Towards a Machine Learning-based Decision Support System for Dispatching Helicopters in New Zealand

Tim Rädsch
Karlsruhe Institute of Technology
timraedsch.research@gmail.com

Melanie Reuter-Oppermann
Technical University of Darmstadt
oppermann@is.tu-darmstadt.de

Dave Richards
St John New Zealand
Dave.Richards@stjohn.org.nz

Abstract

Helicopters play an important role in emergency medical service systems worldwide. In sparsely populated countries like New Zealand with long distances between hospitals, helicopters are often the best way to help critically injured patients. As helicopters are extremely costly, they should only be dispatched when really necessary. In this paper, we use data from the South Island of New Zealand to test several Machine Learning approaches and show that they can be used to support dispatchers by identifying emergencies likely to require a helicopter response. We follow a non-static dataset, as the information is successively available during an emergency, and demonstrate that even a limited approach, based only on geographic incident information, can yield an Average Precision of 94% for highlighting critical emergencies. In the latter parts of this paper, we investigate different compositions of training data to assess the impact of a potential concept drift.

1. Introduction

The healthcare sector is undergoing seismic shifts towards data driven approaches [1]. Artificial Intelligence (AI) is moving from theoretical knowledge to practise, with a steady increase of accredited AI algorithms in the USA [2]. In some healthcare areas Machine Learning (ML) based algorithms were able to surpass the average performance of medical experts, such as classification of skin cancer [3] or pleural effusion [4]. One of the main challenges is the need for high quality and large volume training data [1].

Besides medical use cases ML approaches can also help improve logistical tasks [5]. This is not only the case for hospital logistics, but for all healthcare areas, for example emergency medical services (EMS). EMS systems worldwide have the challenging task of providing fast treatment and first care to emergency patients whenever and from wherever they call. The area of Operations Research offers models and algorithms for many different planning problems arising for EMS logistics to improve patient care. An overview of EMS logistics can be found in [6] or [7].

In New Zealand, a medical emergency can be serviced either with a road ambulance or a helicopter (air ambulance), depending on the severity and location of the emergency. While the number of road ambulances far exceeds the number of helicopters, some regions in New Zealand can be accessed significantly faster by a helicopter. Helicopters can also be used to transport patients much quicker to the most appropriate hospital. Highly trained staff dispatch all helicopters in New Zealand from an Air Desk based in Auckland, manually identifying emergencies that might need a helicopter in the operation control system. As a mission using a helicopter response is significantly more expensive than using a road ambulance and impedes the helicopter from assisting in another potentially even more critical emergency, a careful dispatching is crucial. Algorithms based on ML approaches can help to identify the relevant emergencies that dispatchers at the Air Desk can then decide on. The current system is only improved if the algorithm provides an increased performance and enables the human-machine collaboration [8].

The country of New Zealand consists of three main islands, the North Island, the South Island and Stewart Island. While the majority of the population can be found on the North Island, the South Island is highly frequented by tourists and for many *emergency hot spots* driving distances to the nearest ambulance base and/or hospital are very long [9]. Therefore, an efficient use of helicopters is especially crucial and we will focus on the South Island in this work. This poses the following research question:

How can we use Machine Learning to automatically assist the dispatching of helicopters on the South Island of New Zealand?

In the scientific ML community an aversion on the application of ML to real-world problems can be detected [10]. This trend has worsen in the last eight

URI: https://hdl.handle.net/10125/70822
978-0-9981331-4-0

years [11]. Thus, in this paper we investigate several ML approaches and different input sets and compare their performance. Given this research question, this work and its findings contribute to quality improvements in EMS, for both patients as well as EMS staff.

The remainder of this work is structured as follows. In Section 2, the foundations on the New Zealand EMS system, helicopter EMS (HEMS) logistics and ML are summarized. Section 3 presents the methodology that is used. The results are shown and analyzed in depth in Section 4. The paper finishes with conclusions and an outlook on future research in Section 5.

2. Foundations

2.1. Helicopters in New Zealand's EMS System

In New Zealand, St John resources the majority (90%) of road EMS throughout the country and is responsibly for the dispatch of all air EMS. There are currently 6 helicopter bases in the South Island of New Zealand as shown in Figure 1 with 1-2 helicopters located at each base. St John manages all helicopter medical responses for all of New Zealand from the Air Desk in the Ambulance Communications Centre in Auckland. The Air Desk is staffed with two Clinical Support Officers (CSOs) 15 hours a day, seven days a week. The CSOs monitor all current emergencies in the computer aided dispatch (CAD) system that have been entered by call handlers whenever someone calls 111 and that have then been managed by dispatchers who assign ambulances to calls as necessary. The CSOs aim to identify all those emergencies that could benefit from a helicopter response by manually checking the so-called ANTS criteria (Access, Number, Time Saving and Skill) for all entries in the CAD system. In 2019, St John has responded to over 440,000 incidents, leading to a high workload for the CSOs when checking all incidents manually. The CSOs are paramedics trained to intensive care level and the six employees appointed to the role all have air sector experience. Call priority differentiates into 5 urgency categories, ranging from purple, i.e. life-threatening, to grey, which marks a time-uncritical patient transport. Further call priorities display inter hospital transport, road patient transfer services and private hires. In general, only high priority calls are served by helicopters.

2.2. EMS forecasting and HEMS logistics

In EMS logistics, ML approaches have been used primarily for demand forecasting [5, 12]. For example, Setzler et al. have tested Artificial Neural Networks (ANNs) and came to the conclusion that ANNs can be capable of producing accurate forecasts for small areas [13]. Moving Average, Artificial Neural Network, Linear Regression, and Support Vector Machine approaches have been compared by Chen and Lu and Chen et al. [14, 15]. An overview of EMS logistics can be found in [6] or [7], for example. For helicopter emergency medical services, the most studied problem considers the location of helicopters and bases. Erdemir et al. for example investigate the problem of co-locating ambulances and helicopters [16]. Røislien et al. use a coverage model to optimize helicopter locations in Norway [17]. The dispatching of helicopters has been studied only in a few publications. For example, Laatz et al. investigate criteria for dispatching helicopters to emergencies in South Africa [18]. The design of a potential decision support system for helicopter dispatching in Australia is presented by Atyeo et al. [19]. Recently, Eaton et al. have presented a review on HEMS dispatching literature from a medical point of view highlighting the need for further research on HEMS dispatching [20]. The existing literature motivates investigating the helicopter dispatching problem as well as the use of ML methods, but so far, an analysis and comparison of ML approaches to support helicopter dispatching is missing.

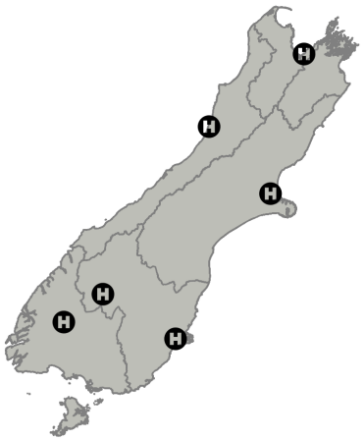


Figure 1: Location of helicopter bases on the South Island of New Zealand.

2.3. The rise of Machine Learning and the need for understandable models

Amongst others, an increase in the availability of computation power as well as data has lead to the so-called "Rise of AI" [21]. Schmidhuber provides a comprehensive overview of the history of neural networks [22]. One of the most relatable advances is autonomous driving [23]. With the recent success of AI,

the demand for understandable decisions and results is steadily growing [24, 25].

The incorporation of AI in daily business builds on the dismantling of barriers [26] and the effective human-machine collaboration [8]. Furthermore, AI needs to be easily accessible and interactive to foster adoption [27]. Thus interpretable solutions improve the cooperation with business stakeholders.

The majority of explainable AI focuses on Deep Learning due to their wide adoption which comes at the cost of the often referred "black-box" model [28]. Doran et al. differentiate between three notions of explainable AI that can be identified across research fields [29]. Adadi and Berrada distinguish between intrinsic and post-hoc explainability methods [30]. Post-hoc methods are model agnostic and generally used as an addition to ANNs. Intrinsic methods are by definition model specific and relate to their architecture. Tree based ML models can outperform Deep Learning models, in particular on tabular data sets [31]. Most classic approaches require less data and less computing. At the same time, tree based models offer a better global interpretability of how input features are translated into predictions [32]. Currently, the local interpretability of single instances is getting more attention. Those methods include the decision path, heuristic approaches as well as model-agnostic approaches [33]. Thus, we focus majorly on intrinsic methods.

3. Methodology

In the beginning of an incoming call, a minor amount of information is available, which increases successively over the duration of an emergency call. Earlier predictions could save valuable time for the treatment of time-critical patients. Thus, we are interested in the performance along the temporal availability of the features. Our system pipeline focuses on four steps and is displayed in Figure 2.

3.1. System pipeline

The goal of the first step, the data pre-processing, is to increase the quality of the data. Real world data is often flawed and the pre-processing consumes a major part of the time invested in a project [34]. We initially utilize the geographic incident information to calculate the distance to the nearest helicopter base. Afterwards, we clean our data, because faulty data can lead to inaccurate analytics and unreliable decisions [35]. Lastly, we obtain the training data set [36], by converting our features.

For the second step, we start with the selection of fitting ML models. Due to the limited availability

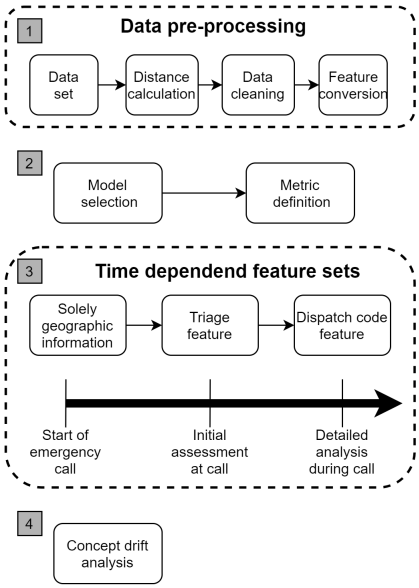


Figure 2: Overview of system pipeline.

of features at the start of an emergency we need to select models that can work with this limited amount of information. Then, we define metrics that reflect our research question.

In the third step, we follow the temporal availability of the features during an emergency call. We begin with a feature set solely comprising of geographic incident information. For the second feature set we add the triage information with the feature call priority. The third feature set explores the additional impact of the dispatch code feature and the chief complaint.

For the last step, we analyze a potential concept drift in the data set. We dissect the data set into yearly subsets to work out their impact on the performance of our selected models.

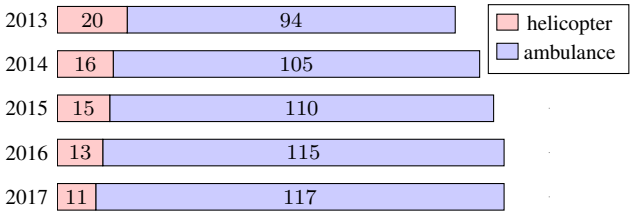
3.2. St John data set

We are using a data set of 616,718 emergency medical responses performed by St John on the South Island of New Zealand from 2013 until 2017. Figure 3b displays the yearly distribution of the responses. 540,620 (87.66%) responses were served by an ambulance and 76,098 (12.34%) by a helicopter. Thus, the data set has a moderate imbalanced distribution of their target label. From 2013 to 2017 the yearly total amount of responses increased by 12.6%. One possible explanation might be the rising number of tourists visiting New Zealand, which increased by around 40% during that period [37]. In the same time the dispatching of helicopters declined by around 44%.

For each incident 14 features are present, which can

geographic information	temporal information	descriptive information
call_location_x	call_answer_time	call_priority
call_location_y	call_coded_time	dispatch_code
District	canceled_time	ID
Territory	booked_time	incident_number
response_area	completed_time	—

(a) Initial features provided in the data set by St John.



(b) Incident type distribution from 2013 until 2017 (in thousand).

Figure 3: Overview of the rescue missions conducted by St John from 2013 until 2017.

be batched into three distinct information categories. The first category contains five features with *geographic information*, such as the location of the incident. The second category includes *temporal information*, such as the time of the call and the completion time of the incident. The last category includes four features with *descriptive information* of the emergency medical responses, such as the priority. The overview of the categories is displayed in Figure 3b.

Since St John is the sole provider of EMS, the data set combines two advantages for the further analysis of the data. It can be assumed that the data set is complete — containing all emergency medical responses in the duration 2013 until 2017 — and that it follows a coherent data format.

3.3. Distance calculation

The range of a helicopter is limited to 138.9 km. The travel time for EMS vehicles can vary significantly, depending on an *aerial route* or *terrestrial route* .

For the aerial route, we calculate the direct distance from the incident site to the nearest helicopter base. The coordinates of the call location are given in a New Zealand specific format, which uses the South Pole as their lateral reference. The longitudinal reference is shifted by 180°. The approximation into the international GPS format (WGS84) is calculated with the following formula:

$$\text{gps\_x} = 180 - \frac{\text{call\_location\_x}}{10^6}$$

$$\text{gps\_y} = -90 + \frac{\text{call\_location\_y}}{10^6}$$

The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes [38]. The new feature *distance_to_heli* is defined by the distance to the nearest helicopter base. For each emergency in the data set, the distance to all helicopter bases is calculated and the smallest distance is saved.

$$\Delta\text{lon} = lon_h - lon_e$$

$$\Delta\text{lat} = lat_h - lat_e$$

$$v = \sin\frac{\Delta\text{lon}}{2}^2 + \cos lat_e \cdot \cos lat_h \cdot \sin\frac{\Delta\text{lat}}{2}^2$$

$$\min km_h = 6371 \cdot 2 \cdot \arcsin\sqrt{v}$$

$$e = \text{call location}$$

$$h \in \{\text{helicopter\_base\_1}, ..., \text{helicopter\_base\_6}\}$$

For the terrestrial route, we multiply the aerial route with a country specific factor, accounting for the local terrain information. In the following, we refer to this factor as the ERAS distance multiplier.

3.4. Data cleaning

The data cleaning includes the identification of mislabeled instances and their treatment [39]. The methods for identification vary broadly and can include both manual and automated approaches [40, 41]. We use an automated outlier detection to remove faulty instances from the data set. In the data set some call locations are located outside of New Zealand and map to locations such as India. These instances were removed. An analysis of the value distribution for each feature reveals that the feature *canceled_time* presents the same value *Null* for every emergency response. Consequently, the feature is removed.

3.5. Feature conversion

The majority of the features are categorical features. A portion of the classical ML approaches follow an algebraic approach, which inhibits the direct usage of categorical variables [42]. Other ML algorithms function with categorical input. However in certain cases, their implementation in standard libraries does not support that [43].

The feature *dispatch_code* incorporates the meta information of the emergency in the first two digits (referred to as the chief complaint) and more granular information in the remaining part. Some codes, e.g. POLICE, use more than two digits to display the meta information. Their chief complaint, e.g. PO, is still

unique and does therefore not reduce the amount of unique dispatch code values (1710). This results in 46 different chief complaints.

Each categorical feature is encoded with one-hot encoding. Thus each feature with n categories is transformed into n-1 features. The feature *distance_to_heli* is min-max normalized. Finally, the binary target label *heli*, whether a helicopter was send, is extracted from the feature *booked_time*. To address the imbalance of the data set, we performed upsampling the minority class (helicopter), which did not improve the performance. Downsampling the majority class (ambulance) would lead to excluding potentially relevant emergencies. In addition, our metrics capture the performance of both classes.

3.6. Model selection

We work with supervised ML approaches. Starting with limited information, the available information increases during an emergency call. Thus, we need to choose algorithms that can work with limited amounts of data (see Section 2.3) and provide an interpretable architecture. We select Decision Trees, Random Forests and K nearest neighbors (KNN) due to their interpretable architecture and ability to work with limited amounts of data. Furthermore, we choose Adaboost, which generates a weighted sum of weak learners, and Multi Layer Perceptron (MLP) with five layers as an ANN approach.

Following the widely accepted recommendation, our datasplit consists of 70% training data and 30% test data. Additionally, 10% of the training data is allocated as a validation set. For our model validation we use a modified Monte Carlo cross-validation. The data set is split randomly 12 times into training and test data. For each split we fit the model and calculate our metrics. We then compute the mean average of the four best performing splits for each approach.

3.7. Metric definition

We define our metrics to ensure comparability and the informative value for the stated research problem. The F_β score is generally defined as

$$F_\beta = (1 + \beta^2) * \frac{\text{precision} * \text{recall}}{(\beta^2 * \text{precision}) + \text{recall}}$$

(for positive real β)[44].

The F_1 score represents the harmonic mean of precision and recall and is often recommended for unbalanced data sets. However, in real world scenarios a misclassified instance can incur different costs [45]. Thus, the weights for precision and recall should be defined with care. From a monetary and opportunity cost point of view, a falsely sent helicopter is more expensive than a falsely sent ambulance. A falsely sent helicopter can lead to a delayed response for consecutive emergencies. Thus, we assume that a falsely flagged helicopter by the algorithm is detected by the dispatcher. The same procedure applies for a falsely flagged ambulance. For our experiments we choose β as 0.2 to account for a higher relevance of precision over recall. Especially if the proposed approaches are used early in the process of an incoming call, we want to focus on a higher precision. Furthermore, we evaluate the precision-recall curve of the approaches using the Average Precision (AP). The AP sums-up a precision-recall curve as the weighted mean of precision values at each threshold, with the increase in recall from the previous threshold used as the weight [46]. The metrics across the different approaches are computed on the same randomly selected test set.

The implementation of the code base is written in the Python Programming Language (version. 3.7), including several standard libraries, such as numpy [47], pandas [48] and scikit-learn [49].

4. Results

4.1. Results with exclusive geographic incident information

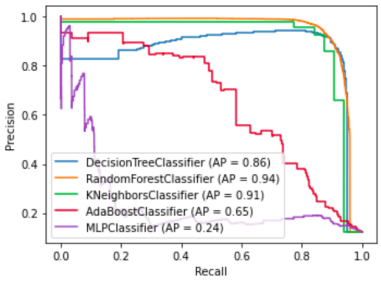
Initially, we focus on the performance of our models by only using geographic incident information that is immediately present for an emergency call. As an input we use the features district, territory and the distance to the next helicopter base. The results of the different approaches are shown in Figure 5a.

The tree based approaches and the KNN approach achieve a $F_{0.2}$ and Precision value over 0.90. For the general performance, the MLP Classifier performs the worst and did not converge. The Adaboost approach reaches a $F_{0.2}$ value of 0.8091 for the dispatching of helicopters. The KNN approach reaches the highest $F_{0.2}$ value with 0.9239. In general, simpler approaches, such as KNN and tree based approaches, perform better with limited amounts of data [50]. Figure 5b displays the precision-recall curve for the different classifiers. Taking this information into account, the Random Forest classifier reaches the highest AP of 0.94.

The results suggest that even at the early stages of an emergency, when limited information is available, KNN or tree based approaches can predict a relevant subset of the emergencies in need for a helicopter. Furthermore, the models could indicate and help to guide early attention to the potentially predicted critical

Approach	Incident type	training duration (in s)	Precision	Recall	$F_{0.2}$	AP
MLP	ambulance helicopter	317.5	0.8767 0.0000	1.0000 0.0000	0.8809 0.0000	0.24
Adaboost	ambulance helicopter	11.56	0.9180 0.8493	0.9907 0.3709	0.9206 0.8091	0.65
Decision Tree	ambulance helicopter	1.69	0.9855 0.9117	0.9878 0.8968	0.9856 0.9112	0.86
KNN	ambulance helicopter	10.30	0.9826 0.9260	0.9902 0.8752	0.9829 **0.9239**	0.91
Random Forest	ambulance helicopter	63.94	0.9853 0.9088	0.9874 0.8956	0.9854 0.9083	**0.94**

(a) Metrics for each classifier using solely geographic incident information.

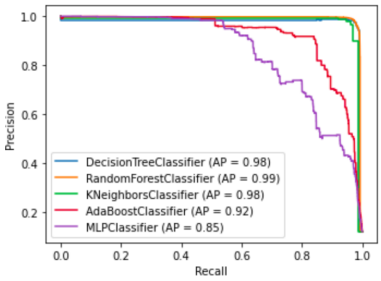


(b) Corresponding precision-recall curves.

Figure 5: Results using solely geographic incident information, which is present at the start of an emergency call.

Approach	Incident type	training duration (in s)	Precision	Recall	$F_{0.2}$	AP
MLP	ambulance helicopter	336.31	0.9513 0.8750	0.9873 0.6381	0.9526 0.8627	0.85
Adaboost	ambulance helicopter	14.66	0.9657 0.9326	0.9924 0.7480	0.9667 0.9238	0.92
Decision Tree	ambulance helicopter	2.18	0.9964 0.9789	0.9971 0.9740	0.9964 0.9787	0.98
KNN	ambulance helicopter	12.78	0.9945 0.9803	0.9973 0.9604	0.9946 **0.9795**	0.98
Random Forest	ambulance helicopter	51.22	0.9965 0.9792	0.9971 0.9747	0.9965 **0.9790**	**0.99**

(a) Metrics - improvement to relying exclusively on geographic information.



(b) Corresponding precision-recall curves.

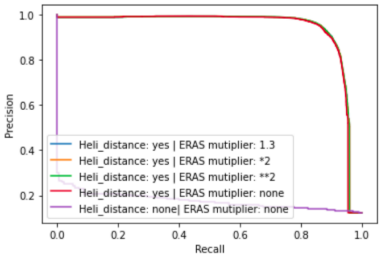
Figure 6: Results using geographic incident information and call priority, which is defined by the triage during an initial assessment in the emergency call.

Approach	Incident type	training duration (in s)	Precision meta code	Recall meta code	$F_{0.2}$ meta code	Difference to $F_{0.2}$ dispatch code	Difference to $F_{0.2}$ geographic information and call priority	AP meta code	Difference to AP geographic information and call priority
MLP	ambulance helicopter	145.25	0.9972 0.9762	0.9966 0.9800	0.9971 0.9763	0.00% 1.20%	4.46% 11.63%	0.99	0.14
Adaboost	ambulance helicopter	24.25	0.9986 0.9620	0.9944 0.9902	0.9984 0.9631	-0.10% 0.91%	3.18% 4.08%	0.99	0.07
KNN	ambulance helicopter	25.27	0.9974 0.9815	0.9974 0.9820	0.9974 0.9816	0.01% 0.04%	0.10% 0.29%	0.99	0.01
Decision Tree	ambulance helicopter	4.53	0.9970 0.9871	0.9982 0.9787	0.9970 0.9868	0.01% 0.13%	0.24% 0.73%	0.98	0.00
Random Forest	ambulance helicopter	56.91	0.9970 0.9872	0.9982 0.9790	0.9971 0.9869	0.01% 0.12%	0.06% 0.80%	0.99	0.00

Figure 7: Results using geographic information, call priority and two different dispatch code features. The latter are getting available with a detailed analysis during the call. Furthermore the performance difference to the previous figure is displayed.

Training data	Reduction of training data	Incident type	Precision	Recall	$F_{0.2}$	AP
2013- 2017	-/-	ambulance helicopter	0.9994 0.9858	0.9986 0.9934	0.9993 0.9861	-/-
2014- 2017	-21.82%	ambulance helicopter	0.9994 0.9899	0.9990 0.9934	0.9994 0.9900	0.00% 0.40%
2015- 2017	-45.10%	ambulance helicopter	0.9993 0.9930	0.9993 0.9931	0.9993 0.9930	0.00% 0.71%
2016- 2017	-69.25%	ambulance helicopter	0.9993 0.9946	0.9995 0.9930	0.9993 0.9945	0.00% 0.86%
2017	-93.86%	ambulance helicopter	0.9995 0.9960	0.9996 0.9948	0.9995 0.9960	0.02% 1.01%

(a) Random Forest classifier trained on different yearly subsets of emergencies. The oldest year is removed successively from the training pool.



(b) Impact of different ERAS multiplier on the Random Forest precision-recall curve.

Figure 8: Results of training data alternations and removed features.

emergencies. At the same time an early prediction could be used to trigger an early preparation for the helicopter takeoff. Once the necessity is confirmed, an already triggered and prepared takeoff can save valuable time for the patient in need. We did not include the features response area and the terrestrial distance in the final feature set for the training, as they did not increase the performance of the models. Regarding the response area we compared the same pipeline for the models with and without the feature. Both performed similarly. Thus, we removed the feature response area. For the terrestrial distance we evaluate the performance with different ERAS distance multipliers. Figure 8b displays the resulting precision-recall curves for a Random Forest Classifier, which is trained separately on four different ERAS distance multipliers. The resulting AP scores are almost identical with the classifier trained without the terrestrial distance approximation.

4.2. Results with the added call priority

In a next step, we add the feature call priority, which incorporates the triage based system that is an initial assessment early into an emergency call. As input we now use the features district, territory, distance to the next helicopter base and the added call priority of the emergency. The same pipeline as in the previous section is used. The results are displayed in Figure 6a. Overall, every performance metric, regardless of the approach improves. MLP and Adaboost have the largest relative performance improvement. The tree based approaches and the KNN approach achieve both a $F_{0.2}$ value over 0.97 and an AP value over 0.98. Figure 6b visualizes these AP values with an close to maxed out corner.

With the additional triage information the automated assistance proposed in the previous section could be improved. Once the triage is completed, the initial indication based on geographic incident information can get updated. Alternatively, the automated assistance could only allow predictions once the triage is completed for a given emergency. This would reduce the potential time saving for a time critical emergency, but increase the quality of the indication via the automated support system. Another possible interpretation of these results might be an unconscious bias present at the assignment of helicopters. Thus, an evaluation of the current practices for the dispatching can yield new results and provide further insights.

4.3. Results with the added dispatch codes

Now, we include the two varieties of the dispatch code as detailed emergency information that is available as a detailed analysis during the call. We explore both varieties of the dispatch code separately, which differ in their granularity of information. We first utilize the same input features as in the previous Subsection 4.2. Then, we train our models once with the dispatch code and once with the chief complaint. The results are displayed in Figure 7. They show that the overall performance improves with the usage of any dispatch code variation in comparison to the previous subsection. The MLP approach improves the most with an 11.63% increase on the $F_{0.2}$ value and an increase of 0.14 for the AP score in comparison to the approach from the previous Subsection 4.2. The Adaboost approach improves around 4%. The improvement for the remaining three approaches is marginal and below 1%. Furthermore, the performance difference between the regular dispatch code and the chief complaint is almost zero. Only the MLP approach reaches a small improvement using the finer dispatch code feature. However, the $F_{0.2}$ value of both MLP and Adaboost is still smaller than for the other approaches, with the gap decreasing.

The results suggest that using the chief complaint only marginally improves the overall performance. By choosing a KNN or Random Forest based approach with using only the geographic information and the call priority, similar results with marginally worse metrics can be achieved. For building an automated assistance for the dispatchers, the chief complaint should receive higher attention than the regular dispatch code, because it is available earlier and provides similar results. Furthermore, the results imply that the additional finer granularity of the dispatch code does not contribute much to the decision of the dispatcher.

4.4. Concept drift analysis

The surrounding conditions of the emergencies during the five year period, such as the structure of the ambulance stations or internal policies for assigning a helicopter, might have changed during that time. This so called concept drift is discussed in the literature, especially in the context of data streams [51] or predictive services [52].

We chose instance selection to address the issue of a potential concept drift, where instances are selected that represent the current concept [53]. To resemble upcoming emergencies the most, we create a test set solely containing emergencies from the most recent year 2017 with data available. The test set is defined as 75% of 2017's emergencies or 15.56% of the total number of emergencies. Comparing strategies with varying training data compositions, we initiate with the remaining data from 2013 until 2017 as our training data. For the consecutive strategies we drop the

emergencies from the oldest year present until the last strategy solely relies on the remaining 25% of the 2017 emergencies. For each strategy the same pipeline with a Random Forest Classifier and the same test set is used to ensure comparability. The results are displayed in Figure 8a.

The strategy of solely using the remaining emergencies of 2017 as input data for the training produced similar results to the other strategies utilising additional emergencies from previous years. We even observe a small increase of 1% for the $F_{0.2}$ score, while only using 6.14% of the overall data set. The given $F_{0.2}$ scores for the dispatching of the helicopters improve while the training data is reduced successively by around 20%. Given the test set being representative for upcoming emergencies, the results suggest that a concept shift happened over the five year period. Furthermore it implies that a sub-sample of last year, for which data is available, can result in similar or slightly improved predictions for upcoming emergencies. For a scenario of potential restructuring of the ambulance and helicopter infrastructure or overall policy changes this has certain implications. Using an ML based support system could be adapted most likely within a year. If the circumstances of the emergencies do not change over the years, the results imply that the proper selection of a sub-sample of the data can slightly boost the performance of the model. At the same time, the computation cost and time can be reduced due to a smaller amount of training data.

The question about the best composition of the test set remains. The test set solely based on 2017 emergencies should reflect current policies and the most recent status of the infrastructure. Additionally, it can be argued that it incorporates the seasonal fluctuations for this year. On the other side, a test set from the entire range of years might resemble the yearly fluctuation throughout the seasons more precisely. Testing on up to-date data (2020) is necessary to meet today's requirements.

5. Conclusions

5.1. Discussion and limitations

Our principle findings suggest that the dispatching of helicopters can be assisted with different ML models. From our observation criteria KNN and Random Forest achieved the highest performance. With our pre-processing we enrich the data set by adding new features and removing faulty instances. At the start of an emergency call, using only geographic incident information enriched with the haversine formula can reach a precision of over 90%. The call priority is the next available feature and increases the precision to over 97.9%. It is the largest absolute contributing factor after the start of the emergency call. The exact information about the emergency in the form of a dispatch code marginally contributes to the performance, especially if we account for the point in time when this information is available. Furthermore, our findings indicate a concept drift during the observed time period.

For a long time, academia was primarily focused on developing new architectures for ML Algorithms [22]. We now observe a trend towards training data quality. With our findings we show that classic approaches of ML, such as Random Forest and KNN, perform well if the conditions are met. The rising trend of explainable AI highlights the need for transparent decisions [24]. If a given problem can be solved with a simpler algorithm, this algorithm should be chosen. It does not always have to be the newest model architecture. Classic ML algorithms often offer the transparency that business stakeholders are looking for in particular.

Already existing data sources should be examined. This could lead to monetary savings or quality improvements for patients and staff. Furthermore, setting up proper data pipelines, in accordance with adequate security measures, might be key to automatically aggregate data for future predictions.

Initially, an automated assistance system could be implemented that assists the dispatchers and highlights emergencies that the system deems worthy of sending a helicopter. For situations in which several emergencies compete for the attention of the dispatcher, this could reduce the stress factor on the dispatcher and potentially improve the joint decision. Furthermore, the system and the approach using solely geographic information in particular could be used to alert the helicopter pilots in advance of an upcoming departure. Thus, valuable time could be saved resulting in a higher chance of patient survival. The presented models could also be extended to train new dispatchers.

We acknowledge several limitations of this research. First, the data set might exclude information that is used for determining the correct response for an emergency. The data set does not reflect the distribution and locations of ambulance bases on the South Island of New Zealand. Their travel distance is only approximated via the helicopter distance multiplied with a ground multiplier. Furthermore, we assumed no change in the policies or locations for both ambulances and helicopters over the considered time frame. The data set does not reflect the amount of helicopters available at each base and does not account for the temporal unavailability of helicopters, such as maintenance or

another active mission. We also neglect the potential impact of weather on helicopters. Second, the test set and validation set are sampled randomly from the overall data set and thus provide the same quality level as the training data. Lastly, there might exist a bias in the decision making process of helicopter dispatching. By using the data set as an input our models could learn and apply this bias as well. Finally, we assume that the Air Desks decisions are of good quality and can be used to train future decision making.

5.2. Outlook

The work presents several opportunities for further research. The given data set can be extended with further potentially relevant information, such as the locations of ambulance bases or weather input. This might yield insights into which factors contribute most to the decision making process and thus should be paid more attention. Consequently, those factors should be evaluated whether they provide medical relevant additions and do not just represent a bias. Also, the test set and the validation set offer possibilities for further research. The performance of the models is based on the assumption that these sets are correctly classified. The data quality could be improved by reevaluating these emergencies. In retrospect and given less time pressure, the decision of a dispatcher might change. Alternatively, several dispatchers could vote separately and pool their decisions, e.g. with a majority voting, to create a high quality test and validation set. Simulations could get tested as well to form a higher quality data set. In addition, intermediate solutions for changes in policies or base locations could be explored.

It remains an open question whether the concept drift occurred suddenly, e.g. due to the closure of a base, or gradually, e.g. with evolving skills of the dispatchers. Lastly, the interaction of dispatchers with the assistance system could result in new insights for the human-computer interaction field. Future research could also investigate the performance of ML methods for helicopter dispatching in other countries. As pre-existing ML methods were used together with standard input features from HEMS practice, transferring the approaches should be easy. Nevertheless, the performance cannot be easily predicted as additional input features might be necessary, for example.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42.

[2] L. D. Minor, "Stanford medicine 2020 health trends report: The rise of the data-driven physician. stanford," *Health Trends Report*, 2020.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Feb. 2017.

[4] J. Irvin, P. Rajpurkar, *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, July 2019.

[5] M. Reuter-Oppermann and N. Kühl, "Artificial intelligence for healthcare logistics: An overview and research agenda," in *Artificial Intelligence and Data Mining in Healthcare* (M. Masmoudi, B. Jarboui, and P. Siarry, eds.), Springer, 2020.

[6] M. Reuter-Oppermann, P. L. van den Berg, and J. L. Vile, "Logistics for Emergency Medical Service systems," *Health Systems*, vol. 6, no. 3, pp. 187–208, 2017.

[7] V. Blanger, A. Ruiz, and P. Soriano, "Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles," *European Journal of Operational Research*, vol. 272, no. 1, pp. 1 – 23, 2019.

[8] H. J. Wilson and P. R. Daugherty, "Collaborative intelligence: humans and ai are joining forces," *Harvard Business Review*, vol. 96, no. 4, pp. 114–123, 2018.

[9] D. Stanford, "exceptional visitors: Dimensions of tourist responsibility in the context of new zealand," *Journal of Sustainable Tourism*, vol. 16, no. 3, pp. 258–275, 2008.

[10] H. Kerner, "Too many ai researchers think real-world problems are not relevant," *MIT Technology Review*, 2020.

[11] K. Wagstaff, "Machine learning that matters," *arXiv preprint arXiv:1206.4656*, 2012.

[12] M. Reuter-Oppermann and C. Wolff, "Towards a unified understanding of data-driven support for emergency medical service logistics," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[13] H. Setzler, S. Park, and C. Saydam, "EMS call volume predictions: A comparative study," *Computers & Operations Research*, vol. 36, pp. 1843–1851, 2009.

[14] A. Y. Chen and T.-Y. Lu, "A gis-based demand forecast using machine learning for emergency medical services," *Computing in Civil and Building Engineering (2014)*, pp. 1634–1641, 2014.

[15] A. Y. Chen, T.-Y. Lu, M. H.-M. Ma, and W.-Z. Sun, "Demand forecast using data analytics for the preallocation of ambulances," *IEEE journal of biomedical and health informatics*, vol. 20, no. 4, pp. 1178–1187, 2016.

[16] E. T. Erdemir, R. Batta, P. A. Rogerson, A. Blatt, and M. Flanigan, "Joint ground and air emergency medical services coverage models: A greedy heuristic solution approach," *European Journal of Operational Research*, vol. 207, no. 2, pp. 736 – 749, 2010.

[17] J. Røislien, P. L. van den Berg, T. Lindner, E. Zakariassen, K. Aardal, and J. T. van Essen, "Exploring optimal air ambulance base locations in norway using advanced mathematical modelling," *Injury Prevention*, vol. 23, no. 1, pp. 10–15, 2017.

[18] D. Laatz, T. Welzel, and W. Stassen, "Developing a south african helicopter emergency medical service activation screen (sahas): A delphi study," *African Journal of Emergency Medicine*, vol. 9, no. 1, pp. 1 – 7, 2019.

[19] S. Atyeo, A. Sinha, and K. Young, "A decision support system for helicopter emergency medical service operations," in *ICAS 2005*, pp. 4048–4056, International Council of The Aeronautical Sciences (ICAS), 2006.

[20] G. Eaton, S. Brown, and J. Raitt, "Hems dispatch: A systematic review," *Trauma*, vol. 20, no. 1, pp. 3–10, 2018.

[21] N. Kühl, M. Goutier, R. Hirt, and G. Satzger, "Machine learning in artificial intelligence: Towards a common understanding," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015.

[23] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, pp. 163–168, IEEE, 2011.

[24] European Commission, C. a. T. Directorate General for Communications Networks, and High-Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI*. 2019. OCLC: 1128948729.

[25] A. Holzinger, "From machine learning to explainable ai," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 55–66, IEEE, 2018.

[26] T. Fountaine, B. McCarthy, and T. Saleh, "Building the ai-powered organization," *Harvard Business Review*, vol. 97, no. 4, pp. 62–73, 2019.

[27] A. Chander, R. Srinivasan, S. Chelian, J. Wang, and K. Uchino, "Working with beliefs: Ai transparency in the enterprise.," in *IUI Workshops*, 2018.

[28] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, p. 2, 2017.

[29] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

[30] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[33] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence*, vol. 2, no. 1, 2020.

[34] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, pp. 375–381, May 2003.

[35] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data Cleaning: Overview and Emerging Challenges," in *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, (San Francisco, California, USA), pp. 2201–2206, ACM Press, 2016.

[36] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised learning," *International Journal of Computer Science*, vol. 1, pp. 111–117, 01 2006.

[37] UNWTO, "Compendium of tourism statistics dataset," 2019. [Online; accessed 2020/07/15].

[38] C. C. Robusto, "The Cosine-Haversine Formula," *The American Mathematical Monthly*, vol. 64, p. 38, Jan. 1957.

[39] C. E. Brodley and M. A. Friedl, "Identifying Mislabeled Training Data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 131–167, Aug. 1999.

[40] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data preprocessing: Experiments in medical domains," *Applied Artificial Intelligence*, vol. 14, pp. 205–223, Feb. 2000.

[41] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Proceedings of 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Springer, June 2003.

[42] U. Maulik, S. Bandyopadhyay, and I. Saha, "Integrating Clustering and Supervised Learning for Categorical Data Analysis," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, pp. 664–675, July 2010.

[43] "sklearn.tree.DecisionTreeClassifier scikit-learn 0.23.0 documentation," [Online; accessed 2020/07/15].

[44] N. Chinchor, "MUC-4 evaluation metrics," in *Proceedings of the 4th conference on Message understanding - MUC4 '92*, (McLean, Virginia), p. 22, Association for Computational Linguistics, 1992.

[45] D. Hand and P. Christen, "A note on using the F-measure for evaluating record linkage algorithms," *Statistics and Computing*, vol. 28, pp. 539–547, May 2018.

[46] "sklearn.metrics.average_precision_score scikit-learn 0.23.1 documentation."

[47] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[48] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[49] Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[50] T. M. Mitchell, *Machine Learning*. McGraw-Hill series in computer science, New York: McGraw-Hill, 1997.

[51] J. Gama, I. liobait, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, pp. 1–37, Apr. 2014.

[52] L. Baier, N. Kühl, and G. Satzger, "How to cope with change?-preserving validity of predictive services over time," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[53] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.