Examensarbete inom medicinsk teknik

Avancerad nivå, 30 hp

Supervised Machine Learning for Identification of Severely Injured Trauma Patients

Övervakad maskininlärning för identifiering av svårt skadade traumapatienter

SALVAR ANDRI JÓHANNSSON

I would like to express my gratitude to my supervisor, Oscar Lapidus, and my co-supervisor, Martin Jakobsson, for granting me the opportunity to work on this project and for their guidance and support throughout its duration. I am also deeply appreciative of Maksims Kornevs for his constructive feedback during the writing of this thesis. Additionally, I extend my thanks to my fellow students, Egill and Lydia, for their insightful feedback during the final stages of my report.

I am profoundly grateful to my family and my partner, Sara, for their unwavering support, encouragement, and belief in me throughout my studies.

Abstract

English

Trauma is one of the leading causes of mortality, and is the leading cause of death among young people. The rapid activation of a trauma team for severely injured trauma patients is critical for improving patient outcomes. Severely injured patients must be rapidly identified upon arrival at the emergency department to activate a trauma team, making the need for triage tools that can rapidly and accurately assess the need for trauma team activation (TTA) important. This project explores the use of supervised machine learning models to predict the need for TTA using data from the Swedish Trauma Registry (SweTrau). The research focuses on the use of two machine learning models. A fine-tuned Large Language Model (LLM) for analyzing triage notes, and an AdaBoost ensemble model utilizing decision trees for structured data selected from the SweTrau by consulting an expert. The models were trained to predict the necessity of TTA, defined by a New Injury Severity Score (NISS) greater than 15. The results demonstrate that both models outperform the current triage practices in terms of sensitivity, ability to discern between patients severely injured patients and non-severely injured patients, with a combination of the models output offering the best performance. Although neither model is yet suitable for clinical use, the findings suggest that machine learning has the potential to enhance trauma triage protocols, particularly in identifying severe trauma cases that require TTA that could otherwise be overlooked.

Svenska

Trauma är ledande dödsorsak och är den vanligaste dödsorsaken bland unga. Aktiveringen av ett traumateam är avgörande för att förbättra utfallet för svårt skadade traumapatienter. Svårt skadade patienter måste snabbt identifieras vid ankomst till akutmottagningen, vilket kräver ett triageverktyg som snabbt och exakt kan bedöma patientens behov av traumalarm (TTA). Detta projekt utforskar användningen av övervakade maskininlärningsmodeller för att förutsäga behovet av TTA med hjälp av journaldata och register från Svenska Traumaregistret (SweTrau). Forskningen fokuserar på användningen av två maskininlärnings modeller. En finjusterad Large Language Model (LLM) för att analysera triagejournaler och en AdaBoost-ensemblemodell som använder beslutsträd för att utvärdera strukturerad data som SweTrau. Modellerna tränades för att förutsäga behovet av TTA, definierad av en New Injury Severity Score (NISS) högre än 15. Resultaten visar att båda modellerna överträffar nuvarande triage-praxis när det gäller känslighet, förmåga att skilja mellan lindrigt skadade och skadade patienter, där en kombination av modellerna gav bäst resultat. Även om ingen av modellerna ännu är lämpliga för klinisk användning, tyder fyndet på att maskininlärning har potential som triageverktyg för att identifiera patienter i behov av TTA.

Abbreviation	**Definition**
TTA	Trauma Team Activation
SBP	Systolic Blood Pressure
RR	Respiratory Rate
GCS	Glasgow Coma Scale
NISS	New Injury Severity Score
SweTrau	Swedish Trauma Registry
ED	Emergency Department
Pt	Patient
MCC	Motorcycle Crash
MVC	Motor Vehicle Crash
LLM	Large Language Model
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
SBAR	Situation, Background, Assessment, Recommendation
NaN	Not a Number

Table 1: List of Abbreviations

Contents

1 Introduction

Trauma is one of the leading causes of death in the world, being the primary cause of death among young people. Fast and effective treatment of trauma patients can be a deciding factor in preventing trauma deaths [1]. Trauma team activation (TTA) is a protocol that varies between different trauma systems, but generally consists of a preparation of a multidisciplinary team of experts that can rapidly determine the condition of a patient and decide on a appropriate course of action [2]. The implementation of trauma team activation can reduce mortality rates and improve patient outcomes, by allowing time for the multidisciplinary trauma team to assemble and prepare for the arrival of the severely injured patient [3]. In order to provide fast and effective treatment for the most severely injured patients, hospitals utilize triage protocols to quickly identify patients that are in need of urgent medical attention based on information available on arrival.

Trauma centers are hospitals that are able to provide specialized trauma care compared to non-trauma center emergency hospitals. Trauma centers are therefore better suited for the care of severely injured trauma patients and generally have better patient outcomes and lower mortality rates. Regional trauma systems are composed of multiple non-trauma center emergency hospitals and a single trauma center. In mature trauma systems, severely injured trauma patients are often redirected to be transported directly to the trauma center, and non-trauma center emergency hospitals are bypassed even if they are closer. A retrospective study on the mortality rates of trauma patients in Sweden found a substantial survival benefit for severely injured trauma patients treated at trauma centers compared to non-trauma center emergency hospitals [1].

Patients that are transported directly to the regional trauma centers are identified during pre-hospital triage. Pre-hospital undertriage is when a severely injured patient is transported to a facility that lacks the ability to provide appropriate level of care. This could pose a problem for non-trauma center emergency hospitals, as trauma patients that are easily identified as needing urgent care are likely to be transported to trauma centers, likely leaving the more obscure and occult cases to non-trauma centers. This presents a problem during in-hospital triage in identifying when procedures, such as trauma team activation, are warranted.

The New Injury Severity score (NISS) is a scoring system for the severity of trauma injuries [4]. According to national guidelines, undertriage happens when a patient with NISS > 15 does not prompt a full TTA upon arrival to the hospital [5]. Previous investigations have found that in clinical practice the adherence to national guidelines regarding trauma team activation is relatively low. The criteria for trauma team activation, which can be found in table 14, may also favor patients with apparent injuries and be unsuitable for trauma cohorts where the selection is towards patients with non-apparent injuries [6]. A study that determines the need for a in-hospital trauma team activation using artificial intelligence trained on dataset from local trauma cohort has yet to be conducted.

The aim of this thesis is to create and evaluate the feasibility and accuracy of machine learning models trained to predict the need for trauma team activation using data from the Swedish Trauma Registry (SweTrau). To achieve the aim of the thesis, the thesis goes through the following steps.

1. Prepare data from SweTrau and extract relevant features. The thesis begins with the preparation of the data available from SweTrau. This consists off deciding on what variables available in the dataset are of importance and relevant. Changing data formats to be suitable for machine learning, such as categorical data. Feature extraction and engineering.

2. Train machine learning models to predict when NISS $>$ 15. The thesis revolves around predicting the need for trauma team activation. For this thesis the need for trauma team activation was based on national guidelines as having a NISS score is greater 15. Therefore the models will be trained to predict when the NISS is > 15, which means a trauma team activation is warranted.

3. Estimate model performances. The model performances was estimated using a separate testing dataset to accurately evaluate the accuracy of the models.

4. Compare the performances with previous triage performance. The machine learning performances were compared to previous triage performances to see how the models are compare to the current triage system.

2 Background

This section will provide an overview of information relating to the thesis.

2.1 Trauma

Trauma is an tissue injury due to accident or violence. Trauma can be broadly categorized into three groups based on the mechanism of injury; penetrating trauma, blunt trauma, and deceleration trauma [7]. Trauma can vary greatly in severity and can be life threatening. In emergency rooms there are treated various conditions, including a range of trauma injuries. Different severity of trauma injuries require different hospital responses. The most severe injuries and life threatening injuries require immediate and effective treatment. In order to provide these treatments these severe injuries need to be rapidly identified during triage. Various triage tools have been developed in order to rapidly identify said patients. As this project will not be attempting to improve on the currently available triage tools, but rather to increase the accuracy when using the current tools, the only triage tools of interest for the thesis are the injury severity score scales.

2.1.1 The New Injury Severity Score

The New Injury Severity Score (NISS) is a modification on the Injury Severity Score (ISS), which are a scoring system for injury severity for predicting patient outcomes. The NISS and ISS are both based on the Abbreviated Injury Scale (AIS), where each injury is ranked on a scale from 1 (Minor) - 6 (Unsurvivable). The ISS provides an overall injury score for patients based on their injuries. Each injury is assigned an AIS score and allocated to one of six body regions based on its location. Using only the highest AIS score for each region, the three body regions with the highest AIS score have their scores squared and summed to provide the ISS. ISS scores range from 1 to 75, and a patient is assigned a score of 75 if any injury has an AIS score of 6. As the AIS score is usually assigned there can be problems with accurate assessment of the injury assessment, and inter- and intra-observer reliability. [4]

The ISS has limitations when it comes to patients that have received multiple injuries to the same body region or more then three regions, this has the results that a patients score are often underestimated, particularly for penetrating trauma. The ISS also gives equal importance to all body regions. The NISS was developed to improve upon the ISS by reducing these limitations. The NISS is calculated by squaring the three highest AIS scores, regardless of the body region. The NISS has improved performance for estimating the severity of penetrating trauma when compared to the ISS, and research suggests the same for severe blunt force trauma [8]. The NISS is, however, also affected by limitations of accurate AIS assessment by the healthcare professional [4].

2.1.2 The Swedish Trauma Registry

The Swedish Trauma Registry (SweTrau) is a registry of trauma patients in Sweden. The data is gathered from 48 hospitals, and collects data from patients that have experienced traumatic event and where a trauma alarm has been activated at the hospital, inpatients with NISS $>$ 15 even if they did not trigger a trauma alarm, and patients that have been transferred to the hospital within 7 days from the traumatic event and have a NISS $>$

15. Patients who are excluded include those who triggered a trauma alarm without experiencing a traumatic event, as well as patients whose only traumatic injury is chronic subdural hematoma, which is a collection of blood and fluid on the surface of the brain [9]. While a subdural hematoma can result from trauma, it is not always the case, and the trauma can be minor. It is likely the result of various factors [10]. The variables collected in the SweTrau registry are based on The Utstein Template, which allows for uniform collecting of data following a major trauma incident in order to allow for comparisons of trauma systems and improved prediction models [11].

The SweTrau registry contains within a wide range of data for each patient, where only a small part of which is relevant to the project based on the intended aim of the project. These include, but are not limited to, age, gender, prehospital report by paramedics, early vital signs, mechanism of injury, time of trauma call, time at arrival and departure to and from scene, time of arrival to hospital, and, of course, the NISS of the patients. The prehospital notes made by paramedics is of particular interest, as compared to the other variables which are provided as structured categorical date or numerical data, the prehospital records are in the form of free text. As the prehospital records can provide valuable information and indicators of the patients condition, a method for feature extraction from the text or a model that can use free text as a variable is needed.

The project has been approved by the Ethical Review Authority in Sweden, DNR: 2024-02482-01, for the use of trauma data. The project makes use of trauma data from 3902 individual cases registered in the SweTrau database, originating from a single hospital. The SweTrau dataset consists of a wide range of data, most of which are not of any relevance because, among other things, would not be available during triage, do not have any relevance (eg. intra-hospital patient number or date of discharge). This is done by looking at what data is available at triage and is of relevance by consulting an expert. The data that was determined to be relevant is displayed in table 2.

Pre SBP Value*	ED SBP Value*
Pre RR Value*	ED RR Value*
Pre GCS Motor*	ED GCS Motor*
Pre GCS Sum*	ED GCS Sum*
Dominant Injury*	Intention of Injury*
Mechanism Of Injury*	Type of transportation*
Pt age*	Pt Sex*
Pre Cardiac Arrest*	Trauma Alarm*
Date/time Trauma	Date/time Arrival At Scene
Date/time Leave Scene	Date/time Arrival At Hospital
Reason for visit	Type of accident
Triage notes	

Table 2: SweTrau selected variables. Pre (Pre-hospital), ED (Emergency Department), Pt (Patient), SBP (Systolic Blood Pressure), RR (Respiratory Rate), GCS (Glasgow Coma Scale), NISS (New Injury Severity Score). Utstein trauma template variables denoted by an asterisk (*).

Vital signs can be used to estimate patients condition. The Vitals consist of prehospital Systolic Blood Pressure (SBP) and Respiratory rate (RR), as well as SBP and RR measured at arrival to emergency department (ED). The SBP is measured in mmHg, left at 0 for patients in cardiac arrest, and left blank in case of missing or unknown value. The RR is measured in

breathes per minute and is left blank in case of missing or unknown values. The vital signs are measured at arrival to scene and arrival to ED, this could provide additional data in the form of change in vitals, which could be an indicator of the patient condition [12].

Glasgow Coma Scale (GCS) is a scale used to describe a patients state of consciousness. The scale assess the consciousness of the patient using responsiveness of eye-opening, motor, and verbal responses. The responses are shown, with their respective scores, in table 3. The combination of the scores can be used to provide a summary of the patients state of consciousness [13]. GCS sum is the summation of the three GCS scores score the patient on a scale of 3-15, and 999 for unknown. GCS Motor is the the motor response and is on a scale from 1-6, with 999 if the score is unknown. A value of 99 is used if the patient is under general anesthesia, is intubated, or curarized on arrival, that is has been put into a state of reduced consciousness by medical personnel. This provides an insight into the patients statues and is therefore of interest for the project. As the GCS is evaluated at scene and arrival to the hospital, the change in GCS can also proved valuable information, as well as if the patient has been put under general anesthesia, is intubated, or curarized on arrival [12].

Best eye response (4)	**Best verbal response (5)**	**Best motor response (6)**
1. No eye opening	1. No verbal response	1. No motor response
2. Eye opening to pain	2. Incomprehensible sounds	2. Abnormal extension to pain
3. Eye opening to sound	3. Inappropriate words	3. Abnormal flexion to pain
4. Eyes open spontaneously	4. Confused	4. Withdrawal from pain
	5. Oriented	5. Localizing pain
		6. Obeys commands

Table 3: Glasgow Coma Scale: Eye, Verbal, and Motor Responses [13].

Dominant Injury is the main type of injury due to the trauma, classified as either penetrating injury or blunt injury. Blunt injuries are a result of a patient collision with an outside object and penetration injury is a result of penetration of patient tissue. This information can be important when assessing the severity of the trauma. Code 999 is used when the dominant injury is unknown [12].

Intention of Injury is the human intent behind the cause of the injury, the classification shown in table 4. The intent behind the injury can provide valuable information behind the severity of the trauma [12].

Score	**Description**
1	Accident (unintentional)
2	Self-inflicted (suspected suicide, incomplete suicide attempt, or injury attempt)
3	Assault (suspected)
4	Other
999	Unknown

Table 4: Injury Intention [12].

Mechanism of injury explains the external factor that caused the incident. Nominal data that is detailed in table 5. The mechanism of injury

is important information that can help predict the severity of the patient injury [12].

Code	Description
1	Traffic: motor vehicle accident – not motorcycle (the injured patient is an occupant or passenger of a motor vehicle; i.e., car, pickup truck, van, heavy transport vehicle, bus)
2	Traffic: motorcycle accident (the injured patient is an occupant or passenger of a motorcycle)
3	Traffic: bicycle accident (the injured patient is an occupant or passenger of a bicycle)
4	Traffic: pedestrian (the injured patient is a pedestrian)
5	Traffic: other (the injured patient is an occupant or passenger of other means of transport; i.e., ship, airplane, railway train)
6	Shot by handgun, shotgun, rifle, other firearm of any calibre
7	Stabbed by knife, sword, dagger, other pointed or sharp object
8	Struck or hit by blunt object (i.e., tree, tree branch, bar, stone, human body part, metal, other)
9	Low energy fall (fall at the same level)
10	High energy fall (fall from a higher level)
11	Blast injury (the injured patient is involved in an explosion)
12	Other
999	Unknown

Table 5: Injury Type Codes and Descriptions [12].

Type of transportation shows the method of transportation which was sued to get the patient to the emergency department. The method of transportation can provide an indicator to the distance from the scene, severity of trauma, and state of the patient. The variable is not found in the Upstein Template, but was an extra variable found in the dataset.

Patient age is the patient reported age at time of injury. Numerical data in years, rounded down to the next integer (except for patients under the age of 1 year). If the data is missing, the field is left blank. Patient demographic data that can be used for training of machine learning model [12].

Patient sex is the sex of the patient. Categorical data where patients are classified as male, female, or unknown. Patient demographic data that can be used for training of machine learning model

Pre-hospital Cardiac Arrest is a variable which tells if the patient suffered an injury-related cardiac arrest before arrival to emergency department. This is an important indicator to the patients state. Code 999 if unknown [12].

Trauma Alarm provides the information on there was a TTA prior to, or upon arrival of patient to the emergency department. This provides a method to compare the performance of actual triage to models that will be developed. There are three categories, level 1 (full) TTA, level 2 (limited) TTA, and no TTA. In the context of the project, limited TTA is not considered TTA, as severely injured (NISS $>$ 15) should prompt full TTA and limited TTA is considered undertriage according to current guidelines. In this project TTA will refer to full TTA, unless specifically stated otherwise. The variable is not found in the Utstein Template, but is a combination of TTA information already prepared within the provided dataset.

The date and time variables show the time of events in the sequence from incident to transporting a trauma patient to the emergency department. The data and time can provide information on what time of day the trauma occurs, wether it is a weekend or a workday, and the time between each event can provide meaningful data. The time it takes for a patient to receive first treatment, the time it takes to finish treatment at the scene, and the time it takes to bring the patient back to the hospital. The field is left blank if the information is missing [12].

Reason for visit includes a categorized reason for the visit to the emergency room, categories shown in table 6. This can provide important information on the nature of the trauma. The variable is not found in the Upstein Template, but was an extra variable in the dataset. The categories shown in the table are the ones that were used within the dataset, but might not be a complete list of all categories. In some cases, the listed problem may not be classified as a trauma but rather as a separate reason that occurs concurrently with or potentially serves as the indirect cause of the trauma.

Abdominal pain	Nasal injury
Abdominal/pelvic injury	Nausea/vomiting
Abuse	Neck injury
Ankle injury	Neck pain
Arrhythmia	Neurological decline
Back injury	Poisoning
Back pain	Shoulder injury
Blood in urine	Social failure
Bloody stools	Speech impairment
Bone	Suicidal
Breathing problems	Thorax
Burn	Toe, finger injury
Chest injury	Trauma
Chest pain	Traumatic amputation
Confusion	Unspecified
Cramps	Upper arm injury
Dizziness	Weakness, paralysis
Elbow injury	Wound damage
Extremity swelling/pain	Wrist injury
Eye damage	Fainting
Feeling of illness	Lapse of consciousness
Flank pain	Lower leg injury
Foot injury	Lowered alertness
Forearm injury	Hypothermia
Genital injury	Inhalation of gas/smoke
Hand injury	Knee injury
Head injury	Hip injury

Table 6: Reasons for visit - Categories.

Type of accident includes the accident that caused the injury. The cause of the injury can be important in determining the severity of the trauma. The variable is not found in the Upstein Template, but was an extra variable found in the dataset.

Triage notes are notes that have information regarding the patient and situation. They are concise notes that usually follow a structured framework. The structure of the triage notes often follow SBAR, which stands for Situation, Background, Assessment, and Recommendation. The SBAR

is considered best practice when communicating information during patient handover and during critical situations [14]. An example of a SBAR communication technique in practice (altered from [14]):

> S: Trouble breathing
> B: 54 year old man with chronic lung disease who has been sliding downhill, and now he's acutely worse.
> A: No breath sounds in his right chest. Suspected pneumothorax.
> R: Might need a chest tube

Other information that can be found in the triage notes are patient allergies, if the patient has multi resistant bacteria strain, and permission for viewing of the patients medical notes. After consulting with a medical professional, it was determined that the information that would be relevant to the project was limited to the information found withing the SBAR communication technique. The SBAR should contain all the relevant information regarding each case that is available during the triage of the patient. The triage notes in the dataset provided by the SweTrau are in Swedish.

2.2 Machine learning

Machine learning is using data in order to learn patterns. Machine learning is used for various tasks in a wide range of industries. A common machine learning task is prediction, which is the task machine learning models will be used for in this project. Machine learning is often split into two paths, supervised and unsupervised learning. These paths are named after the method in which the model learns. Supervised machine learning methods are used for prediction or classification of a specific variable of interest. The model is trained using a data set where the variable of interest is known or already classified and uses that to determine a pattern for prediction or classification. The model is supervised by having a determined target variable that the model learns to predict. Unsupervised learning is, however, a machine learning method where there is not a predetermined variable the model is learning to predict, but is supposed to find relationships within the data without having to relate it to a specific target variable [15]. This project will make use of supervised learning, as the goal of the project is to improve triage by predicting the need for TTA for patients. The prediction problem laid out in the project is a binary classification problem, for which there are various well known machine learning algorithms.

2.2.1 Loss function

Loss functions are the functions for calculating the error or loss of the machine learning model output when compared to the correct or desired output. The loss function is used to improve the performance of the model by training the model to minimize the output of the loss function[16]. Depending on the task at hand and what the machine learning model is being trained to achieve, different loss functions can be used. Binary classification is a common problem within the field of machine learning, and a common loss function for training a model to achieve improved classification is cross-entropy.

Cross-entropy is a function that is commonly used as a loss function for classification problems in machine learning. Cross-entropy is a measure of the difference between distributions. In the case of machine learning, it is the difference between the true label and the predicted label. Where the true label is the true value or class of the sample being predicted, and the predicted probability distribution of the model. Minimising the difference between the two means an increase in similarities between the distributions and increased accuracy of the model.

Binary cross-entropy is the version of cross-entropy where cross-entropy is used for binary classification, as is the case with the project at hand. The formula for binary cross-entropy, where N is the number of samples, y_i is the true label at i, p_i is the predicted probability of positive label at i:

Binary cross-entropy

$$-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i)\right] \tag{1}$$

Weighted binary cross-entropy is a version of binary cross-entropy, where there have been added weights to the classes being predicted. The weights can be used to emphasize a class by making a false positive or false negative have a greater impact on the loss function by penalizing misclassification of one class more then the other. The weights can be picked depending on the need and dataset. This can be useful when training a model on datasets that do not have a balance between the two classes and have more or less of

one class. The lack of balance in the dataset can cause the model to have a bias towards the more frequent class, using the weights in the loss function can be used to counteract against the model picking up on the bias found in the dataset. The formula for weighted binary cross-entropy, where N is the number of samples, y_i is the true label at i, p_i is the predicted probability of positive label at i, and w_0 is the weight for the negative class and w_1 is the weight for the positive class:

Weighted binary cross-entropy

$$-\frac{1}{N}\sum_{i=1}^{N}\left[w_1 \cdot y_i \cdot \log(p_i) + w_0 \cdot (1-y_i) \cdot \log(1-p_i)\right] \tag{2}$$

2.2.2 Dataset split

In machine learning, the standard practice is to have three separate datasets: a training dataset, a validation dataset, and a testing dataset. The training dataset is used to train the machine learning model, the validation dataset is used to evaluate the model during training and tune parameters, and the testing dataset is used to evaluate the model's performance. The dataset for testing is kept separate from the dataset for training so that when the model is validated and tested, it will be on data the model has not seen or been trained on. This ensures that the performance evaluation can be generalized to other unseen datasets, and that the model has not been trained to only work within the specific training dataset. It is common practice to randomize the order of the dataset and split it randomly into training, validation, and testing datasets. A commonly used split is 70% of the data for training, 15% for validation, and 15% for testing. In cases where the dataset is imbalanced and small, it might be considered good practice to ensure that all datasets have similar distribution samples, as is the case in this project.

2.2.3 Validation

Validation is an important part of machine learning. Validation is used to track models progress in training and can be used to prevent the model from overfitting to the training data. Validation can be used to track the progress of training along with the loss function. The validation data is separate from the training data, and the model has not been trained on the data within the dataset. This allows tracking of performance of the model on previously unseen data, and shows if the model is overfitting or if it is learning patterns that can be generalized to previously unseen data. This allows for fine-tuning of model parameters and selection of model. There are different methods for validation, the model can be tested on parts of the validation dataset and performance metrics for the unseen data can be measured at regular intervals between periods of training using hold, allowing to track changes as the model is being trained. This allows tacking progress as the training unfolds, and see if the models starts loosing its ability to generalize. This also allows for methods that automatically select model based on validation performance and automated stopping to prevent overfitting. This method is called hold-out validation and has some drawbacks, such as being dependant on a separate dataset reducing the data available to train the model. The results of the hold-out method can also be dependent on which observations happen to be included in either training or validation datasets [17]. Other methods have been developed for validation such as k-fold validation.

k-fold validation is a validation method, in which the data samples are split into k groups, or folds, that are equal in size. One of the folds is

used as a validation dataset, and the remaining folds are used for the training of the machine learning model. After training the models performance on the validation fold is calculated. The process is repeated so that the model is retrained so that each fold gets used for validation. The average of the performance metric chosen from all the folds, providing the cross-validation estimate [18]. This method can be more costly in computations then the previously mention hold-out validation method, but it does not require separate validation and training dataset. This can be especially useful for smaller datasets. The k-fold validation method is also less sensitive to the distribution of observations like the hold-out validation method [17].

2.2.4 AdaBoostM1

Ensemble learning is a method in machine learning that makes use of combination of multiple smaller models, or "weak learners", in order to create a larger model with improved performance. One such method is adaptive boosting (AdaBoost), in which the multiple weak learners in a sequence and each model tries to correct the the errors the previous model made. This is done by fitting a model, and giving increased importance, in the form of weights, to any observation the model misclassified and the subsequent model is trained using the weights. The following models learn to better predict the previously misclassified observations. The trained weak models are then combined into a stronger model as a weighted sum of the weaker models [19]. This focuses on the harder to classify cases, which might cause issues with noisy data that the model might start overfitting. The method is however capable of producing improved accuracy when compared to single models. The method can be good for using with imbalanced datasets as it focuses on the misclassified observations which is often the underrepresented class.

2.2.5 Large language models

Large language models (LLM) are machine learning models that have the ability to perform various general tasks relating to generating and understanding languages. These models are created with deep learning methods and vast amounts of data. Large language models have shown proficiency in many language related tasks, such as creating summaries, answering questions, and analysis of texts [20]. The success of these models raises the question of their possible usefulness for healthcare and healthcare related tasks. Previous study showed that LLM can be used with medical text data for prediction of patient outcomes when researching seizure recurrence in using routine clinical notes [21]. LLM can be fine-tuned for tasks such as classification.

GPT-SW3 is a large language model developed for the Nordic languages, among them being the Swedish language [22]. Having a model that has the capabilities to understand and work with the Swedish language is of importance for the project as the free text variables in Swedish. The GPT-SW3 has multiple LLM of varying size and performance, all of which have a vocabulary of 64,000 words [22].

2.3 Performance evaluation

Receiver operating characteristic (ROC) curve is commonly used for estimating the performance of a binary classification methods. The curve represents the relationship between two parameters, *sensitivity* (true positive rate) and $1 - specificity$ (false positive rate) of a classification. Where

sensitivity, or true positive rate, is the rate for which a classification correctly identifies subjects that have the variable of interest. For this project it is the ratio of patients correctly identified as needing trauma team activation against total number of patients needing trauma team activation. Specificity is the rate at which the classification correctly identifies the subjects that lack the variable of interest, or the number of patients correctly identified as not needing trauma team activation against the total number of patients not needing trauma team activation. For the ROC curve it is not the specificity that is plotted but rather the inverse specificity, represented by $1 - specificity$, also known as false positive rate, as it provides a good visualization of the trade-offs that are common for binary classification methods. As higher sensitivity often results in lower specificity and vice versa. Sensitivity and specificity are defined in equations 3 and 4, respectively.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{Total Positives}} \tag{3}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{Total Negatives}} \tag{4}$$

ROC curves can be a helpful tool when identifying a threshold value for classification because it shows the trade-offs between the sensitivity and specificity, depending on which is more important for the task at hand. It is also useful for estimating the overall quality of the model by calculating the Area Under the Curve (AUC), where higher area means that the model has a higher capabilities of differentiating between the two classes. Figure 1 shows an explanation of an ROC curve. The red dotted line across the graph shows the line of a classifier with no ability to discern between the classes, the equivalent of a random predictions. As the curve moves towards the upper left corner the higher the capabilities of the classifier for differentiating between the two classes. A perfect classifier, having no trade-offs, being a dot in the top left corner of the graph.

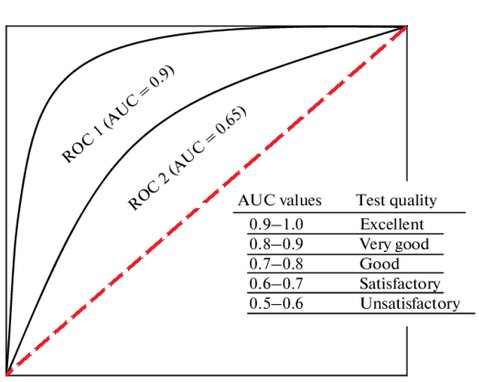

Figure 1: ROC example figure, edited from [23].

2.4 F1 metric

F1 score is an evaluation metric common in machine learning. It is the harmonic mean of precision and sensitivity (recall). Where precision is the accuracy of the positive predictions. The F1 score is proportional to true positive values so increased recall increases the F1 score but is balanced by the precision. The formula for the F1 score is shown in equation 5. [24]

$$F1 = 2 \times \frac{precision \times \text{recall}}{precision + \text{recall}} \tag{5}$$

2.5 Previous work

Machine learning models have been developed for triage of patients, often as a prediction of in-hospital mortality, critical care admission, and patient hospitalization. These model show promise as they outperform the traditional triage tools with a statistical significance. There are other benefits from utilizing machine learning models, such as reducing to variability due to physicians experience and time needed to perform triage [25]. Studies, such as [26], have successfully utilized machine learning to develop models for trauma mortality prediction, focusing on variables easily available in prehospital settings. Variables such as ability to follow commands (GCS-motor <6), age, pulse rate, SBP and, peripheral oxygen saturation were used to develop a neural network that outperformed other trauma scores available for prehospital settings [26].

Similarly, large language models (LLMs) have shown promise in clinical applications. For instance, [21] explored the predictive power of LLMs for seizure recurrence using routine clinical notes. Their findings revealed that LLMs outperformed structured data models, particularly when enhanced through domain-specific pre-training. This study shows the potential of utilizing text data, such as trauma notes, to improve outcome prediction in specialized healthcare contexts.

Further, [27] demonstrated the value of utilizing machine learning in predicting trauma team activation levels for pediatric patients. The models were trained on a extensive set of variables, such as demographic data, mechanism of injury, actual trauma activation level, medical comorbidities, interventions performed in the pre-hospital setting, pre-hospital GCS, prehospital SBP, along with other clinical scores and variables. The study outperformed actual trauma triage performance in determining the appropriate level of trauma team activation. The study made use of data from pediatric trauma patients who triggered a trauma team activation. The study made use of ISS scores > 15 and if any of specific criteria were filled as warranting full trauma team activation, otherwise it would be categorized as partial trauma team activation. This approach reflects the approach used in this project, but this project differs in the use of data from cases that should have warranted TTA but did not and the use of large language model.

Machine learning has been utilized to predict various parameters during triage, both in-hospital and pre-hospital, including trauma team activation levels. However, it has not been applied to determine the need for trauma team activation using trauma notes and data easily available upon arrival. This project will use supervised machine learning to identify patients requiring trauma team activation, applying a large language model trained on free-text trauma notes and a model trained on structured variables, to assess the feasibility and accuracy of these methods.

3 Methodology

In this section the methodology will be explained. The data is processed to be used for training of machine learning models and estimation of performance of the models and the real world triage rates. The data consist of the target variable which the models will be trained to classify, a dataset of various variables from the SweTrau database, and the triage notes. The triage notes will be used to fine-tune a large language model for the purpose of classification with regards to trauma team activation. The dataset will be used for training a separate model for the same classification task. A third model will also be trained, incorporating both the dataset and the triage notes. The performance of the models will be calculated, along with the performance of the real world triage with regards to trauma team activation. The performances will be compared, to see whether the machine learning models can perform as well, or better then the real world triage, and to see which model performed the best.

3.1 Data processing

The data consisted of 3902 individual trauma cases, split into two separate datasets. The datasets were merged into single dataset by each respective trauma case identification number. This was done to prevent possible issues with matching information between the two datasets and that the order would be the same across the datasets.

The ordinal and nominal variables were only assigned numerical labels for categories that did not already have one, without any additional preprocessing. Additional ordinal variable of change in GCS was created, this was done by subtracting the score taken at arrival to the emergency department from the one taken at the scene.

The numerical variables were normalized to the range [0,1]. Variables with a value of 999, used to indicate unknowns, were excluded from the normalization process and replaced with empty fields. Additional variables were created in the form of change in vitals, this was calculated by subtracting the values at the time of arrival from the values taken at the scene and normalized.

There were four date and time variables, each for a single event in the sequence of a injury to receiving treatment at the emergency department. They included the time and date of receiving the injury, emergency department personnel arrival to the scene, leaving the scene, and arrival to emergency department. The time between the trauma occurring and first medical assistance, the time on scene, and the time to travel to the emergency department can all be important information. This was all calculated by subtracting the minutes of the prior event from the subsequent event, this was then normalized within the range [0,1]. Other information extracted from the date and time variables include categorical data on whether the trauma occurred on a weekend, and what time of day. The time of day was categorized depending on the hour the traumas occur as "Night", "Morning","Day", or "Evening".

The data was then collected into a single dataset, to be used for training of machine learning models. The supervised machine learning models in the project aim to predict the necessity of a Trauma Team Activation (TTA) based on the severity of trauma, which is measured using the New Injury Severity Score (NISS). The NISS ranges from 0 to 75, with scores of 15 or higher indicating severe trauma. In this project, trauma cases with NISS

$>$ 15 are those that require a TTA. Therefore, the target variable for the machine learning model is binary that indicates whether the NISS score is 15 or greater, warranting a TTA (represented by a 1), or less than 15, not warranting a TTA (represented by a 0). The exact NISS value is not relevant to the project, the only interest is in whether or not a trauma case warrants TTA. The binary data was extracted from the NISS scores provided in the SweTrau database. This was used to create a target variable dataset.

The trauma team activation data is the data regarding whether a trauma team activation occurred prior to or upon patient arrival to emergency department. This data is necessary in order to allow for a comparison between the models performances and the real world triage performance in identifying severe traumas for TTA. The SweTrau dataset has a variable in which shows if TTA occurred, and if so, if it was limited or a full activation. This data was used to create a separate dataset for comparison.

Triage notes are the notes written by emergency room personnel during triage. The triage notes contain information regarding the patient, such as specific reason for hospitalization or transport, mechanism of injury, preliminary vital signs, and patient mental status. These notes are in the form of semi-structured free text, as demonstrated in 2.1.2. As the reasons for admittance to the ER can be varying and the specifics of each case are unique there are not always the same information being relayed for each patient, but rather focusing on relevant factors for each case. The notes are generally written in short and concise manner, and might make use of abbreviations. As the notes contain a lot of valuable information regarding the patient and are available during triage stages of patient admittance, they are of interest for this project. Due to the nature of the medical notes there is a need for extra processing of that particular variable, as it is not in the form of a categorical or nominal data that is normally used for machine learning. The triage notes will be used for fine-tuning a large language model for classification of the notes based on the need for TTA. The notes make use of the SBAR structure, shown in 2.1.2, which helps communicate information about the trauma case by splitting it into 'Situation', 'Background', 'Assessment', and 'Recommendation'. In the triage notes it was attempted to split the notes into their respective SBAR components in order to remove irrelevant information to the trauma (such as allergies, Multidrug-resistant bacteria, etc.) and to test different methods for fine-tuning the model based on the individual components. While the notes were structured to include these components, there were cases where information was outside any component and differences in structure which caused difficulties. The splitting of the triage notes was therefore not done, as determining the correct component for individual cases requires a trained medical professional and not ideal to remove data from the small dataset. The notes were therefore used as they were written, the only change being the removal of characters indicating a new line, '<nl>', which were replaced by a space for syntax purposes. There were two triage notes which did not have triage notes, which were removed from all datasets as the project uses the triage notes for predicting the need for TTA. The data was put into a separate dataset for the use of training the LLM.

This resulted in four separate datasets that had the same order of trauma cases. In order to use the datasets for further training and comparison, each dataset was split into three different datasets. One dataset to train the machine learning models, one dataset for validation of the model, and one dataset for final testing of performance. The split was performed with a

ratio of 0.7:0.15:0.15, where 70% of the data was for training, 15% for validation and 15% for testing. Since the data was highly biased for cases where TTA was not needed, the datasets were split using stratified splitting. The datasets were split so the relative proportion of the target variable would be equal across the split dataset. The stratified splitting was done by finding the indices for cases where NISS > 15 and NISS ≤ 15. The ratio between the two was calculated and the 0.7:0.15:0.15 split was used to calculate the number of cases needing TTA for training, validation, and testing. A random seed was set in order to allow for replication of results, and random indices from the positive (NISS > 15) and negative (NISS ≤ 15) were selected to create the training datasets. The selected indices were then removed from the pool, and this was repeated for the validation while the remainders were used to create the testing datasets. The training datasets were then randomly shuffled.

3.2 Real-world triage

For a comparison to the model performances, the instances when trauma team activation occurred were compared to when they were warranted based on the NISS > 15. The dataset from SweTrau includes a trauma team activation variable. The variable tells the extent of a TTA for each case. There is a TTA, a limited TTA, and no TTA. All severely injured patients (NISS > 15) should activate a full (level 1) TTA, so limited (level 2) TTA is considered to be undertriage according to current guidelines. Comparing the full TTA to the NISS > 15, shows the performance of the triage. The triage performance was looked at both across the entire dataset, as well as in the testing subset as that is the data the models will be evaluated by. A Confusion matrix was calculated, as well as the performance metrics in the form of sensitivity, specificity, and F1 score.

3.3 Fine-tuning of LLM

The triage notes, along with the NISS > 15 data, were used to fine-tune the GPT-Sw3 356M large language model. In order to use the triage notes for the training, validation, and testing of the model, the notes need to be in the right format. The notes were tokenized, so that they were in a format the LLM can process, and padded to the length of the longest note.

A loss function for the classification fine-tuning of the LLM was defined. A weighted binary cross-entropy loss function was selected. The loss function was chosen to minimise the similarities between the two classes and to compensate for the highly imbalanced dataset. The binary-cross entropy uses the predicted probability of the positive class (NISS being > 15) and compares to the true label, which is either 1 or 0 depending on if it's positive or not. The weights are calculated based on the distribution of the training dataset before the training, and are inverse proportional to the frequency of each class. This causes the loss function to give higher score to misclassification of the minority class and lower score for misclassification on the majority class, compensating for the imbalance. The loss function also makes use of label smoothing, where instead of using the labels integers 1 and 0, it uses $1 - \epsilon$ and ϵ. This essentially introduces noise to the prediction of the model, reducing the risk of overfitting and increasing the models ability to generalize.

The model consist of the base model, GPT-Sw3 356M, which is a pretrained LLM. The model has additional untrained layers appended to it so

it can be used for classification. The output from the LLM's classification model is fed into a custom model architecture that is further appended to the original model architecture. Experimenting with additional layers depending on performance provided the final architecture, which consists of the following layers:

1. A linear layer,
2. A batch normalization layer,
3. A ReLU activation function,
4. A 40 % dropout function,
5. A linear classification layer,
6. A sigmoid function

The linear layer provides a linear transformation to its input, which is the LLM original classifier output, allowing it to learn by adjusting its weights. The normalization layer then normalizes the features in the model, The ReLU function effectively put a threshold to the features, introducing non-linearity to the model. The dropout function randomly removes 40% of the neurons, preventing the model from being able to focus on specific features and forces it to develop a more robust method for classification allowing for improved generalization. The second linear layer reduces the inputs down to two logits, which is fed into the sigmoid function which uses the logits two calculate and outputs the probability of the positive class, which is NISS $>$ 15.

The models parameters were determined by manual testing and selected based on the model performance during training, validation, and testing, based on the loss values, AUROC, recall, and accuracy. The parameters can be seen in table 7. In order to prevent overfitting to the training data, an early stopping method was implemented so once the validation loss stops improving after 3 epochs, the training ceases and the model that produced the best loss on the validation set is used.

Parameter	Value	Description
`per_device_train_batch_size`	`2`	Training batch size per device.
`per_device_eval_batch_size`	`32`	Evaluation batch size per device.
`num_train_epochs`	`15`	Number of training epochs.
`weight_decay`	`0.02`	L2 regularization parameter to prevent overfitting.
`remove_unused_columns`	`True`	Automatically remove unused columns from the dataset during training.
`fp16`	`True`	Enable mixed precision training for faster computations and reduced memory usage.
`gradient_accumulation_steps`	`2`	Number of steps to accumulate gradients before updating the model weights.
`load_best_model_at_end`	`True`	Load the best model (based on the metric monitored) at the end of training.
`metric_for_best_model`	`"eval_loss"`	Metric used to determine the best model during training.
`greater_is_better`	`False`	Indicates whether a lower value for the metric is considered better.
`learning_rate`	`5e-5`	Initial learning rate for training.

Table 7: Training Configuration Parameters

The resulting model was tested using the testing data and had its performance evaluated using ROC curve. A threshold for the probabilities was determined by iterating through different thresholds to maximise for the F1 score. Performance metrics were calculated using the threshold and a confusion matrix was created for evaluating and comparing the model performance.

3.4 Decision tree model

A decision tree based model was trained for the purpose of classification. The reason for the use of the decision tree based model was their ability to be trained with a lot of features of varying data types with ease, as well as handling of missing values. The dataset has a lot of missing data for the features collected and the model needs to be able to handle that, as well as having both categorical and numerical data. The model uses AdaBoostM1 ensemble method for binary classification, that combines multiple decision trees into a singular model. The method makes use of multiple decision trees for predictions based on individual features and then focuses on the trees that classify incorrectly and increases the weights associated to those trees. The trees are combined into a single, better performing, classifier

that uses the trees to vote for a class based on their weights. The loss function used for training the model is an exponential loss function, with an added weight of 4 for misclassification of the minority class due to the dataset imbalance. For training the model, the dataset that were used for training and validation in the LLM classification model, were combined and shuffled into a new training dataset. The training of the model used a different validation method. The validation method was a 5-fold Cross-validation, where the new training data is split into five different parts, each part containing equal distribution of the classes. The model is trained on the four other parts and validated on one. The process is repeated for all parts, and the average is used for validation of the model, calculating AUROC, accuracy, recall. Using the performance from the validation using manual iteration, the models parameter were calibrated and can be seen in table 8.

Parameter	**Value**	**Description**
MaxNumSplits	20	The maximum number of splits in each tree.
NumLearningCycles	30	The number of decision trees in the ensemble.
LearnRate	0.1	The learning rate.

Table 8: AdaBoostM1 Parameters.

The model was then tested on the testing dataset, the ROC curve was calculated from the probabilities. A threshold for the probabilities was determined by iterating through different thresholds to maximise for the F1 score. The performance of the model was calculated with sensitivity, specificity, AUROC, and F1 score.

3.5 Combination of the models

The two models were trained using different datasets and utilized different features and information. The models were combined to see how it would affect performance. In order to make use of the training done by both models in a simple way, the average of the probability output was used to create a new probability. as shown in equation 6.

$$p_{\text{NEW}} = \frac{p_{\text{LLM classifier}} + p_{\text{AdaBoost classifier}}}{2} \tag{6}$$

The resulting probability was used to calculate a ROC curve. A threshold for the probabilities was determined by iterating through different thresholds to maximize for the F1 score. Using the selected threshold value, the models performance was calculated in the form of sensitivity, specificity, AUROC, and F1 score.

4 Results

Here the results will be presented, explained, and analyzed.

4.1 Real-world triage

Comparing the actual TTA against severely inured patients gives the following results. Figure 2 shows the Confusion matrix when NISS $>$ 15 is compared against the actual trauma triage across the entire dataset.

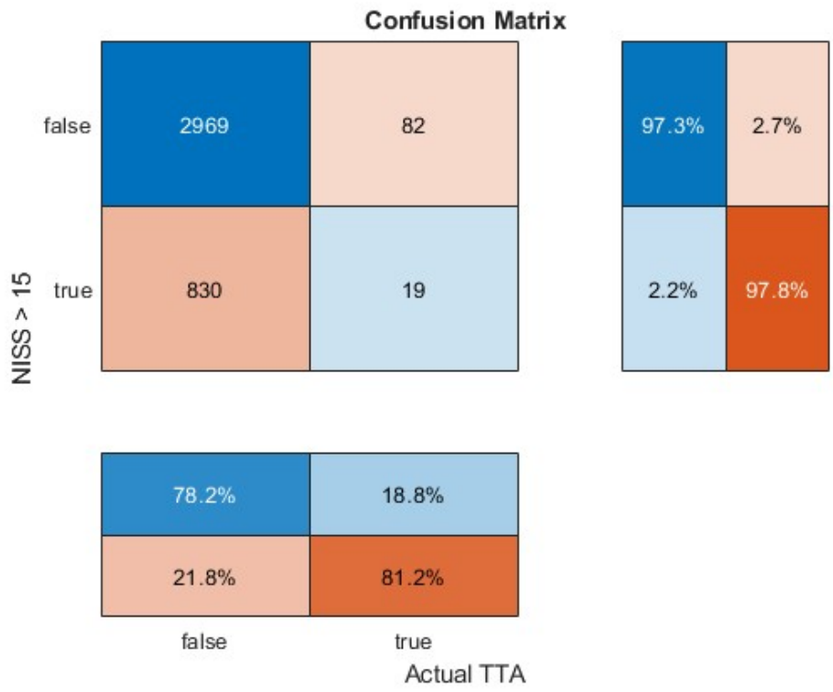

Figure 2: Trauma Team Activation (entire dataset) - Confusion matrix.

There are 3900 individual trauma cases in the dataset. The figure shows that out of the 3900 trauma cases, there have been 19 cases where the NISS $>$ 15 and a level 1 trauma team was activated, compared to 830 cases where the NISS $>$ 15 which were undertriaged and received either no or limited TTA. There were 82 cases where there was a level 1 trauma team activation for a patient that was not severely injured (NISS $<$ 15), and 2969 patients that were not severely injured and had limited TTA. That makes up 2969 non-severely injured patients and 19 severely injured patients that had the correct levels of triage. There were 830 severely injured patients that were undertriaged and 82 non-severely patients that were overtriaged. This information can be used to calculate performance metrics, shown in table 9. The table shows a relatively high accuracy and specificity, but low precision and only a recall of 2.2%. The low recall causes a low F1 score of 0.04.

Metric	**Value**
Accuracy	76.6%
Recall (Sensitivity)	2.2%
Specificity	97.3%
F1 Score	0.04

Table 9: TTA Performance Metrics

As the models will be evaluated on their test score performance, a confusion matrix for the performance of the TTA on the cases within the test dataset can be seen in figure 3.

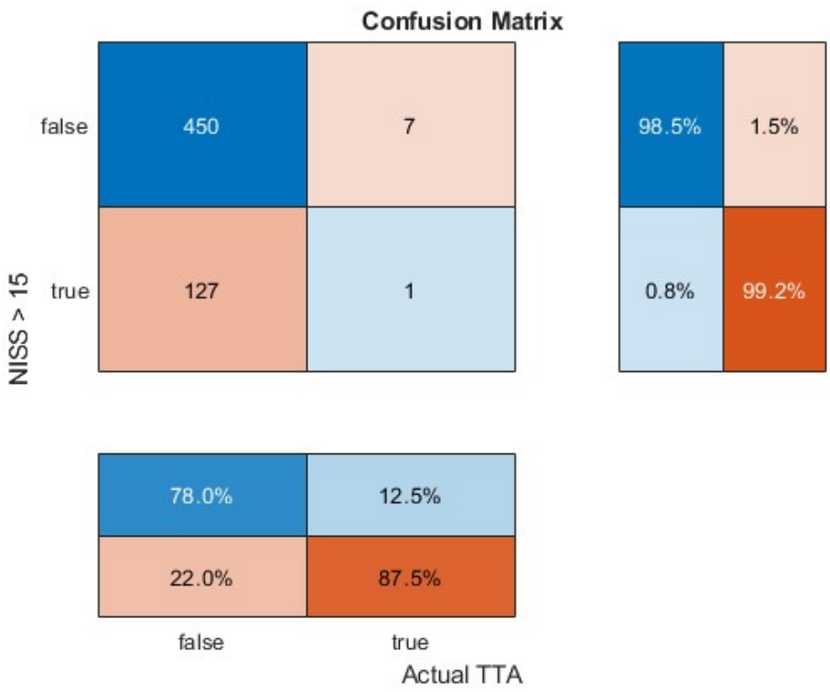

Figure 3: Trauma Team Activation (test dataset) - Confusion matrix.

The number of cases where severely injured patients that were classified correctly is 1, and 127 severely injured patients that were undertriaged and did not trigger a TTA. There were 450 patients correctly identifies as not needing full TTA, and 7 patients that were overtriaged and triggered a TTA despite not being severely injured. The performance metrics can be seen in table 10. The recall and prevision within the testing dataset are only 0.8% and 12.5%, respectively. The accuracy and specificity remain high, with the accuracy at 77.1% and the specificity at 98.5%. The F1 score is only 0.01, which is due to the low sensitivity and precision.

Metric	**Value**
Accuracy	77.1%
Recall (Sensitivity)	0.8%
Specificity	98.5%
F1 score	0.01%

Table 10: TTA Performance Metrics

4.2 Large language model

The performance of the fine-tuned LLM classification model on the testing set can be seen in the confusion matrix shown in figure 4. The matrix shows that 78.1% of severely injured patients were correctly identified as requiring a TTA, and 49.0% of patients were correctly identified as not needing a full TTA. 21.9% of the severely injured patients were undertriaged, and 51.0% of the non-severely injured patients were overtriaged.

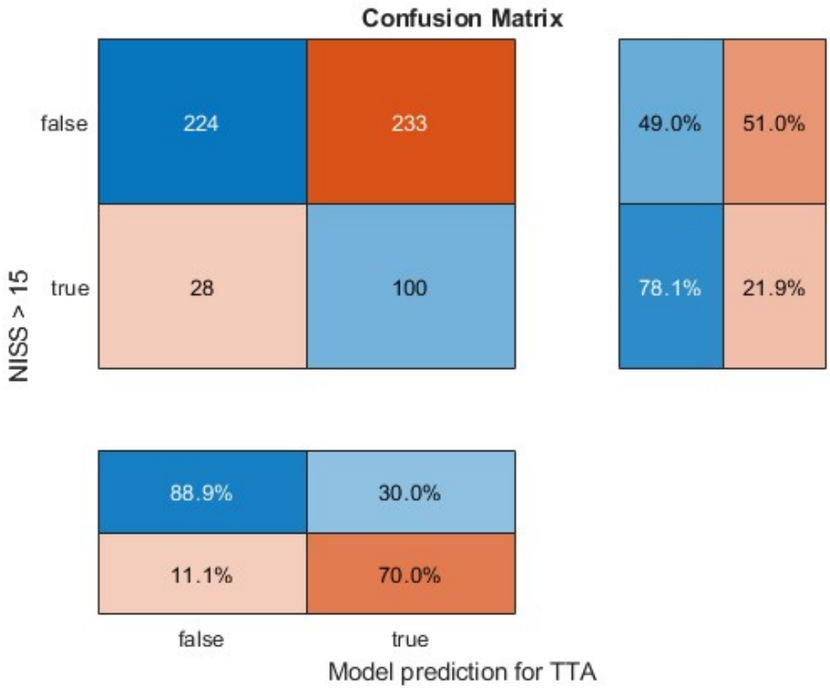

Figure 4: Fine-tuned LLM - Confusion matrix.

The output of the model is the form of probabilities of each case needing a TTA, where a threshold determines a trade-off between accuracy and recall performance. An ROC analysis can be applied to the output of the model, to determine how effective the model is at separating the different classes using AUROC. The AUROC of the LLM classification model is 0.6733, shows some ability to discern between the classes. Figure 5, shows the ROC curve. The figures shows how the model performs at the chosen threshold. The figure also shows the actual triage performance as a point of comparison. The actual triage performance having very low false positive rates (high specificity) but also very low sensitivity.

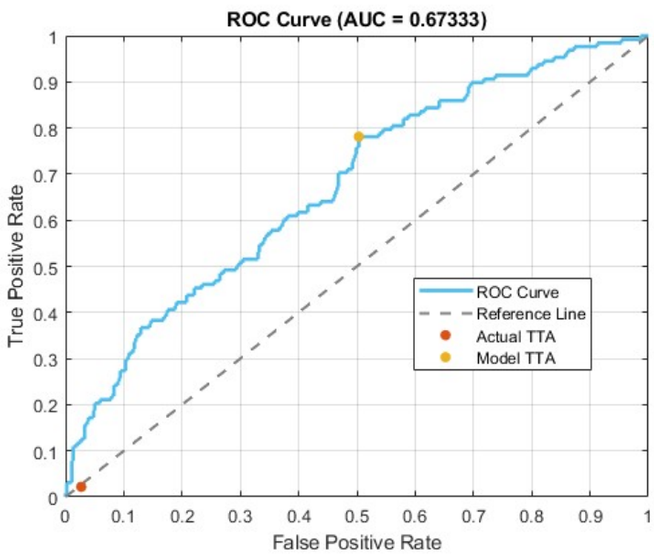

Figure 5: Fine-tuned LLM - ROC.

Table 11 shows the performance metrics for the LLM classification model. Showing a sensitivity of 78.1% and a accuracy of 55.6%. The F1 score is 0.4338, which shows the imbalance between sensitivity and precision. The model is performing better then the actual triage in terms of sensitivity,

and is also more balanced as shown by the F1 score. The model, however, performs worse in accuracy and specificity.

Metric	**Value**
Accuracy	55.4%
Recall (Sensitivity)	78.1%
Specificity	49.0%
AUROC	0.6733
F1 Score	0.4338

Table 11: LLM Performance Metrics.

4.3 AdaBoost model

The performance of the AdaBoost model is shown in figure 6. The figure shows that 75.8% of severely injured patients were classified as needing a TTA, and that 76.6% of non severely injured patients were correctly identified as not needing a full TTA. 23.4% of patients were overtriaged and 24.2% of patient were undertriaged.

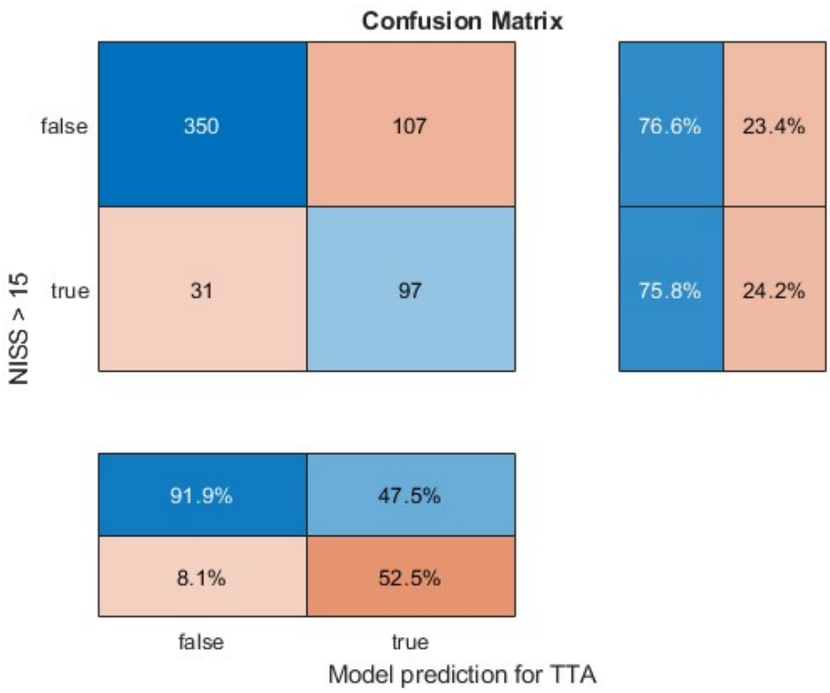

Figure 6: AdaBoost Model - Confusion matrix.

Figure 7 shows the ROC curve for the AdaBoost model, as well as the model performance at the chosen threshold. The AUROC is higher then for the LLM classifier model. The specificity is higher, as well while maintaining the same sensitivity. This shows that the model has a greater ability to discern between the classes.

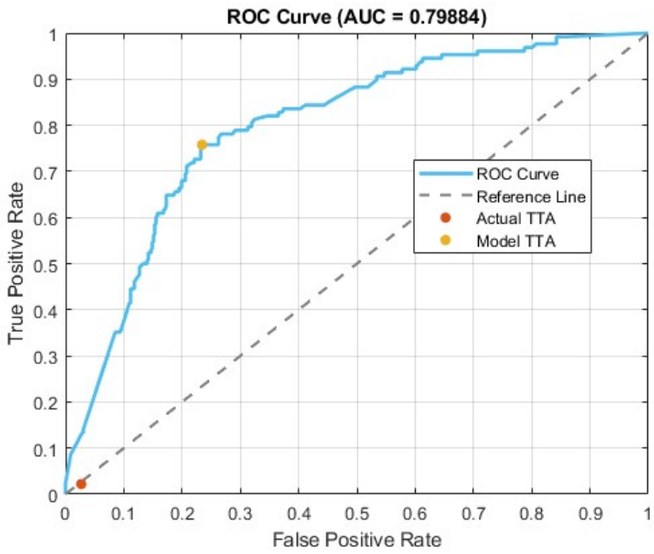

Figure 7: AdaBoost model - ROC.

The performance metrics for the AdaBoost model are shown in table 12. The table shows slightly less accuracy than the actual triage, and worse specificity. The sensitivity is the same as for the LLM classifier model without sacrificing specificity and accuracy to the same extent, which is reflected in the higher F1 score.

Metric	**Value**
Accuracy	76.4%
Recall (Sensitivity)	75.8%
Specificity	76.6%
AUROC	0.7988
F1 Score	0.5843

Table 12: AdaBoost Performance Metrics.

4.4 Combined prediction

The performance of the combined prediction can be seen in figure 8. The sensitivity is slightly better then in the standalone models, and a slightly worse specificity then the AdaBoost model, but much better then the LLM classifier model.

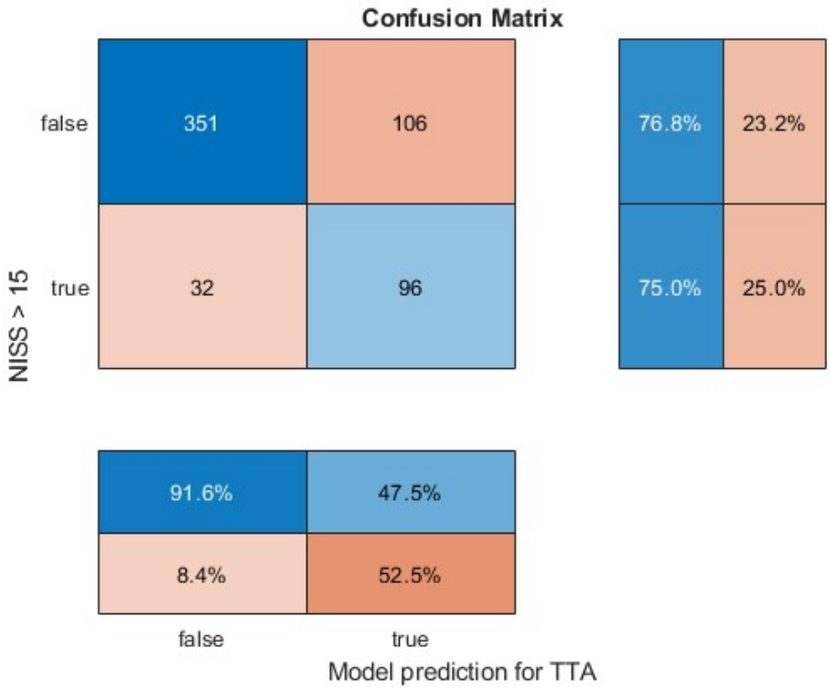

Figure 8: Combined model - Confusion matrix.

The ROC curve for the combined probabilities can be seen in figure 9. The AUROC is higher than both models, which suggests a better ability to distinguish the classes then either model by itself.

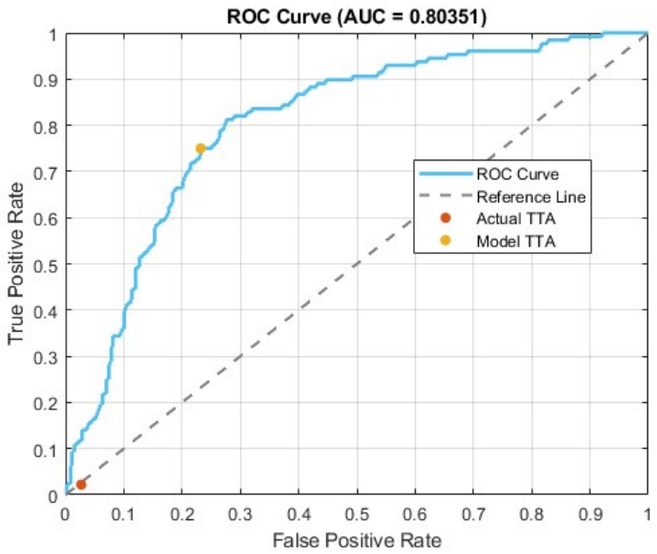

Figure 9: Combined model - ROC.

Table 13 shows the performance metrics of the combined probabilities. The accuracy is the same as for the AdaBoost model. The model has a lower sensitivity then AdaBoost, meaning it has worse ability to recognize the cases warranting a TTA. While the combination of the models probabilities is slightly worse at identifying the cases warranting TTA, it has a slightly higher specificity which would result in lower amount of false TTA.

Metric	**Value**
Accuracy	76.4%
Recall (Sensitivity)	75.0%
Specificity	76.8%
AUROC	0.8035
F1 Score	0.5818

Table 13: Decision Tree Performance Metrics.

5 Discussion

In this section the results will be analyzed and compared to the actual triage rates. The limitations of the project will be discussed, along with future work, and ethical considerations.

5.1 Model performances

Two models were trained to evaluate their performance in trauma triage, achieving an AUROC of 0.6733 for the LLM model and 0.7989 for the AdaBoost model. Combining these models resulted in a final model with an AUROC of 0.803, indicating a slightly improved performance over either model alone.

From the results it can be seen that the triage of severely injured patients could be improved. With a recall of only 2.2% means that 97.8% of severe trauma cases are undertriaged. While the accuracy is relatively high, that is due to the imbalance of the dataset. The data also shows that TTA is very conservative with only 101 TTA when there were 849 cases of severely injured patients that warranted TTA according to guidelines, which results in high specificity of 97.3%. These results show that the current triage is not very sensitive to severely injured patients, and has difficulties differentiating them from non-severely injured patients. This is also shown in the testing subset, which has a worse performance. While it is important to have a high accuracy, it is more important to be able to properly identify severely injured patients. While high accuracy caused by high specificity can lead to less wasted resources, but with low recall more patients get undertriaged which can negatively impact patient outcomes as they do not receive appropriate treatment.

The results from the LLM classification model has a decent recall of 75.0%. Which is a big increase in the recall when compared to the actual TTA. The increase comes at a cost to accuracy, as the accuracy is only 55.4% due to overtriage. The model is not as reserved in classifying TTA when compared to the actual triage, and classifies more TTAs then not. Despite the low accuracy, the model has some ability, though not great ability, to predict the need for TTA. This can be seen by the AUROC of 0.67, which means that the model has the ability to distinguish between the two classes to some extent. While the model outperforms the real triage in recall and precision, it is still not a good classification model by any standard. The model has low accuracy and high rates of overtriage. Despite this the LLM classification model has managed to find some patterns to identify trauma cases where TTA is warranted using only the triage notes. This shows that there are possibilities for trauma triage with the help of a LLM.

The Decision tree model managed to out-perform the real triage both in recall and accuracy. It does not perform as well as the LLM classification model, but it has a better balance between accuracy and recall, and higher precision meaning less overtriage. The model has a stronger ability to distinguish between the two classes as can be seen from the AUROC of 0.798. While there is room for the model to improve, it still performs decently. The model clearly shows the possibility for using supervised machine learning for trauma triage.

The use of the average of the models probabilities for classification proved to provide a slightly better model when it comes to discerning the between the classes as reflected in the AUROC score. This suggest that there might be some value in using LLM for triage as it can be used to improve the

model performance. The combined outputs of the models have a increased specificity, resulting in lower false TTAs.

While the models varied in performance, both models demonstrated an ability for triage classification. While the models do not perform well enough for clinical practice, they show that there is a possibility for further development. With more studies, improved models, and more data, they could prove to be useful for triage. The project also showed that a LLM model was able to extract features from triage notes.

5.2 Limitations

While the models managed to show ability to discern between the two classes, there are some considerations regarding the project. The project was entirely based on data from a single hospital, which can result in the models to learn things that might not be able to generalize to a broader dataset or learning unwanted patterns. The triage notes were used as they were written and sometimes included identification codes, names or numbers of emergency room personnel, and other information. When using a larger dataset from varying sources, this might not be an issue, and might even improve on the generalization of the model, as triage notes are varied, but in this context the model might be learning patterns specific to the hospital unrelated to the patients condition.

Another limitation that is worth mentioning, is that the data used for the project is all provided from the SweTrau, that comes with specific inclusion and exclusion criteria for trauma cases that are included in the database. This prevents the models to both be trained on, and to have their performance evaluated on dataset that represent the entire range of different trauma cases. The cases that are included in the database are pre-selected and exclude most lesser trauma cases, as the SweTrau database is made up of cohort with relatively high probability of being severely injured when compared to an unselected cohort of patients. This means that the effectiveness of the models in unselected emergency department cohort is unknown, and would need further research for the validation of machine learning triage tools as a real-world triage tool.

5.3 Future work

The results of the project show that there is potential for using machine learning for the triage of trauma patients. There is, however, a need for further studies before it can be actualised. The project made use of an imbalanced and relatively small dataset of only 3900 trauma cases, all of which originated from the same hospital. Having more data from varied sources will be important for future work as it would mean that there is more data to be used for training of the machine learning models, and more examples of the minority class in the dataset. Having more data could therefore increase the machine learning models performance for the classification task. The fact that the data all originates from the same hospital could also negatively impact the models ability to generalize based on the training data and might underperform on data originating from outside that particular hospital.

There is a lot of room for further improvement aside from increased data. The project trained two separate models for the classification task, one making use of the triage notes and the other using features from the SweTrau dataset. As can be seen by the combination of the model outputs, there might be values in a combination model. Creating a single model that makes use of both the triage notes and the SweTrau features could have

better performance. This could be done adding the features as inputs to a custom head for the classification from the LLM, or using the outputs of the LLM classifier as features for a different model.

The selection of a different LLM can improve results. The project made use of the GPT-Sw3 356M LLM, which is small LLM compared to many of the current LLM's. There could be an increase in performance of the classification when using a better LLM to fine-tune for the classification. This would increase the computing power and time needed for the training of the model, but could improve the performance of the model. LLM's can also be trained for particular tasks, so training a LLM specifically for triage notes or medical texts before fine-tuning for classification could produce interesting results.

There could also be an approach where a LLM is trained to extract features from triage notes. A possible method would be to train a LLM for token classification in order to identify the text tokens that represent important medical data, such as vitals or patient state of consciousness.

While the project did not produce models that would be acceptable for clinical practice, it has managed to show that there are many possibilities in using machine learning for the triage of trauma patients.

5.4 Ethical considerations

The project received an ethical permit for the access and use of patient trauma data acquired by the SweTrau database for the training of a machine learning model. Ethical permit DNR: 2024-02482-01.

All data storing, processing, and training with the data received from SweTrau was done locally in order to ensure data privacy.

There are several ethical considerations that follow any project that deals with medical information and decisions. The data used to train the models is sensitive patient data which has to be handled with care. There is the danger of a model learning to any biases of the dataset being used, that could negatively affect patient triage or care. Depending on the type of model, there could also be a problem with a lack of transparency in the decision making process of a machine learning model, such as a LLM. This could lead to lack of trust.

While there are many possible issues and ethical considerations that have to be taken into account, there can also be a lot of potential benefits. Possible improved speed and accuracy of patient triage, does not rely on the experience of the person in triage and could improve on identification of less obvious severe trauma cases.

6 Conclusion

The project aimed to create and evaluate supervised machine learning models for discerning the need for TTA during triage using data from the SweTrau database. Using variables selected by consulting an expert, two separate supervised machine learning models were trained, one using AdaBoost ensemble with decision trees, and another fine-tuning a LLM for classification. The AdaBoost model was trained using categorical and numerical data available at triage, while the LLM classification model was trained using triage notes. The LLM classification model had a lot of room for improvement, but showed some ability to be able to discern between the classes, showing a possibility for the use of LLMs for triage. The AdaBoost model performed better and had a demonstrated an ability to discern between the two classes. The output of the two models were also combined to create a new prediction that performed slightly better then the AdaBoost model by itself, indicating that there might be value in a combined classification model with an LLM. The results were compared to the actual trauma triage, which all models outperformed.

Both models produced promising results, outperforming the real-world triage when tested on unseen data. The results indicate that there is some promise in using LLMs and supervised machine learning for discerning the need for TTA using data available at triage. There is, however, a need for further development, studies, and validation. Despite outperforming the real-world triage when it came to sensitivity and ability to discern between severely injured patients and non-severely injured patients, there is a lot of room for improvement. Possible improvements include, the use of a different LLM, mixed models that makes use of all parameters, and a larger dataset and more varied dataset. The project had several limitations when it came to the data used, as it originates from a database pre-selected from trauma cases with a bias towards more severely injured patients, meaning that the performance of the models on real-world data that has not been curated is unknown. While further work is needed, the results from this project show promise for the use of machine learning in determining trauma severity during triage.

References

[1] Stefan Candefjord, Linn Asker, and Eva-Corina Caragounis. "Mortality of trauma patients treated at trauma centers compared to non-trauma centers in Sweden: a retrospective study". In: *European Journal of Trauma and Emergency Surgery* 48.1 (July 2020), pp. 525–536. ISSN: 1863-9941. DOI: 10.1007/s00068-020-01446-6. URL: http://dx.doi.org/10.1007/s00068-020-01446-6.

[2] Marcus C-K Tai, Raymond C-H Cheng, and Timothy H Rainer. "Trauma systems: Do trauma teams make a difference?" In: *Trauma* 13.4 (Oct. 2011), pp. 294–299. ISSN: 1477-0350. DOI: 10.1177/1460408611405294. URL: http://dx.doi.org/10.1177/1460408611405294.

[3] S. A. Deane et al. "IMPLEMENTATION OF A TRAUMA TEAM". In: *Australian and New Zealand Journal of Surgery* 59.5 (May 1989), pp. 373–378. ISSN: 0004-8682. DOI: 10.1111/j.1445-2197.1989.tb01589.x. URL: http://dx.doi.org/10.1111/j.1445-2197.1989.tb01589.x.

[4] M.N Chawda et al. "Predicting outcome after multiple trauma: which scoring system?" In: *Injury* 35.4 (Apr. 2004), pp. 347–358. ISSN: 0020-1383. DOI: 10.1016/s0020-1383(03)00140-2. URL: http://dx.doi.org/10.1016/S0020-1383(03)00140-2.

[5] Oscar Lapidus et al. "72103 - Undertriage of severely injured patients at one non-trauma center hospital in Stockholm". In: *British Journal of Surgery* 111.Supplement_7 (Aug. 2024). ISSN: 1365-2168. DOI: 10.1093/bjs/znae175.098. URL: http://dx.doi.org/10.1093/bjs/znae175.098.

[6] Oscar Lapidus et al. "Triage accuracy and adherence to Swedish guidelines for in-hospital trauma team activation at one non-trauma center emergency hospital in Stockholm". In: *UNPUBLISHED* X.Y (), pp. Z–ZZ.

[7] J Dumovich and P Singh. "Physiology, Trauma". In: (Jan. 2024).

[8] Hui Li and Yue-Feng Ma. "New injury severity score (NISS) outperforms injury severity score (ISS) in the evaluation of severe blunt trauma patients". In: *Chinese Journal of Traumatology* 24.5 (Sept. 2021), pp. 261–265. ISSN: 1008-1275. DOI: 10.1016/j.cjtee.2021.01.006. URL: http://dx.doi.org/10.1016/j.cjtee.2021.01.006.

[9] *Swedish Trauma Registry (SweTrau)*. Accessed on 2024-02-15. 2020. URL: http://rcsyd.se/swetrau/.

[10] Ellie Edlmann et al. "Pathophysiology of chronic subdural haematoma: inflammation, angiogenesis and implications for pharmacotherapy". In: *Journal of Neuroinflammation* 14.1 (May 2017). ISSN: 1742-2094. DOI: 10.1186/s12974-017-0881-y. URL: http://dx.doi.org/10.1186/s12974-017-0881-y.

[11] Kjetil G. Ringdal et al. "The Utstein template for uniform reporting of data following major trauma: a joint revision by SCANTEM, TARN, DGU-TR and RITG". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 16.1 (2008), p. 7. ISSN: 1757-7241. DOI: 10.1186/1757-7241-16-7. URL: http://dx.doi.org/10.1186/1757-7241-16-7.

[12] Swetrau. *Revised Utstein Template – Data Dictionary v1.1.1*. Accessed: 2024-08-26. Jan. 2021. URL: https://rcsyd.se/swetrau/wp-content/uploads/sites/10/2021/01/Revised-Utstein-Template-%E2%80%93-Data-Dictionary-v1.1.1.pdf.

[13] Shivani Jain and Linda M. Iverson. *Glasgow Coma Scale*. [Updated 2023 Jun 12]. Treasure Island (FL): StatPearls Publishing, 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK513298/.

[14] Martin Müller et al. "Impact of the communication and patient hand-off tool SBAR on patient safety: a systematic review". In: *BMJ Open* 8.8 (Aug. 2018), e022202. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2018-022202. URL: http://dx.doi.org/10.1136/bmjopen-2018-022202.

[15] Tammy Jiang, Jaimie L. Gradus, and Anthony J. Rosellini. "Supervised Machine Learning: A Brief Primer". In: *Behavior Therapy* 51.5 (Sept. 2020), pp. 675–687. ISSN: 0005-7894. DOI: 10.1016/j.beth.2020.05.002. URL: http://dx.doi.org/10.1016/j.beth.2020.05.002.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. ISBN: 978-0387310732.

[17] Umberto Michelucci. *Fundamental Mathematical Concepts for Machine Learning in Science*. Cham, Switzerland: Springer Nature Switzerland AG, 2024. ISBN: 978-3-031-56430-7. DOI: 10.1007/978-3-031-56431-4.

[18] Gareth James et al. *An Introduction to Statistical Learning, With Applications in Python*. 1st. Springer Nature Switzerland AG, 2023. ISBN: 978-3-031-38747-0. DOI: 10.1007/978-3-031-38747-0. URL: https://www.statlearning.com.

[19] Ammar Mohammed and Rania Kora. "A comprehensive review on ensemble deep learning: Opportunities and challenges". In: *Journal of King Saud University - Computer and Information Sciences* 35.2 (Feb. 2023), pp. 757–774. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2023.01.014. URL: http://dx.doi.org/10.1016/j.jksuci.2023.01.014.

[20] Muhammad Usman Hadi et al. "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects". In: (Nov. 2023). DOI: 10.36227/techrxiv.23589741. URL: http://dx.doi.org/10.36227/techrxiv.23589741.

[21] Brett K Beaulieu-Jones et al. "Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study". In: *The Lancet Digital Health* 5.12 (Dec. 2023), e882–e894. ISSN: 2589-7500. DOI: 10.1016/s2589-7500(23)00179-6. URL: http://dx.doi.org/10.1016/S2589-7500(23)00179-6.

[22] Ariel Ekgren et al. *GPT-SW3: An Autoregressive Language Model for the Nordic Languages*. 2023. DOI: 10.48550/ARXIV.2305.12987. URL: https://arxiv.org/abs/2305.12987.

[23] O.P. Trifonova, P.G. Lokhov, and A.I. Archakov. "Metabolic profiling of human blood". In: *Biomeditsinskaya Khimiya* 60.3 (2014), pp. 281–294. ISSN: 2310-6972. DOI: 10.18097/pbmc20146003281. URL: http://dx.doi.org/10.18097/pbmc20146003281.

[24] Steven A. Hicks et al. "On evaluation metrics for medical applications of artificial intelligence". In: *Scientific Reports* 12.1 (Apr. 2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-09954-8. URL: http://dx.doi.org/10.1038/s41598-022-09954-8.

[25] Oluwasemilore Adebayo, Zunira Areeba Bhuiyan, and Zubair Ahmed. "Exploring the effectiveness of artificial intelligence, machine learning and deep learning in trauma triage: A systematic review and meta-analysis". In: *DIGITAL HEALTH* 9 (Jan. 2023). ISSN: 2055-2076. DOI: 10.1177/20552076231205736. URL: http://dx.doi.org/10.1177/20552076231205736.

[26] Yun Li et al. "Development and Validation of a Simplified Prehospital Triage Model Using Neural Network to Predict Mortality in Trauma Patients: The Ability to Follow Commands, Age, Pulse Rate, Systolic Blood Pressure and Peripheral Oxygen Saturation (CAPSO) Model". In: *Frontiers in Medicine* 8 (Dec. 2021). ISSN: 2296-858X. DOI: 10.3389/fmed.2021.810195. URL: http://dx.doi.org/10.3389/fmed.2021.810195.

[27] Catherine W. Liu et al. "Machine Learning Improves the Accuracy of Trauma Team Activation Level Assignments in Pediatric Patients". In: *Journal of Pediatric Surgery* 59.1 (Jan. 2024), pp. 74–79. ISSN: 0022-3468. DOI: 10.1016/j.jpedsurg.2023.09.014. URL: http://dx.doi.org/10.1016/j.jpedsurg.2023.09.014.

[28] Fredrik Linder et al. "A prospective stepped wedge cohort evaluation of the new national trauma team activation criteria in Sweden – the TRAUMALERT study". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 27.1 (Apr. 2019). ISSN: 1757-7241. DOI: 10.1186/s13049-019-0619-1. URL: http://dx.doi.org/10.1186/s13049-019-0619-1.

[29] Lina Holmberg et al. "Trauma triage criteria as predictors of severe injury - a Swedish multicenter cohort study". In: *BMC Emergency Medicine* 22.1 (Mar. 2022). ISSN: 1471-227X. DOI: 10.1186/s12873-022-00596-7. URL: http://dx.doi.org/10.1186/s12873-022-00596-7.

[30] Landstingens Ömsesidiga Försäkringsbolag. *Nationella traumalarm-skriterier*. Accessed: 2024-12-10. 2024. URL: https://lof.se/patientsakerhet/vara-projekt/saker-traumavard/.

Trauma team activation

Physiology

- Need for ventilatory support
- Respiratory rate <10 or >29/min
- Children: Signs of respiratory compromise
- Systolic BP <90 mmHg or no palp radial pulse
- Children: Capillary refill > 2 seconds
- Children: Pulse <90 or >190 if 0-1 years, <70 or >160 if 1-5 years
- GCS $\leq$ 13

Specific injuries

- Penetrating trauma to the abdomen, chest, neck or extremities above.
- $\geq$ 2 fractures on long bones
- Severe pain in pelvis
- Amputation above hand or foot
- Burn $\geq$ 18% TBSA or inhalation burn
- Deformed chest wall
- Suspected spinal cord injury
- Massive external hemorrhage
- Open skull fracture/impression fracture
- Face or neck trauma with threatened airway

Limited trauma team activation

Mechanism of injury

- MVC > 50 km/h without seatbelt
- Extrication time > 20 min
- Thrown out of vehicle
- MCC (or equivalent) > 35 km/h
- Children: Hit/run over by vehicle
- Fall > 5m in height
- Children: Fall > 3m in height

Cautions - (May influence TTA)

Ongoing deterioration, anticoagulatory medications, extremes in age, severe preexisting conditions, hypothermia (<35°C), intoxication, or pregnancy.

BP: Blood pressure, GCS: Glasgow coma scale, MVC: Motor vehicle crash, MCC Motorcycle crash, TBSA: Total burn surface area, children: age < 15. Each item on the list should result in full or limited TTA.

Table 14: Trauma team activation criteria in Sweden.[28] [29] [30].

TRITA – TRITA-CBH-GRU-2024:338
Stockholm, Sweden 2024

www.kth.se