2021 International Conference on Data Mining Workshops (ICDMW) | 978-1-6654-2427-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDMW53433.2021.00068

# Multimodal Machine Learning for 30-Days Post-Operative Mortality Prediction of Elderly Hip Fracture Patients

Berk Yenidogan*, Shreyasi Pathak*†, Jeroen Geerdink†, Johannes H. Hegeman*† and Maurice van Keulen*
*University of Twente, Enschede, Netherlands,†Hospital Group Twente (ZGT), Almelo, Netherlands
*berkyenidogan@hotmail.com,{s.pathak,m.vankeulen}@utwente.nl,†{J.Geerdink,h.hegeman}@zgt.nl

***Abstract*—Hip fractures in the elderly are a major health care problem in the society. In the clinic, it is important for medical specialists to identify high-risk patients to assist them in the decision-making process in choosing the right treatment. In this paper, we propose a multimodal machine learning model for prediction of 30-days mortality of elderly hip fracture patients. The paper addresses both the clinical problem of identifying high-risk patients and the specific risks involved, as well as the technical problem of how to fuse information from different modalities as input to one predictive model. Our model uses features from three modalities: hip X-ray images, chest X-ray images and structured data and fuses them based on early fusion and late fusion techniques for the prediction task. Our model uses a convolutional neural network to extract features from the chest and hip images before combining them with the structured data. The fused features are further passed through a fusion classifier for the final prediction. The proposed model outperforms a replicated version of Almelo Hip Fracture Score (AHFS-a) with an AUC score of 0.786 vs 0.717. Finally, by the analysis of feature importance, we found that chest X-ray images contain important signs to predict the 30-days mortality of elderly hip fracture patients. We also found that structured and chest X-ray modalities were more important in predicting high-risk patients as compared to hip X-ray modality, though the final results on the test set show that structured, hip and chest X-ray modalities together are needed to get the best performance for predicting 30-days mortality. Further, we achieved the best performance using early fusion with random forest technique, though late fusion achieved a competitive performance.**

## I. Introduction

With changing socio-economic conditions, the life expectancy of humans is increasing. Apart from other consequences, this also results in a higher number of elderly with an acute hip fracture. The estimated number of hip fractures in 1990 throughout the world was 1.26 million and this number is expected to reach 2.6 million and 4.5 million in 2025 and 2050 respectively [1]. Due to the frailty and comorbidities of these elderly hip fracture patients, there is a high risk on early post-operative mortality for some patients. It has been reported that the 30-days mortality rate of hip fracture patients can be as high as 13.3% [2].

It is important for medical specialists to continuously improve the quality of care, reduce the costs involved, and inform patients and their relatives as accurately as possible about their diagnosis, treatment options and possible complications including mortality after hip fracture surgery. In the case of hip fracture patients, it is important to identify the high-risk patients to enable consideration of different treatment pathways, preventive measures in order to guide them more safely through the peri-operative process, and other surgical decisions. A machine learning-based prediction model with sufficient performance would be instrumental for this purpose and of added value during the shared decision-making process.

Research has shown, however, that existing prediction models have not reached sufficient performance yet [3]–[8]. We believe that the problem lies in the limited information that is available to the model, typically only images or only a limited set of structured variables are used in these models. A model capable of harnessing the hidden indications in all available data of the patient may have a better chance of predicting 30-days mortality with sufficient performance. This, however, poses a technical challenge: *multimodal machine learning*, i.e., the ability of using data of different data types (modalities) such as structured variables, images, and text, as combined input for one prediction model.

In this paper, we report on a first step towards these goals leading to the following contributions;

- Designs and experimental comparison for alternative machine learning architectures for combining structured and image data as input for a multimodal model. The experiments are in the context of the 30-days mortality prediction task for hip fracture patients showing increased performance when combining multimodal inputs.
- Local and global post-hoc interpretation of our multimodal model.

The paper is structured as follows. Section 2 summarizes the related work, Section 3 describes the dataset used, Section 4 explains the methodology, followed with experimental setup in Section 5. Section 6 shows the results and Section 7 discusses the results. Finally, Section 8 concludes the paper.

## II. Related Work

### A. 30-day mortality of hip fracture patients

Several research groups have studied the mortality of hip fracture patients after surgery. Nijmeijer et al. [3] developed the Almelo Hip Fracture Score (AHFS) to predict the early mortality risk factor of hip fracture patients after surgery. They trained and validated a logistic regression model on 850

978-1-6654-2427-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00068

patients from Hospital Group Twente (ZGT) using features related to patient characteristics and comorbidities and achieved an AUC of 0.82. Karres et al. [4] used logistic regression to develop the Hip fracture Estimator of Mortality Amsterdam (HEMA). The established model achieved an AUC of 0.81 and 0.79 in the development cohort and validation cohort respectively. Van de Ree et al. [5] developed the Brabant Hip Fracture Score (BHFS-30) using manual backward multivariable logistic regression on 925 patients, achieving an AUC 0.71 and found the significant factors of early mortality were patient characteristics and comorbidities.

Jiang et al. [6] aimed to define the factors affecting in-hospital and 1-year mortality after hip fracture. Using a multivariable backward selection procedure for logistic regression on 3981 patients, they also found patient characteristics and comorbidities to be most important. Predictions for in-hospital mortality achieved AUC of 0.82 on the validation set. Maxwell et al. [7] developed the Nottingham Hip Fracture Score (NHFS) using logistic regression. The study on 4967 patients, achieved an AUC score of 0.719 in the validation set. Marufu et al. [8] recalibrated the Nottingham Hip Fracture Score (NHFS) and the Surgical Outcome Risk Tool (SORT) by using a national dataset of 9017 patients and NHFS achieved an AUC of 0.71 for NHFS and 0.70 for SORT.

Our work differs from the above studies in multiple ways. Firstly, we are using a much higher number of features including image modality which has not been used in other studies. As a consequence, we are not only employing logistic regression but also different machine learning algorithms including deep learning. Most of these algorithms have built-in feature selection capabilities and therefore, we do not follow any pre-feature selection process in our methodology.

### *B. Multimodal Machine Learning*

A modality implies how something happened, and a research problem is identified as multimodal when it consists of more than one modality in its nature [9]. The multimodal machine learning concept has the goal to develop models that can process and associate information from different modalities. Baltrusaitis et al. [9] categorizes multimodal fusion into early, late and hybrid fusion. In the early fusion, modalities are fused after the feature extraction whereas late fusion combines modalities after there is a decision by each modality. So, late fusion combines output of unimodal predictors. Lastly, hybrid fusion incorporates both early fusion and late fusion outputs.

Suk et al. [10] diagnosed Alzheimer's disease with a multimodal approach using deep learning. The fusion of multimodal information from Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) data was done using multimodal Deep Boltzmann Machines (DBM) and outperformed baseline methods. They further investigated the trained model visually and found that their method can hierarchically expose the latent patterns in MRI and PET.

Kenneth et al. [11] fused ECG, blood pressure, saturated oxygen content and respiratory data in order to improve clinical diagnosis of patients in cardiac care units. They concluded that better results can be achieved with the fusion system which can be modeled as a multi-dimensional process. Moreover, Bramon et al. [12] proposed an information-theoretic approach for multimodal data fusion, and showed promising results on medical datasets. Nunes et al. [13] aimed to optimize the efficiency of indexing radiology exams. They developed a multimodal approach that integrates a convolutional neural network (CNN) for processing images with a bidirectional Recurrent Neural Network (RNN) for processing text and showed promising results for the multimodal processing of radiology data. Results showed improvement when the fusion model is pre-trained with large datasets and the accuracy increased from 84% to 98% [13].

In our work, we define modality as a different type & source of the data: structured modality, chest X-ray modality and hip X-ray modality. We extract features using CNN and fuse features from different modalities using early and late fusion techniques [9] and provide a comparison of the methods.

## III. Dataset: Multimodal Data of Hip Fracture Patients

The dataset used in this work was collected from Hospital Group Twente (ZGT), Netherlands checked for quality, and anonymized by the innovation manager of the hospital. Furthermore, all the experiments and explorations have been performed through remote access to a research compute cluster inside the hospital to protect the patient information. The study period used in this dataset is between 01-04-2008 and 31-01-2020. All patients who are 70 years or older and were admitted to the Emergency Room (ER) with an acute hip fracture were included in the dataset. The selection is based on diagnostic treatment code (DBC) for hip fracture which is '218 Femur, proximal (+collum)'. Patients with a femur fracture, periprosthetic fracture, or pathological fracture were excluded. Patients who had total hip arthroplasty or deceased before the surgery were excluded. Furthermore, patients without thorax or hip/pelvis X-rays were also excluded. With all inclusion and exclusion filters, the complete dataset ended up with 2404 patients. In total, the dataset includes 2211 patients who have survived and 193 patients who have deceased within the 30-days period after the operation, hence the 30-days postoperative mortality rate (positive sample rate) is 8%.

Our 3 modalities are as follows. The **structured modality (S)** contains tabular data with each row corresponding to a patient and each column containing data like physical examination, recording of vital signals, electrocardiography, nutrition, mobility, activities of daily living, cognitive problems, comorbidities and living situations resulting in a total of 99 columns. The **hip X-ray image modality (H)** contains the X-ray image of the Anterior-Posterior (AP) view of the hip (pelvis) and the **chest X-rays image modality (C)** contains the X-ray image of the AP view of the chest. Our dataset contains one image for hip X-ray and one image for chest X-ray per patient. All images were resized for the deep learning architectures, i.e., 299x299 for chest images and 224x224 for hip. Images were augmented using rotation 20, width shift

0.2, height shift 0.2, shear 0.2, zoom 0.2, channel shift 10 and flipped randomly on both horizontal and vertical axis.

## IV. Methodology

In this section, we discuss our model architecture, how it handles each data modality and how we fuse the information learnt from each modality to predict 30-days mortality of hip fracture patients. Fig. 1 illustrates our unimodal and multimodal designs. We define 30-days mortality prediction as a binary classification problem, where each patient is to be classified into the survive or deceased class based on their chest X-ray images, hip X-ray images and structured data.

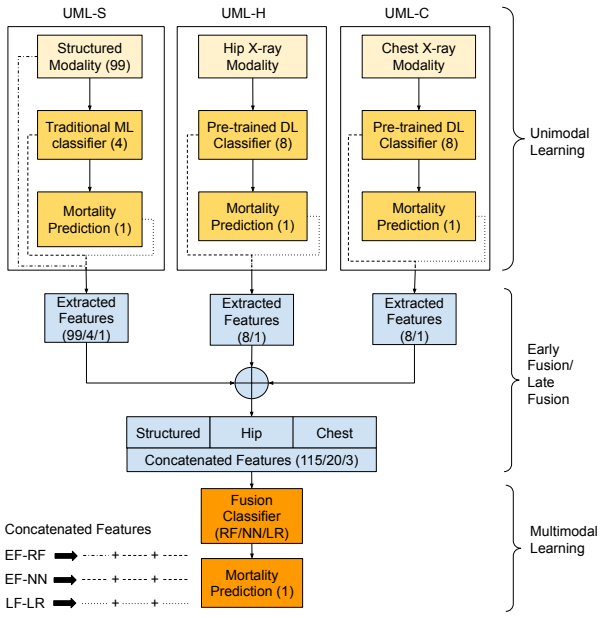


Fig. 1: Unimodal and multimodal learning architectures: 3 separate pipelines specific to each modality for unimodal learning - UML-S, UML-H, UML-C. The feature representation from the unimodal learning used for EF and LF are shown with dashed lines and their variants, EF-RF, EF-NN and LF-LR are summarized in the bottom left. Each variant has 3 dashed lines with a '+' showing horizontal concatenation of 3 modalities S, H and C, respectively. The type of dashed line indicates the feature representation used for the concatenation. The feature dimensions resulting from each block are shown in brackets, e.g. 8 for pre-trained deep learning classifier. *Extracted features* shows the corresponding dimensions depending on the feature representation being used. *Concatenated features* shows the dimensions for EF-RF (99+8+8), EF-NN (4+8+8) and LF-LR (1+1+1) respectively.

### A. Unimodal Learning (UML)

In this subsection, we describe how we perform the training of our model on each data modality.

*1) Structured Modality Learning (UML-S):* We trained 5 traditional machine learning algorithms - Random Forest (RF), Logistic Regression (LR), XGBoost (XGB), Linear SVM, AdaBoost, on the structured data without doing prior feature selection. The goal was to find the best classifier to be used in the later stages of multimodal learning experiments.

*2) Hip X-Ray (UML-H) and Chest X-Ray (UML-C) Image Modality Learning:* We employed deep learning with the help of transfer learning for extracting features from the image modality and then use the extracted features to predict 30-days mortality. Training at this stage was done with 30-days mortality labels. CNN models that were trained on ImageNet [14] were selected and fine-tuned with chest and hip X-ray images separately resulting in a chest X-ray modality model and a hip X-ray modality model, respectively.

### B. Multimodal Learning (MML)

We use the unimodal models trained above to fuse multiple modalities using early & late *fusion techniques (F-tech)*.

*1) Early Fusion (EF):* We design an early fusion architecture by horizontally concatenating the features of all modalities as input for the final *fusion classifier (F-clf)*. We use the following *F-clf* to train the multimodal models - i) *Neural Network variant (EF-NN)*, where *F-clf* is a Neural Network (NN). Training is done simultaneously, which means that back-propagation from the loss function was done throughout the full network such that consequently all modalities are trained at the same time, ii) *Traditional Machine Learning variant (EF-RF)*, where the *F-clf* is RF. Image models were trained separately as two independent networks (unlike EF-NN).

For H & C modality, feature vector from the layer preceding the output layer is taken for concatenation. For S modality, no feature extraction is performed for EF-RF, but for EF-NN variant, we used the feature vector from a small NN. We horizontally concatenate the image feature vector with structured modality features and feed that to the fusion classifier.

*2) Late Fusion (LF):* In this ensemble-like approach, we combine decision outputs of the unimodal models and train a classifier, Logistic Regression *(LF-LR)*, by using the final decision from different modalities as features. We use logistic regression to learn the importance of the three modalities instead of just taking an average as the latter would assume all the three modalities have equal importance.

### C. Explainability of Machine Learning

In health care, the purpose of applying machine learning is not just to obtain predictive models, but often also to gain insight into the predictive influence of the various features. The prediction model that we attempt to explain in our paper is the multimodal EF-RF.

- *Global Explanation*: A global explanation method attempts to explain the reasoning of the whole model. We determine the globally most important features driving the 30-days mortality prediction of hip fracture patients by calculating the contribution of each feature in the

reduction of impurity with respect to the Gini criterion in the RF model.

- *Case Specific Explanation*: Although a global explanation gives an idea about importance of variables, it does not suffice to understand the why of a certain outcome for an individual patient. Therefore, we also employ the case-specific explanation method *Tree Interpreter* [15] which calculates the contribution of variables in the decision path for the prediction of a particular class: the contribution of a variable at a decision node is the change in probability of the instance for the class. For each variable, the average of its contributions is taken over all trees of the random forest.
- *Grad-CAM*: We additionally attempt to explain the extracted features from the image modalities. We use Gradient weighted - Class Activation Mapping (Grad-CAM) [16] to understand the most important features from the CNN models. Grad-CAM uses the gradients of any target class with respect to the final convolutional layer to highlight the important regions responsible for the prediction. We customized the Grad-CAM implementation[1] to get a neuron-specific explanation of the layer preceding the output layer instead of the output layer.

## V. Experimental Setup

In this section, we describe the experimental setup of the following models: (i) the unimodal models, (ii) the multimodal model for finding the contribution of different modalities, (iii) various multimodal models for finding the best design for multimodal fusion, (iv) the baseline model, and (v) the explainability experiments to extract the importance of features from our structured and image modalities. We used python, sklearn packages and tensorflow/keras for implementation. The code can be found in the github repository.[2]

The **hyperparameter** values of the deep learning models were set as follows. The maximum number of training epochs were set as 100, with a patience of 10 epochs for early stopping. We trained with the adam optimizer with an adaptive learning rate, starting at 0.001 and then reducing by a factor of 0.05 with a patience epoch of 5. Our model minimized the binary cross entropy loss between the predicted and the true label. We used a batch size of 10 for training.

### A. Experimental Setup Unimodal Learning (UML)

*1) Structured Modality Experiments:* As a significant amount of missing values existed in the structured modality, we applied the missing value imputation method, Iterative Imputation with a KNeighbors Regressor.[3] We validated these classifiers by 5-fold stratified cross-validation. We did hyperparameter optimization on the best classifier (random forest) with a cross-validated grid-search. We call this model, optimized random forest (O-RF). The optimal hyperparameters for the random forest classifier were: number of trees 50, bootstrap true, minimum sample split 40, split criterion as gini impurity, maximum depth 5 and maximum leaf nodes 100.

*2) Image Modality Experiments:* We wanted to compare transfer learning methods to investigate the transferability of the features of a pre-trained model to the medical images. The motivation being deep learning models learn generic features like edges, corners in the early layers and task specific features in the deeper layers [17]. This can be employed to identify generic features from medical images in the early layers. We experimented with partial and full training. Full training refers to fine-tuning all the weights of a neural network whereas partial training refers to freezing some of the early layers and fine-tuning the remaining (deeper) layers of the model. The selected pre-trained models that fit the best are also used in the remaining experiments for multimodal learning.

To find the best architecture for our task, 4 pre-trained CNN networks are fine-tuned: Xception, DenseNet169, ResNet152, InceptionV3. The default adjustment made for each model was the removal of the final fully connected layer for classifying imagenet objects. We added 2 fully connected layers (we call them Dense_0 and Dense_1) having 256 and 8 neurons respectively, with ReLU activation function and 1 neuron output layer with sigmoid activation function. The number of neurons in the fully connected layers have been selected after experimenting with some values. For each model, there is a different freeze point for the partial training experiments. The layer numbers 116, 369, 483 and 249 were respectively used as the freeze point for models Xception, DenseNet169, ResNet152 and InceptionV3.

### B. Multimodal Learning: Contribution of Different Modalities

The aim of this experiment was to find contributions of different combinations of the modalities: (i) structured and hip (S+H); (ii) structured and chest (S+C); (iii) chest and hip (C+H); and (iv) structured, chest, and hip (S+C+H). The architectures for these combinations are illustrated in Fig. 1. To keep the comparison fair, we used the same fusion technique (EF-NN) in all models when combining multiple modalities. For both the image modalities, the features are extracted from a pre-trained CNN model. For structured modality, the feature vector is extracted from the last fully connected (FC) layer of a NN, consisting of 3 FC layers with 16, 8, 4 neurons, respectively, with all layers having ReLU activation functions. Extracted features are concatenated horizontally and passed to *F-clf*, NN, with 1 FC layer of 4 neurons with ReLU and 1 output layer with sigmoid activation function.

### C. Multimodal Learning: Comparing Fusion Techniques

The aim of this experiment is to find the best design for multimodal fusion. For early fusion, we experimented with a fully connected neural network (EF-NN) and a random forest (EF-RF) as fusion classifier, and late fusion with logistic regression (LF-LR) on the modality combination of S+C and S+H+C. The explanation of designs are based on Fig. 1:

*1) EF-NN:* These are the same experiments from Section V-B, and are used here for the comparison with other designs.

[1] https://www.pyimagesearch.com/2020/03/09/grad-cam-visualize-class-activation-maps-with-keras-tensorflow-and-deep-learning/
[2] https://github.com/byenidogan/30_days_mortality_hip_fracture
[3] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborRegressor.html

*2) EF-RF:* Chest and hip X-rays' features are extracted via pre-trained deep learning. The output of Dense_1 layers of the CNN were taken as the extracted features, then horizontally concatenated to structured modality, and fed into the RF classifier. We further optimized this RF by considering the additional input coming from features extracted from chest and hip modalities. This optimization was done with cross-validated grid-search as when normal validation is done during RF hyperparameter search, it is very likely to overfit the validation data. Due to training and validation of the image modality in a normal way (not cross-validation), the data points used in the training set for the deep learning model, will also be present in the validation set of some folds of the cross validation. We believe there is no concern for a fair comparison due to a leak of information, because the final model is not selected based on the cross validation results, rather only the hyperparameters are selected. New hyperparameter values are as follows: number of trees 30, minimum sample leaf 2, minimum sample split 50 & maximum depth 12.

*3) LF-LR:* Final outputs of each unimodal classifier are concatenated and fed into a LR model. Unimodal classifier of structured modality is the O-RF whereas for the images, they are the best performing pre-trained CNN (cf. Table Ib). The same training set of the unimodal classifiers was used for training the LF-LR. No hyperparameter optimization was performed for the fusion classifier. The validation was done by aggregating the result of 5 models that was trained with a different seed value.

### *D. Baseline Model*

We used the Almelo Hip Fracture Score (AHFS) [3] addressing the same prediction goal as baseline model. Due to the lack of some variables used in AHFS, we adjusted it to the variables of our dataset; we call this adjusted version AHFS-a. Used variables are: age, gender, CCI score, prone to delirium, memory problems, KATZ ADL score, ASA score, pre-fracture living situation, pre-fracture mobility, cancer, hematocrit, prone to under-nutrition, unintentional loss of weight, decreased appetite, drink or tube feeding, and SNAQ score. The classifier used in AHFS-a is the Logistic regression method from sklearn library with solver "sag" and without any penalty for regularization.

### *E. Evaluation*

Since our data exhibits class imbalance, the evaluation of the models are based on the Area Under the Receiver Operating Characteristic Curve (AUC). For the deep learning experiments, we split the dataset into 50% train, 25% validation and 25% test set. Moreover, we repeat the same experiment 5 times, in which we change the seed value to add randomization and then report the mean and standard deviation of AUC of these 5 experiments. When evaluating the traditional machine learning classifiers, we apply 5-fold stratified cross-validation, although the 25% test split remains unseen during these experiments.

TABLE I: Unimodal learning results

(a) Structured modality

| Classifier | Val AUC |
|---|---|
| O-RF | **0.775±0.039** |
| RF | 0.750±0.038 |
| LR | 0.731±0.037 |
| XGB | 0.727±0.050 |
| SVM | 0.721±0.018 |
| AdaBoost | 0.552±0.026 |

(b) Chest (C) & Hip (H) X-ray modality

| Model | Training | C Val AUC | H Val AUC |
|---|---|---|---|
| Xception | full | **0.694±0.043** | 0.572±0.039 |
| DenseNet169 | full | 0.618±0.042 | 0.533±0.025 |
| ResNet152 | full | 0.605±0.023 | 0.575±0.029 |
| ResNet152 | partial | 0.605±0.060 | **0.594±0.029** |
| InceptionV3 | full | 0.534±0.038 | 0.539±0.025 |
| InceptionV3 | partial | 0.503±0.007 | 0.500±0.000 |
| Xception | partial | 0.500±0.000 | 0.517±0.025 |
| DenseNet169 | partial | 0.474±0.081 | 0.517±0.036 |

For the final model, EF-RF, we also provide other performance evaluation metrics on the test set. True positive (TP), false positive (FP), true negative (TN), false negative (FN), precision (Pr), recall (Re) or sensitivity, specificity (Sp), negative predictive value (NPV) & accuracy (Acc) based on different decision thresholds can be found in Table IV, where our positive class is 'deceased'.

$$Pr = \frac{TP}{TP+FP},\ Re = \frac{TP}{TP+FN},\ F1 = 2 * \frac{Pr * Re}{Pr + Re}$$
$$Sp = \frac{TN}{TN+FP},\ NPV = \frac{TN}{TN+FN},\ Acc = \frac{TP+TN}{N}$$

## VI. Multimodal Model & Explainability Results

In Table Ia, we show the results of 30-days mortality prediction by using only **structured modality (S)**. We report the mean AUC and the standard deviation of AUC on the validation set. We can see that random forest performs the best with a mean AUC of 0.750. Further optimization of the hyperparameters of this model increased it to 0.775 (O-RF).

In Table Ib, we present the results of 30-days mortality prediction by using **chest modality (C)**. It is clear that Xception model with full training mode performs best on this task. For the **hip modality (H)**, the best performing model is ResNet152 with partial training mode (cf. Table Ib). However, overall performance on chest modality is significantly higher than hip modality. This suggests that chest X-rays carry more relevant information than the hip modality regarding the 30-days mortality of the patients. On comparing Tables Ia and Ib, we conclude that structured modality model (AUC: 0.775) performs best among all unimodal models.

Table IIa shows the contribution of various modalities through **multimodal learning experiments on different modality combinations (MML-Modality)** - structured (S), chest (C) and hip (H) modalities. The best performing candidate of this part is the combination of structured and chest modality. However, it is observed that when we include more than 1 image modality (H+C or S+C+H) in EF-NN, there is a significant drop in the performance compared to one image modality models (S+C and S+H). This performance drop suggests that multiple image modalities do not work well together with the neural network fusion for this prediction task.

Table IIb **compares different ways of multimodal fusion (MML-Fusion)**. Although the winner of the previous stage was the combination with structured and chest modalities, we

TABLE II: Multimodal learning results

(a) Modality contribution (MML-Modality)

| Modalities | Val AUC |
|---|---|
| S+C | **0.714±0.028** |
| S+H | 0.643±0.052 |
| H+C | 0.568±0.011 |
| S+C+H | 0.551±0.063 |

(b) Fusion techniques (MML-Fusion)

| F-tech | F-clf | Modalities | Val AUC |
|---|---|---|---|
| LF | LR | S+C+H | **0.784±0.005** |
| LF | LR | S+C | 0.783±0.005 |
| EF | RF | S+C | 0.768±0.006 |
| EF | RF | S+C+H | 0.764±0.007 |
| EF | NN | S+C | 0.714±0.028 |
| EF | NN | S+C+H | 0.551±0.063 |

TABLE III: Performance of the competitive models from unimodal and multimodal learning on the test set

| Phase of experiments | Technique | Modalities | Test AUC |
|---|---|---|---|
| MML-Fusion | EF-RF | S+C+H | **0.786** |
| MML-Fusion | EF-RF | S+C | 0.774 |
| MML-Fusion | LF-LR | S+C+H | 0.780 |
| MML-Fusion | LF-LR | S+C | 0.771 |
| UML-S | O-RF | S | 0.732 |
| Baseline | AHFS-a | S | 0.717 |
| UML-C | Xception (Full) | C | 0.700 |
| UML-H | ResNet152 (Partial) | H | 0.670 |
| MML-Modality | EF-NN | S+C | 0.653 |

should not ignore the combination of all modalities in Table IIb as the reason of poor performance may be a design issue. This can be supported by the observation that on changing *F-clf*, the performance of S+C+H (0.784 using LF-LR) is similar to S+C (0.783 using LF-LR). It can also be observed that hip modality does not offer a significant contribution. The most important finding at this stage is that late fusion is better than early fusion, but the performance is very close to EF-RF.

In Table III, we present the **final results of the competitive candidates** from Tables Ia, Ib, IIa, and IIb on the same **test set**, which is unseen data from all aspects. By competitive candidates, we mean the best performing models and models which are close in performance to them. It is clear that both the EF-RF and LF-LR outperform our baseline model, AHFS-a [3]. On the test set, EF-RF outperformed LF-LR unlike the observation on the validation set. Further, S+C+H outperformed the S+C on the test set. Combining the observation from the validation set (cf. Table IIb) and the test set (cf. Table III), we cannot say for certain that hip X-ray modality does not add any extra information. However, it can be observed that structured and chest X-ray modality both contain enough important information for prediction of 30-days mortality as their performance is close to S+C+H.

We evaluated **our final model EF-RF using some additional metrics** on the test set. In Fig. 2, the ROC curve is presented where the AUC is 0.786. In Table IV, performance metrics based on different decision thresholds are reported to analyze how the model could be used in the clinic. Further discussion on this table will take place in Section VII.

Although the LF-LR was slightly better than EF-RF during the validation, we based the explanation methods on the EF-RF technique for the sake of convenience when it comes to

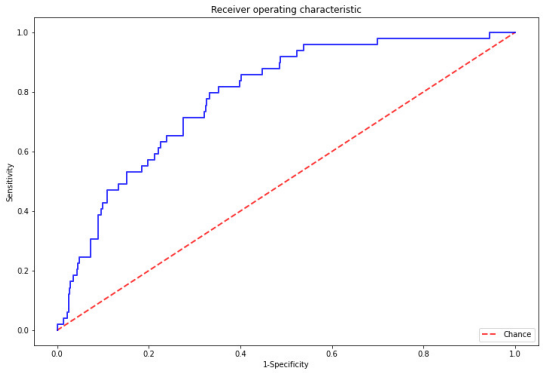


Fig. 2: ROC curve of the EF-RF model on the test set

TABLE IV: Performance metrics for different decision thresholds for EF-RF model

| Decision Threshold | TP | FP | FN | TN | NPV | Sp | Re | Pr | F1 | Acc |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 35 | 166 | 14 | 389 | 0.965 | 0.701 | 0.714 | 0.174 | 0.28 | 0.702 |
| 0.2 | 15 | 44 | 34 | 511 | 0.938 | 0.921 | 0.306 | 0.254 | 0.278 | 0.871 |
| 0.25 | 4 | 14 | 45 | 541 | 0.923 | 0.975 | 0.082 | 0.222 | 0.119 | 0.902 |
| 0.27 | 2 | 9 | 47 | 546 | 0.921 | 0.984 | 0.041 | 0.182 | 0.067 | 0.907 |
| 0.3 | 1 | 1 | 48 | 554 | 0.92 | 0.998 | 0.02 | 0.5 | 0.039 | 0.919 |
| 0.4 | 0 | 0 | 49 | 555 | 0.919 | 1 | 0 | 0 | 0 | 0.919 |
| 0.5 | 0 | 0 | 49 | 555 | 0.919 | 1 | 0 | 0 | 0 | 0.919 |

understanding in medical practice. Further, EF-RF was the best performing model on the test set, leading us to choose this model for explanation. In Fig. 3, we present an example of the **model output dashboard for local explanation** of one patient. Here, the prediction scores of involved models (top), attention points of the neurons calculated with Grad-CAM for chest X-ray (left) and hip X-ray (middle), and case specific explanation of the RF with tree-interpreter (right) are provided. The predicted probability of mortality in 30-days for this patient by our EF-RF model is 3% (i.e. the predicted probability of survival in 30-days is 97%). It is important to note that to keep the image readable, we only show the features that have an absolute contribution higher than 0.005.

In Fig. 4, we present the **global explanation**, showing the overall feature importance of the EF-RF model. Since these importance scores are calculated on the overall model, they do not change for each case.

## VII. Discussion

When we look into the Table IV, we observe that the model tends to lean towards the negative class. Even if we decrease the **decision threshold** the model can only predict few cases as positive. This is not very surprising as the positive class cases happen very rarely. To address the class imbalance problem, we experimented with various class imbalance handling methods such as oversampling, undersampling and adjusting class weights for the loss function, in the early stages of the study. However, it was found that these techniques do not have any significant contribution to the performance and therefore put them out of the scope of this paper. Moreover, we observed that the model learns to predict the negative class in a better

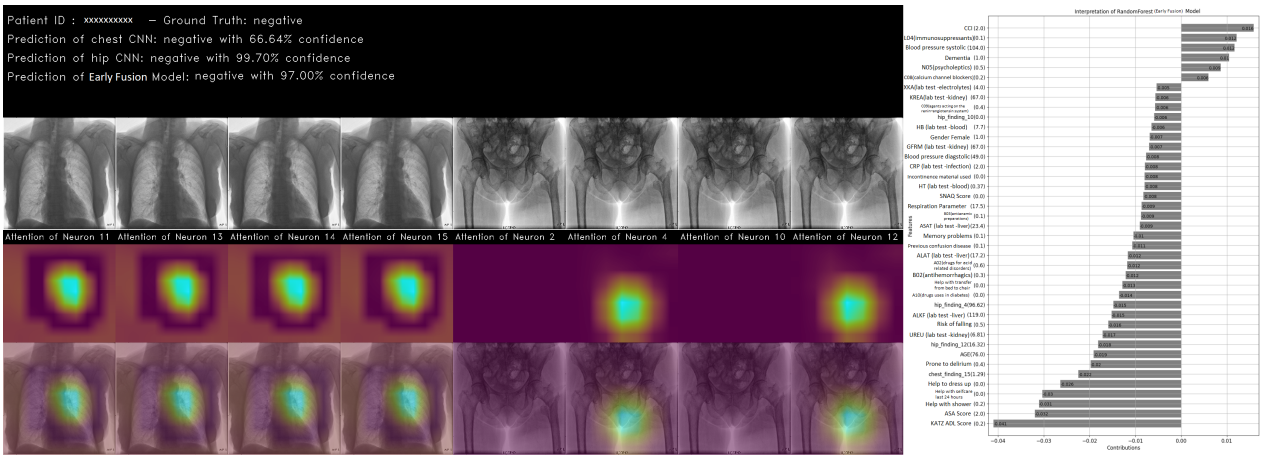


Fig. 3: Example of model output dashboard, case specific explanation of the model

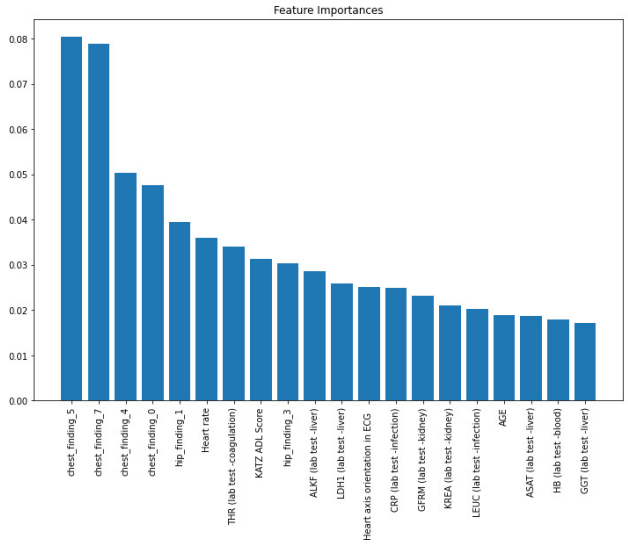


Fig. 4: Global explanation of random forest model

way. For example, if we look at the decision threshold at 0.1, we see that NPV is very high. This means that when the model predicts a case as positive with less than 10% confidence or negative with more than 90% confidence (essentially the same thing), 96.5% of the time this patient will survive. Taking the model's decision support of 90% confidence of being negative potentially can help daily practice to skip some steps which are used to find out if the patient will survive or not. In return, this might save some costs and accelerate the treatment process.

On the other hand, predicting the patients who are going to decease with high precision and recall is a very challenging task. It requires more samples of deceased patients to find out patterns amongst these cases. A possible way to improve the model in this direction is doing a study where the dataset is collected from multiple hospitals. It should be noted that when there are multiple hospitals providing data, the variables that each data source can supply will highly differ. This will result in missing values and incomplete dataset. Introducing more strict standards to the data collection process of health care industry would alleviate this problem.

Before we take a look at our explainability efforts, we feel the urge to mention that due to the dissimilarity of our methodology to the past studies in the literature with the same prediction goal, it is difficult to **compare the features that we found as important, to the ones in the literature**. The studies in the literature [3]–[7] have first applied various statistical analysis to find if a feature has significant effect on 30-days mortality of the patient in order to select their variables that they will include in their model. This kind of methodology lead them to models with only few variables whereas our model is fed with more than 100 variables. Furthermore, none of these studies have included image modality in their models. It is easily noticeable from Fig. 4, that the most important **global features** for our EF-RF model are the ones extracted from chest X-rays. Even though we cannot directly compare this finding to other studies, in some of the past studies [4], [6], radiological signs (visible in the chest X-ray) like pneumonia, enlarged contour of the heart which might suggest congestive heart failure are found to be a significant factors affecting 30-days mortality of the patient. The explanation results are also shown to a trauma surgeon, and their comments were not conclusive. Therefore, it could be that findings from chest X-rays are important due to the fact that they might be showing signs of pneumonia, but this would only be a speculation. Most of the other studies found comorbidities such as renal failure [4]–[6], diabetes [5] as important factors of 30-days mortality, however, we did not find this in our study. This could be due to incompleteness of comorbidity information in our data.

Alongside the heart rate, KATZ ADL score, heart axis orientation and age, there are also various lab tests in the top 20 most important global features of the EF-RF model. KATZ ADL score tells us about how independent a patient

is in their daily activities. If the patient is dependent in many activities, this shows sign of frailty. It should be noted that unlike general usage, in our dataset we calculate KATZ ADL in reverse, where a higher KATZ ADL score indicates that the patient is dependent (normal calculation of KATZ ADL, 6 = patient independent, 0 = patient very dependent). Therefore, a low KATZ ADL score in Fig. 3 means the patient is very independent, contributing to the negative class. A higher age indicates that the patient might be frail and is predictor for higher chance on early mortality (also found by the studies [3]–[7]).

One of important lab tests is Thrombocytes which is related to the coagulation in the blood. If a patient uses anticoagulation medication they usually have a cardiac disease (artrial fibriliation, myocardial infarction etc.) or a neurological disease (TIA = Transient Ischemic Attack or CVA = Cerebro Vascular Accident) which makes them more vulnerable. Moreover, high CRP (C-reactief proteïne) or White Blood Cells count (LEUC) might indicate some sort of an infection, high eGFR (GFRM) or Creatinine (KREA) decreased function of the kidneys. In general, when the result of these lab tests are abnormal, this might indicate a disease. Our global feature importance found these factors to be important (Fig. 4). Even though our comorbidity data are not high quality in terms of completeness, we might be partially compensating this through various lab tests. They have much less missing values and can also lead us to the signs of vulnerability of the patient.

In Fig. 3, when we look at the **case specific important features** and their effects/contributions to the final prediction, we can observe quite some similarities to what we have found with the global feature importances from Fig. 4. Most of the lab tests present themselves as important contributors to the prediction of the EF-RF model for the particular patient. If we take Haemoglobin (HB) as an example, a lower HB seems to be a predictor for higher mortality which was also observed in the study of AHFS [3]. In our example, the patient has a high HB (7.7) and this may lower the chance on early mortality. In line with that, it contributes to the prediction of the negative class. A higher Charlson Comorbidity Index (CCI) means more comorbidities, which may indicate a more frail person and therefore may lead to an increased chance on early mortality. Dementia which is a comorbidity that is counted while calculating CCI also leads to higher probability of mortality and therefore it is contributing to the prediction of the positive class as this patient had Dementia. On the other hand, a lower ASA score (in this case, 2) means a healthier patient which contributes to the prediction of negative class. In the case of KATZ ADL score, we also observe the questions used in the assessment of KATZ ADL such as help to dress up, help with selfcare, help with shower are also present in Fig. 3 individually and they suggest that patient is independent which contributes to the prediction of negative class. Analyzing the interpretation of Random Forest model makes sense in these cases. We also see that hip findings and chest findings are present amongst the important features that EF-RF used to predict the outcome for this patient. In order to make these points more transparent, we add the attention points of the neurons calculated with Grad-CAM for the image modality, so that medical specialists can relate to the chest and hip findings and do not try to interpret the value of the neuron output as this number do not make any sense for them.

A **limitation** of our dataset is the limited data quality of the structured modality. It contains many missing values, including missing comorbidities. Comorbidities were identified as important features by other studies for predicting mortality, but this was not observed in our study. For future work, we think that it can be useful to include text modality in the research. We intend to extract comorbidities and current medication from emergency department reports of hip fracture patients and include a complete list of comorbidities as features for our multimodal model. We believe that this addition might compensate to some of the information that is lacking in the current setup.

## VIII. Conclusion

We have created a multimodal model for predicting 30-days post-operative mortality of elderly hip fracture patients. To the best of our knowledge, this is the first multimodal model for this purpose, harnessing the power of structured data, containing patient characteristics, lab tests and comorbidities, and chest and hip X-ray images. Overall, we achieved a best AUC of 0.786 with early fusion with random forest multimodal model and on data modality combination of structured, chest, and hip X-rays. We outperformed our baseline (AUC: 0.717), an adjusted version of AHFS model.

Among unimodal models, structured modality achieved the best performance with an AUC of 0.732 using random forest model. Unimodal model on chest X-ray performs better than hip X-ray modality showing that chest X-ray contains more important information than hip X-ray to predict 30-days mortality. Among multimodal learning models, the best performance is achieved using all the three modalities, but our experiments on validation set showed that structured and chest X-ray modalities combined can already predict survival of a patient quite well and hip X-ray modality does not contribute much. To conclude, learning from multimodal data improves performance over single modality for 30-days mortality prediction. Our experiments on multimodality fusion techniques showed that early fusion that uses a random forest as fusion classifier achieved the best performance. However, late fusion using logistic regression had similar performance to the best performing early fusion model. From our explainability techniques, it can be seen that the important features found by our model are meaningful. Both case-based and global features show that image modalities features, especially chest X-ray, add significant value to predictions.

All in all, we believe that such a model can be of value to the medical specialists. It can be used as an early warning system indicating the patients for whom extra attention should be given, e.g., accurate monitoring of vital functions, preventive medication before, during, and after surgery or changes in treatment strategy. The model explanation could

play an important role here, but advancement in explainable machine learning is still definitely needed. For future work, we furthermore intend to attempt to predict the various complications that occur re-casting the problem into a multi-class problem, while we also address the class imbalance problem in a more focused way. The main takeaway message of this paper is, however, clear: a multimodal machine learning model can significantly exploit the additional information from other modalities.

## ACKNOWLEDGMENT

We would like to thank the Hospital Group Twente, Almelo for providing us with the data and the medical specialists with the clinical knowledge. We would like to thank Wieke S. Nijmeijer and Ellis C. Folbert for helping us with data understanding, re-coding of categorical variables and metadata.

## REFERENCES

[1] B. Gullberg, O. Johnell, and J. A. Kanis, "World-wide projections for hip fracture," *Osteoporosis International*, vol. 7, no. 5, pp. 407–413, 1997.

[2] F. Hu, C. Jiang, J. Shen, P. Tang, and Y. Wang, "Preoperative predictors for mortality following hip fracture surgery: A systematic review and meta-analysis," *Injury*, vol. 43, no. 6, pp. 676–685, 6 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138311002117

[3] W. S. Nijmeijer, E. C. Folbert, M. Vermeer, J. P. Slaets, and J. H. Hegeman, "Prediction of early mortality following hip fracture surgery in frail elderly: The Almelo Hip Fracture Score (AHFS)," *Injury*, vol. 47, no. 10, pp. 2138–2143, 10 2016.

[4] J. Karres, N. Kieviet, J.-P. Eerenberg, and B. C. Vrouenraets, "Predicting Early Mortality After Hip Fracture Surgery," *Journal of Orthopaedic Trauma*, vol. 32, no. 1, pp. 27–33, 1 2018. [Online]. Available: http://Insights.ovid.com/crossref?an=00005131-201801000-00006

[5] C. L. van de Ree, T. Gosens, A. H. van der Veen, C. J. Oosterbos, M. W. Heymans, and M. A. de Jongh, "Development and validation of the Brabant Hip Fracture Score for 30-day and 1-year mortality." *Hip international : the journal of clinical and experimental research on hip pathology and therapy*, p. 1120700019836962, 3 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30912455

[6] H. X. Jiang, S. R. Majumdar, D. A. Dick, M. Moreau, J. Raso, D. D. Otto, and D. W. C. Johnston, "Development and Initial Validation of a Risk Score for Predicting In-Hospital and 1-Year Mortality in Patients With Hip Fractures," *Journal of Bone and Mineral Research*, vol. 20, no. 3, pp. 494–500, 2005. [Online]. Available: https://asbmr.onlinelibrary.wiley.com/doi/abs/10.1359/JBMR.041133

[7] M. J. Maxwell, C. G. Moran, and I. K. Moppett, "Development and validation of a preoperative scoring system to predict 30 day mortality in patients undergoing hip fracture surgery," *BJA: British Journal of Anaesthesia*, vol. 101, no. 4, pp. 511–517, 8 2008. [Online]. Available: https://doi.org/10.1093/bja/aen236

[8] T. C. Marufu, S. M. White, R. Griffiths, S. R. Moonesinghe, and I. K. Moppett, "Prediction of 30-day mortality after hip fracture surgery by the Nottingham Hip Fracture Score and the Surgical Outcome Risk Tool," *Anaesthesia*, vol. 71, no. 5, pp. 515–521, 5 2016. [Online]. Available: http://doi.wiley.com/10.1111/anae.13418

[9] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[10] H. I. Suk, S. W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 11 2014.

[11] K. Er, A. U. Rajendra, N. Kannathal, and L. C. Min, "Data fusion of multimodal cardiovascular signals," in *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 7 VOLS. Conf Proc IEEE Eng Med Biol Soc, 2005, pp. 4689–4692.

[12] R. Bramon, I. Boada, A. Bardera, J. Rodríguez, M. Feixas, J. Puig, and M. Sbert, "Multimodal data fusion based on mutual information," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1574–1587, 2012.

[13] N. Nunes, B. Martins, N. André da Silva, F. Leite, and M. J. Silva, "A multi-modal deep learning method for classifying chest radiology exams," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11804 LNAI. Springer Verlag, 9 2019, pp. 323–335.

[14] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[15] A. Saabas, "Tree interpreter," https://github.com/andosa/treeinterpreter, 2015.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" 2014.