2**.

The *p*-values for the three overall tests (likelihood, Wald, and score) for stages 1–4, showed significance, suggesting that the model holds importance. These tests assess the general null hypothesis that all beta coefficients (β) equal 0. In the given instance, the test statistics closely aligned, leading to a firm rejection of the general null hypothesis. That is, the coefficients were existing non zeros.

Secondly, the statistical significance of each covariate in the partially simulated data was checked, which in the below discussion are the hazard ratios or the values under the column with the title 'exp(coefficient)', 'exp(β)' or 'e^{β}' (**Tables 3–6**).

Table 3. Full Cox PH model-hazard ratios and *p*-values for variables of stage 1 lung cancer data.

Covariate	Coefficient	Exp (coefficient)	Standard Err. (coefficient)	*Z*	*Pr* (> |*Z*|)
Stage 1	−0.85790	0.42405	0.38745	−2.342	0.0192
Age	−0.06678	0.93540	0.01762	−1.861	0.0627
Gender male	−1.57594	0.20681	0.32392	−2.051	0.0403
Smoker	1.69714	5.45834	0.32379	2.190	0.0285

Table 4. Full Cox PH model–hazard ratios and *p*-values for variables of stage 2 lung cancer data.

Covariate	Coefficient	Exp (coefficient)	Standard Err. (coefficient)	*Z*	*Pr* (> |*Z*|)
Stage 2	0.89030	2.43585	0.3874	2.298	0.0216
Age	0.01564	1.01576	0.0176	0.888	0.3746
Gender male	0.19913	1.22034	0.3239	0.615	0.5387
Smoker	0.72066	2.05579	0.3237	2.226	0.0260

Table 5. Full Cox PH model—hazard ratios and *p*-values for variables of stage 3 lung cancer data.

Covariate	Coefficient	Exp (coefficient)	Standard Err. (coefficient)	*Z*	*Pr* (> |*Z*|)
Stage 3	0.04527	1.04631	0.01990	2.275	0.0229
Age	0.20473	1.22762	1.6417	0.125	0.9008
Gender male	−0.27398	0.76035	0.30131	−0.909	0.3632
Smoker	−0.53360	0.58647	0.30069	0.30069	0.0760

Table 6. Full Cox PH model—hazard ratios and *p*-values for variables of stage 4 lung cancer data.

Covariate	Coefficient	Exp (coefficient)	Standard Err. (coefficient)	*Z*	*Pr* (> |*Z*|)
Stage 4	−0.16303	0.84857	0.06953	−2.234	0.0190
Age	0.09225	1.096642	0.04497	2.051	0.0402
Gender male	−0.08372	0.91969	0.56618	−0.148	0.8824
Smoker	−0.39789	0.67174	0.61522	−0.647	0.5178

*where Err is the error.

Table 3 shows that the variables ‘stage 1’ and ‘smoking smoker’ (wherein patient is a smoker) had *p*-values less than 0.05, that is, these values did not exist by chance. The hazard ratio for the variable stage 1 was exp(β) = 2.43585 and for the variable, ‘smoking smoker’, this value was 2.05579, and the *p*-values for other variables were less than 0.05. We can also consider the hazard ratios of those variables for which the *p*-value is greater than 0.05, but they could have occurred by chance, that is it suggests that the observed data is not consistent with the null hypothesis.

Similarly, from **Table 4**, the variables ‘stage 2’, ‘gender male’, and ‘smoking smoker’ had *p*-values less than 0.05, which is to say that they were statistically significant, and in **Table 5**, the variables ‘stage 3’ and ‘age’ had *p*-values less than 0.05, while **Table 6** shows that the variable ‘stage 4’ had a *p*-value less than 0.05.

To conclude, the spread of the cancer as seen on CT scans were converted to quantities and the hazards corresponding to the cancer were calculated. These models can be used to predict the trends in hazard or risk due to lung cancer and the other variables. Some variables had statistical significance and others did not. The variables that were not statistically significant were dropped from the Cox PH model.

The only specific models with statistical significance for the stages of lung cancer were as follows:

For stage 1: $h_1(t/X)/h_0(t) = e^{(0.89030\text{Stage 1}+0.01564\text{Age} + 0.72066\text{Smoking smoker})}$,

where $h_1(t/X)/h_0(t)$ can be interpreted as the ratios of infected to non-infected, smoking

to non-smoking, and female to male.

Step 3. The 3rd step in finding the hazard ratios is using the confidence intervals to check if the parameters, that is, hazard ratios are effective for prediction using the above models for all stages of lung cancer. The following rules of thumb apply only to parameters or hazard ratios that are in the model, which were the variables that were statistically significant. Firstly, the parameters should lie inside the confidence intervals, and we can observe from **Table 7** that the parameters are lying in the corresponding confidence intervals. Secondly, the width of the confidence intervals should be very narrow, and this can also be observed.

Table 7. Full Cox PH model—confidence intervals for stage 4.

Covariate	Exp (coefficient)	Lower limit 0.95	Upper limit 0.95
Stage 4	0.8496	0.7413	0.9736
Age	1.0966	1.0041	1.1977
Gender male	0.9198	0.3032	2.7898
Smoker	0.6717	0.2012	2.2432

The hazard gradually increased as time passes as seen in **Figure 2**. At 45 units of time, the hazard increases very steeply, leading to the next stage of cancer. Here, if the advancement of the spread of lung cancer or abnormality is increased, then the hazard increases with time.

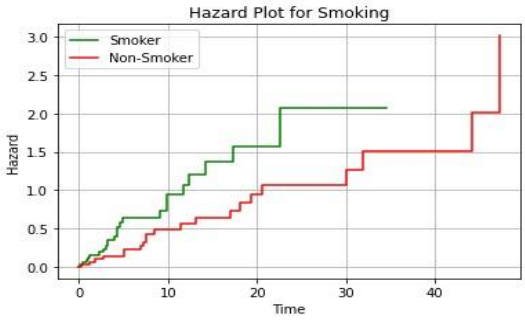

Figure 2. Hazard for smoking status among stage 1 lung cancer patients.

In **Figure 2**, the data for the variable, smoking-status, is available until 35 units of time (what is the specific unit here?). The fluctuation in hazard in males and females can be observed in **Figure 3**. Males have a lower hazard than females. The other variable that had a *p*-value < 0.05 was 'gender male' which had a hazard of 0.2068, and this indicates that males had a decrease of lung cancer by 0.2068 compared to females. Similarly the pattern in hazard can be noticed for the variables in **Figures 2–4**.

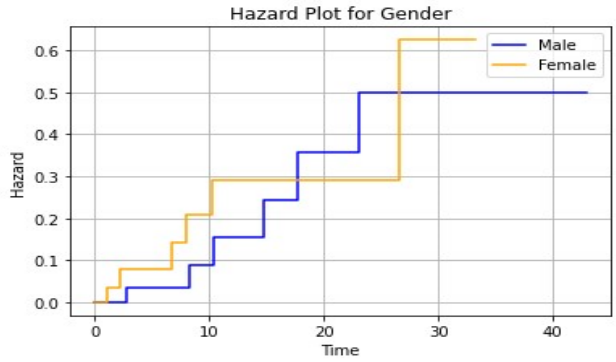

Figure 3. Hazard for the variable, gender.

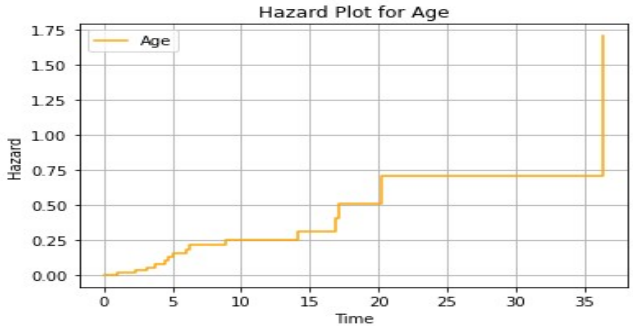

Figure 4. Hazard for the variable age for stage 3 of lung cancer.

3.3. Methods to establish the secondary aim 1.2.2 (c)

Firstly, stage 1 of lung cancer was considered (**Table 3**). The hazard ratio for the variable 'stage 1' was 2.4359 > 1. This implies that for every unit increase in the abnormality, the hazard increases by 2.4359. In addition, this decrease was not by chance as the variable, 'stage 1', was statistically significant. Next, at this stage 1, the other variable, smoking, a categorical variable, had a *p*-value < 0.05 with a hazard ratio of 2.0558 > 1. This also indicates that being a smoker has 2.0558 times greater risk or hazard than being a non-smoker.

Secondly, the variables 'stage 2' (**Table 5**), 'gender male' and 'smoking smoker' had *p*-values < 0.05, and hence, were statistically significant. The hazard ratios and the classification for these variables were as follows:

Variable 'stage 2' had a hazard ratio of 0.4241 < 1, which indicates that the hazard decreases by 0.4241 for every unit of abnormality. Normally, stage 2 should have a hazard greater than 1. This decrease in hazard may be attributed to treatment used if any. The last variable considered in the stage 2 category of lung cancer was the variable, smoking. **Table 4** shows the hazard for the variable, 'smoking smoker', indicating that a person who smokes has a hazard ratio 5.4583 times greater than that of non-smokers at stage 2 as depicted in **Figure 5**.

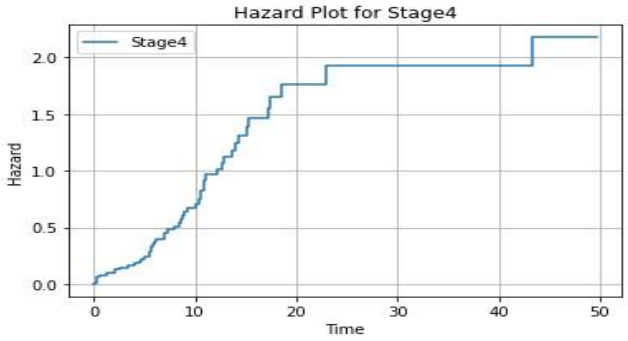

Figure 5. Hazard for the variable, 'stage 4'.

Thirdly, at stage 3 of lung cancer (**Table 5**), the variables for which the *p*-value < 0.05 were for the variables, 'stage 3' and age. The hazard for the variable stage 3 was 0.8496, which means that the hazard is decreased in the patients who were affected. This decrease might be attributed to treatment for cancer. Next, the variable, age, had a hazard of 1.0966, indicating that the hazard was slightly increasing with the age of the patient.

Lastly, the only statistically significant variable is stage 4, for which the *p*-value is < 0.05, with a hazard ratio of 1.0463. This signifies that there is increase in hazard due to cancer.

In this section, lung cancer in CT scans was classified on the basis of the hazard due to the lung cancer as follows:

A hazard ratio of < 1 implies a decrease in the hazard due to cancer.

A hazard ratio > 1 implies an increase in hazard due to lung cancer.

The hazard ratio for stage 1 > 1, and the hazard ratio for smokers at stage 1 lung cancer > 1.

The hazard for stage 2 < 1, and the hazard ratio for gender male of stage 2 < 1.

The hazard ratio for smokers with stage 2 lung cancer > 1.

The hazard ratio for stage 3 < 1, and the hazard ratio for age at stage 3 lung cancer > 1.

The hazard ratio for stage 4 (**Table 6**) lung cancer > 1.

The assumptions of the PH model were checked. For the first assumption, the hazard curves were not intersecting each other for the stage 1 variable. In addition, this has been verified for other stages, and the PH assumptions were met (**Figures 6** and **7**). This indicates that the Cox PH model fit was successful.

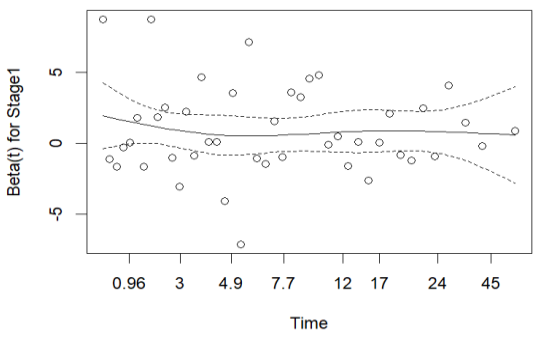

Figure 6. The hazard curves for the variable, 'stage 1'.

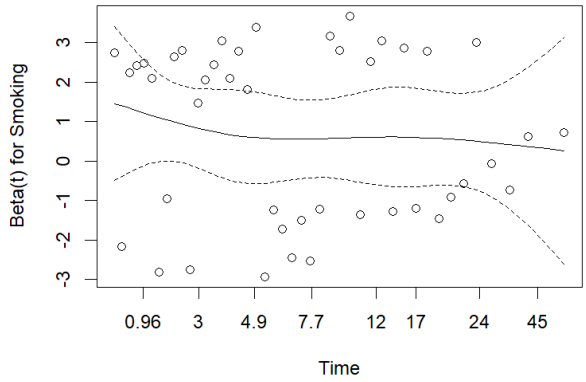

Figure 7. The hazard curves for the variable, 'smoking'.

For the second assumption, the patients were independent of each other, so the second PH assumption was also met.

4. Discussion

The hazard ratio is an important indicator of mortality risk that provides information on disease prognosis and helps inform decisions about treatment made by healthcare professionals. There is limited literature on using biostatistics in combination with AI-based automated image analysis methods, particularly methods that can be used to analyse small data sets without the need for a training data set. The purpose of the current study was to find the hazard or risk due to the abnormality present in the images of a group of people or commodities using a quantified abnormality score calculated from the previously developed MFA and MCM methods [1,2]. This is a novel approach to using AI-based automated image analysis methods to determine the mortality risk associated with disease and was used for the first time.

The MFA or MCM method was used successfully to first find the cancer in the CT scans together with some simulated variables and to classify them on the basis of lung cancer (**Tables 1–4**), which would give the physician information on the cancer spread in the form of a numerical quantity, leading to a better understanding of a patient's disease progression. Secondly, for the group of patients at the different stages of cancer, the hazard ratio was found using techniques from survival analysis in a novel approach. Furthermore, the hazard was studied within subgroups of patients, like, for example, on the basis of smoking. Thirdly, the groups of patients with lung cancer were classified on the basis of hazard due to the cancer affecting them, or any association with their gender or smoking status. There is no other research in the literature like the current study that calculates the hazard corresponding to the converted cancer in image form to a quantified hazard. Moreover, none of the results contradict the hypotheses of this MHM study.

The clinical significance of the current study is that typically in medicine, the hazard due to cancer is attributed to deaths, and here, for the first time, the hazard due to cancer has been estimated using information on cancer in image form before the death of the patients. With this, the physician not only has information about the cancer in image form, but also new information on the numerical hazard due to the cancer, and this could potentially lead to improved treatment. Another significant finding of this study is that the CT scans taken to study the particular patient can be useful to

study the public health of the group. This can impact medicine by allowing healthcare professionals to provide better, earlier treatment and save more lives. Furthermore, the MHM method not only saves a lot of physicians' time, but it can also provide a means of increasing inter-rater reliability. In addition, normally when cancer is studied, other variables or symptoms are mixed with cancer and increase the number of confounding variables. This method does not involve any confounding with other variables, because in the CT scans, the cancer is visible and is converted to number, and then its corresponding hazard is calculated. Hence, there are no confounding variable interactions in this study of cancer.

One limitation of this study is that it needs a considerable amount of data; however, it does not require data sets of a thousand or more, because the Cox PH model only needs enough data points for its assumptions to be met. In addition, the idea of the study has greater generalizability to other fields, but if the images are from a different area of science then the thresholds of classification must be found again. However, the thresholds obtained for cancer in the current study can be used for other CT scans of cancer or other populations with lung cancer. The recommendations for future research are to use the same CT scans or original format of the images instead of conversion to other formats. This will enhance the study and remove extra fuzziness due to the conversion of images.

The key findings of this MHM study are the better understanding of the cancer or abnormality because of its ability to find the hazard ratio corresponding to the cancer. There is also no need to devote extra funding for the study of cancer, as existing CT scans were used in this study. Moreover, the method is simple to write using a CRAN-R package. Overall, MHM is based on the simple idea of converting abnormality in image form to a number, and in turn, finding the hazard due to the abnormality.

5. Conclusion and future directions

To conclude, this is a cross application study where AI, fuzzy systems, and computational intelligence techniques were used to convert abnormality or cancer in the form of an image to a quantity, and subsequently find the hazard resulting from the abnormality or cancer using the Cox PH model of survival analysis, after which the hazard was classified into categories. This was a novel approach that was used for the first time in the literature. Moreover, as very little software was used along with existing images, this study has practical applications. Future research will focus on classifying abnormality in images into survival probabilities.

Author contributions: Conceptualization, KKA; methodology, KKA; software, KKA; validation, KKA, AG and MA; formal analysis, KKA; investigation, KKA; resources, KKA; data curation, KKA; writing—original draft preparation, KKA; writing—review and editing, KKA and MA; visualization, KKA; supervision, KKA, MA, and AG; project administration, KKA. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Akula KK, Gegov A, Arabikhan F. Artificial Intelligence-Based Medical Image Classification using a Multilayer Fuzzy Approach. Wseas Transactions on Computers. 2023; 22: 206-217. doi: 10.37394/23205.2023.22.24
2. Akula KK, Gegov A, Arabikhan F, et al. Medical Image Classification using a Many to Many Relation and a Multilayer Fuzzy Approach. Electronics MDPI 2024 (under review).
3. Wang Z. Multi-scale structural similarity for image quality assessment. In: Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2003.
4. Wulczyn E, Steiner DF, Xu Z, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. Hsieh JCH, ed. PLOS ONE. 2020; 15(6): e0233678. doi: 10.1371/journal.pone.0233678
5. Furrer MA, Sathianathen N, Gahl B, et al. Oncological outcomes after attempted nerve-sparing radical prostatectomy (NSRP) in patients with high-risk prostate cancer are comparable to standard non-NSRP: A longitudinal long-term propensity-matched single-centre study. BJU International. 2023; 133(1): 53-62. doi: 10.1111/bju.16126
6. National Cancer Institute. Available online: https://www.cancerimagingarchive.net/ (accessed on 23 October 2023).
7. John M. Last, A dictionary of epidemiology, Oxford University PESS, 2001.
8. Bradburn MJ, Clark TG, Love SB, et al. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer. 2003; 89(3): 431-436. doi: 10.1038/sj.bjc.6601119
9. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972; 34(2): 187-202. doi: 10.1111/j.2517-6161.1972.tb00899.x
10. Balakrishnan N, Rao CR. Handbook of Statistics 23: Advances in Survival Analysis, North Holland. 2004.
11. Bradburn MJ, Clark TG, Love SB, et al. Survival Analysis Part II: Multivariate data analysis—an introduction to concepts and methods. British Journal of Cancer. 2003; 89(3): 431-436. doi: 10.1038/sj.bjc.6601119
12. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982; 69(1): 239-241. doi: 10.1093/biomet/69.1.239
13. Kleinbaum DG, Klein M. Survival Analysis. Springer New York; 2012. doi: 10.1007/978-1-4419-6646-9
14. Deo SV, Deo V, Sundaram V. Survival analysis—part 2: Cox proportional hazards model. Indian Journal of Thoracic and Cardiovascular Surgery. 2021; 37(2): 229-233. doi: 10.1007/s12055-020-01108-7