Smart Health 18 (2020) 100139

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Machine learning models for synthesizing actionable care decisions on lower extremity wounds

Holly Nguyen [a,*], Emmanuel Agu [a], Bengisu Tulu [a], Diane Strong [a], Haadi Mombini [a], Peder Pedersen [a], Clifford Lindsay [b], Raymond Dunn [b], Lorraine Loretz [b]

[a] Worcester Polytechnic Institute, 100 Institute Road, Worcester, 01609, United States
[b] University of Massachusetts Medical School/UMass Memorial Health Car, 55 N Lake Ave, Worcester, 01655, United States

ARTICLE INFO

Keywords:
Classification
Chronic wounds
Lower extremity ulcers
Machine learning

ABSTRACT

Lower extremity chronic wounds affect 4.5 million Americans annually. Due to inadequate access to wound experts in underserved areas, many patients receive non-uniform, non-standard wound care, resulting in increased costs and lower quality of life. We explored machine learning classifiers to generate actionable wound care decisions about four chronic wound types (diabetic foot, pressure, venous, and arterial ulcers). These decisions (target classes) were: (1) Continue current treatment, (2) Request non-urgent change in treatment from a wound specialist, (3) Refer patient to a wound specialist. We compare classification methods (single classifiers, bagged & boosted ensembles, and a deep learning network) to investigate (1) whether visual wound features are sufficient for generating a decision and (2) whether adding unstructured text from wound experts increases classifier accuracy. Using 205 wound images, the Gradient Boosted Machine (XGBoost) outperformed other methods when using both visual and textual wound features, achieving 81% accuracy. Using only visual features decreased the accuracy to 76%, achieved by a Support Vector Machine classifier. We conclude that machine learning classifiers can generate accurate wound care decisions on lower extremity chronic wounds, an important step toward objective, standardized wound care. Higher decision-making accuracy was achieved by leveraging clinical comments from wound experts.

1. Introduction

Chronic lower extremity wounds ("ulcers") affect 4.5 million individuals in the United States annually (Frykberg & Banks, 2015). These wounds are expensive to treat ($7,439 to $70,000 per wound) with a total annual cost of $25 billion in the U.S. (Sen et al., 2009). Chronic wounds have become more widespread due to an aging population (Flanagan, 2013) and the rise of comorbidities, e.g., diabetes and cardiovascular disease, that can cause ulcers (Flanagan, 2013). Patients with chronic wounds experience reduced mobility, chronic pain, prolonged hospital stays, missed work days (Kirsner & Vivas, 2015; Sen et al., 2009), and negative psycho-social effects (Kirsner & Vivas, 2015). Chronic wounds precede 85% of amputations and may even lead to death (Järbrink et al., 2017).

Accurate and timely diagnoses can reduce the cost of ulcers (Gillespie, 2010), but that assumes access to wound specialists (Flanagan, 2013). Even if initially assessed by a wound specialist, follow-up analysis and treatment is often completed by non-experts, that

* Corresponding author.
E-mail address: hanguyen@wpi.edu (H. Nguyen).

https://doi.org/10.1016/j.smhl.2020.100139
Received 31 July 2019; Received in revised form 25 July 2020; Accepted 8 September 2020
Available online 17 September 2020

is, the majority of domestic wound assessments are conducted by registered nurses who may lack wound expertise (Guest et al., 2015; Zarchi et al., 2014, 2015). The result is inconsistent diagnosis and treatment decisions and poor wound care (Kirsner & Vivas, 2015), potentially resulting in a non-healing chronic wound with a financial consequence ranging from $10,000 to $15,000 per ulcer (Flanagan, 2013). Existing decision support systems are limited to rubrics or questionnaires that have to be filled out manually to generate decisions. Wound care decisions generated autonomously by an ML classifier could provide the necessary support aid for non-expert care providers, ultimately improving care decision consistency and reducing costs. To the best of our knowledge, our work is the first to research and create a machine/deep learning approach to autonomously generate actionable wound care decisions from wound images and assessment notes.

This research investigated using machine learning (ML) to provide such decision support by classifying wounds as one of three decisions: (1) Continue current treatment, (2) Request non-urgent change in treatment from a wound specialist, (3) Refer patient to a wound specialist. We also studied (1) whether clinically-validated important visual wound features are sufficient for generating a decision and (2) whether adding unstructured text from wound experts increased classifier accuracy. Automated decision support via ML classifiers would enable patients to receive standardized wound care from non-expert care providers.

2. Background

Our prior research focused on objective analysis of chronic wounds using visual features (Strong et al., 2014; Wang et al., 2013, 2015, 2016, 2017). Building on this work, we focus on developing a wound Clinical Decision Support System (CDSS), a “smart wound specialist”, that is embedded in a smartphone application. Currently, wound care provided by non-experts may be limited to relying on manual wound assessment tools and paper rubrics for grading wounds. This CDSS aims to surmount the limitations of manual wound assessment by using machine learning and deep learning methods to autonomously recommend actionable chronic wound care decisions.

2.1. Wound assessment tools

The most relevant recent review (Greatrex-White & Moxey, 2015) of current wound assessment tools (WAT) focuses on manual WATs and rubrics, which are clearly different from our machine learning approach to autonomously generate actionable wound care decisions. This review of manual WATs states that there exists no single WAT that completely satisfies a nurse’s needs in wound care management. The study (Greatrex-White & Moxey, 2015) suggested that in the absence of wound experts, a good WAT is the one tool that can help guide non-expert clinicians towards making informed wound care decisions. The most common WATs that are available to non-experts are Bates-Jensen Wound Assessment Tool (Bates-Jensen et al., 1992), Leg Ulcer Measurement Tool (Woodbury et al., 2004), Pressure Ulcer Scale for Healing (Hon et al., 2010), Braden scale (Bates-Jensen et al., 1992), and Photographic Wound Assessment Tool (Thompson et al., 2013). Although these tools grade and score wounds, they do not recommend actionable wound care steps. Non-experts who use such manual WATs still require additional support and more intuitive wound care guidelines to arrive at wound care decisions (Greatrex-White & Moxey, 2015). The additional support from an expert can be provided through remote consultation (telemedicine), official wound care guidelines or CDSS tools. However, telecare is difficult due to experts’ time constraints and wound care guidelines often require expertise to interpret. Thus, our current study seeks to close this gap by experimenting with ML algorithms as the basis for wound decisions in a CDSS tool that provides digital, autonomous, actionable wound care decisions to support non-experts.

For initial assessment of our wound images and to generate ground truth labels for our supervised machine learning and deep learning methods, we graded wound images using an image-based WAT (Thompson et al., 2013; Houghton et al., 2000), the most comprehensive of which is the Photographic Wound Assessment Tool (PWAT) (Thompson et al., 2013). Similar to other available tools, PWAT is a rubric for initial wound assessment, mainly describing the characteristics of a wound, but does not recommend a care decision. It is a validated, state-of-the-art wound assessment questionnaire designed to provide a consistent and quantitative method to represent visual wound attributes evaluated from a wound photograph (see Appendix A) (Thompson et al., 2013). PWAT has moderate to excellent reliability with an Intra-class Correlation Coefficient (ICC) of 0.71 for interrater reliability and 0.89 when comparing bedside assessment to photographic assessment using PWAT. Due to its ability to assess wound features from an image, and its reliability, we used PWAT to quantitatively grade wound characteristics, the results of which were subsequently used as one of many inputs for the ML methods that then autonomously generated wound care decisions.

PWAT scores eight attributes of wounds, (1) size, (2) depth, (3,4) type and amount of necrotic tissue, (5,6) type and amount of granulation tissue, (7) wound edges and (8) periulcer skin viability (Thompson et al., 2013). Each aspect is scored on a scale from zero to four (using aspect-specific severity rubrics), yielding eight PWAT sub-scores. The overall wound assessment is a total of the sub-scores, ranging from 0 to 32, with 32 indicating severe problems, and 0 indicating no wound issues. A decreasing PWAT score over time indicates wound healing.

2.2. Machine/deep learning classification

Machine learning (ML) methods are increasingly applied to clinical issues. Using ML ensembles, Econ et al. (2008)E detected cardiovascular disease from protein expression levels in blood samples. They found that bagged ensembles were more accurate than single classifiers, such as Decision Trees (DTs), Support Vector Machines (SVMs), Multi-layer Perceptron Artificial Neural Networks (MLP ANNs) and Bayesian networks. WeAidU, a CDSS that classified images, also found ensembles more accurate than single classifier types for the task of diagnosing myocardial infarction and heart ischemia (Ohlsson, 2004). Brown and Marotta (Brown & Marotta, 2017) found a gradient boosting ensemble model most accurate in classifying Magnetic Resonance Imaging (MRI). Consequently, in addition to single classifiers, we explored ensembles for classifying wounds.

Deep learning approaches have also been applied to biomedical data. Gao et al. (Gao et al., 2018) used a Hierarchical Attention Network (HAN) consisting of two layers of bidirectional LSTMs/GRUs to identify one of 12 possible International Classification of Diseases (ICD) codes, as well as determine a histological grade classification for cancer pathology reports. Bidirectional LSTMs/GRUs allow retention of past and present contextual information, which is particularly relevant for text data. HANs are used for document classification because its hierarchies reflect the breakdown of a document into sentences, then words. Gao et al. (Gao et al., 2018) found the HAN more accurate than SVM, Random Forest, extreme gradient boosting, RNN, and CNN for classifying pathology reports. Baumel et al. (Baumel et al., 2018) also found a HAN was more accurate (86% accuracy) than SVMs, Continuous Bag of Words (CBOW), and CNNs for classifying 10,000 patient reports with ICD codes. Due to success in prior work with similar medical data, we also implemented a HAN.

2.3. Text mining medical data

To extract information from the unstructured text from wound experts, we explored Natural Language Processing (NLP) and text mining techniques used for Electronic Health Records (EHR). Prior work has combined clinical knowledge with ML to improve classification accuracy. Zhang et al. (Zhang et al., 2018) leveraged clinical knowledge and ML methods (logistic regression) to produce groupings of medical order sets.

A popular approach in prior NLP work involves tokenizing input text and identifying token frequency. Zheng et al. (Zheng et al., 2016) identified cases of Diabetes Mellitus (DM) by mapping tokens to DM risk factors, and indicating presence with a binary encoding. Castro et al. (Castro et al., 2015) identified polycystic ovarian syndrome by outputting the frequency of tokens. Another approach used to transform free text is to use word embeddings. These were used by Gao et al. (Gao et al., 2018) and are considered a powerful tool to represent contextual and semantic meaning. We explored both techniques described to extract meaning from textual EHR data.

3. Methodology

3.1. Use scenario

To illustrate the envisioned functionality of our CDSS, and, thus, our ML classifiers, we present a use scenario. We assume a visiting nurse or a nursing home nurse is following up with a chronic wound patient who has previously seen a wound specialist and has a current treatment plan. The nurse is well-qualified but not a wound expert. The treatments the nurse can provide are limited due to lack of wound-specific training or lack of medical resources in remote settings (see Appendix B). Making these assumptions of situation and minimal wound expertise allows our CDSS to be used by any nurse in typical visiting nurse settings. However, we acknowledge that

Table 1
Treatment types, medical indicators, and wound examples categorized by decision.

	Treatment / Medical Indicators	Wound Example	
Decision 1: Continue with current treatment.	No necrotic tissue (wound is clean) No debridement needed No spreading infection No bone or tendon visible No ischemia or had a prior vascular treatment Size of wound is small enough to not necessitate a skin graft Does not need offloading		For a small, uninfected wound, apply a gauze dressing.
Decision 2: Request order for non-urgent change in treatment from wound specialist.	Change dressing type (if wound is too dry or too moist) VAC (vacuum assisted closure) (if wound is clean but needs closure or granulation assistance) Offloading (if in an area where pressure is an issue) Compression (if venous ulcer) Antibiotic (if signs of infection)		For a dry wound, apply a moist dressing.
Decision 3: Refer patient to a wound specialist.	Debridement (if wound has necrotic tissue) Ascending ischemia (i.e., may indicate a need for revascularization) Wet gangrene Surgery if bone/tendon visible Amputation Skin graft (if wound is clean but of a large size)		For a wound needing surgical cleaning, recommend debridement.

there are many possible scenarios. Thus, to provide a rigorous grounding for our ML algorithms' evaluation, we experiment under two specific clinical decision making scenarios:

1) ML actionable support is provided with no expert involvement (i.e., there is limited input from a wound expert).
2) ML actionable support also utilizes as input EHR clinical assessment notes (i.e., free text comments from an expert available as EHR data).

We address scenario 1 by assuming that the only data available is the visible wound. The nurse would capture an image of the wound with our CDSS smartphone app and be prompted to enter some wound information such as wound location, appearance, and other clinical characterizations that could be easily assessed visually by a nurse. These clinically-validated important visual wound features would be the only input to the ML classifier in scenario 1.

In scenario 2, the wound features are still available to the nurse but we also include our wound care experts' clinical assessment notes that serve the same purpose and contain similar information as EHR notes. These notes simulate a scenario in which an expert has previously conducted a remote consultation or assessment that has been documented in the EHR.

Thus, the final aspect of the use scenarios is the type of actionable care decision that will be autonomously generated by the CDSS smartphone app (i.e., by the ML classifiers embedded in the app). With the aid of wound care experts, we established three actionable wound care decisions our wound CDSS will recommend, see Table 1. These decisions align with the abilities and roles of various care providers within the healthcare system, ensuring standardized wound care, while saving time and money, reducing unnecessary travel and the usage of wound specialists.

For each actionable care decision (target classes), there are typical chronic wound conditions which are shown in Table 1 with example wound images, as collected from the wound experts on our research team. Additional examples of wound decisions and the classification rationale are in Appendix C.

3.2. Wound images and ground truth for the ML study

The 205 wound images used in this study were selected from 2,064 wound images. 1,695 images are from a corpus of IRB-approved UMMS patient data, with another 369 publicly available from the Internet (WPI IRB 18–0148). To ensure image quality, images were excluded if they were too dark or light, the full wound was not in view, or the image was a duplicate of another image (see Appendix D).

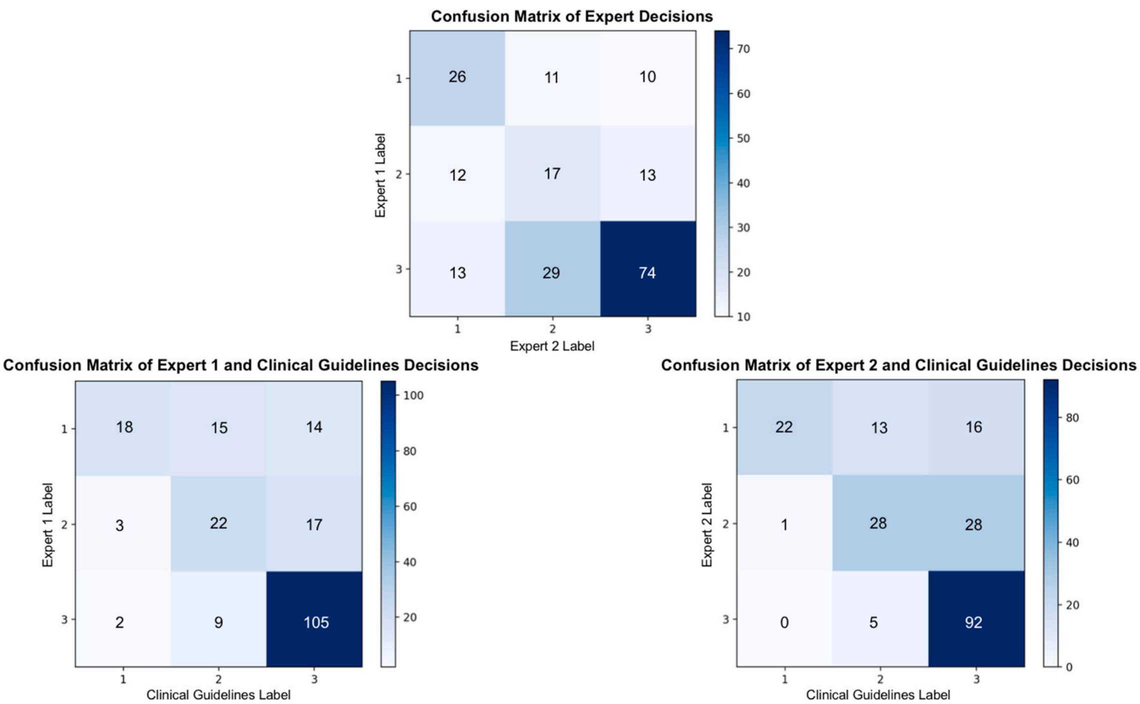

Fig. 1. Confusion matrix between wound experts' decision labels.

Table 2
Set of visual features.

Wound Type					PWAT sub-scores								PWAT Total Score
Diabetic foot	Venous	Arterial	Pressure	Surgical	Sub-score 1	Sub-score 2	Sub-score 2	Sub-score 4	Sub-score 5	Sub-score 6	Sub-score 7	Sub-score 8	Total score
0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0–4	0–4	0–4	0–4	0–4	0–4	0–4	0–4	0–32

High quality images were integral to ensuring that experts could accurately view images and assess the wound to provide ground truth.

After exclusion of poor quality or duplicate images, we generated a random subset to use for ML experiments by sampling images from four chronic wound types (diabetic foot, pressure, venous, and arterial ulcers) with varied wound characteristics (Kirsner & Vivas, 2015). Surgical wounds that began as an ulcer were included as a fifth type since their treatment is similar to that of the other four wound types. Image examination by our wound experts was a time-consuming process, so we limited our sample to 205 images.

Two wound experts, a plastic surgeon (Expert 1) and a dually credentialed podiatric surgeon/vascular Nurse Practitioner (DPM/NP) (Expert 2), provided ground truth by indicating their treatment decision when shown each wound. Ground truth refers to which wound decision (1, 2 or 3) a wound expert assigned to each wound. The ML classification model attempts to learn these decisions during the model training process.

During wound decision labeling sessions, the wound experts viewed printed wound images, assigned a decision and explained why. To reduce bias, questions were limited to clarifications or requests for further explanation. Each session was video recorded and transcribed. The wound expert explanations were used to emulate EHR text content, which became text inputs for our classifiers.

The two experts provided the same decision for 57% of the wounds (117 images), see the confusion matrix on the left in Fig. 1. Extreme disagreement (Decision 1 and 3 selected) occurred for 11% of the wounds. One wound expert explained that decision disagreement is common due to the different treatment philosophies. For example, for a large but clean wound, one expert may recommend a skin graft to help the wound close more quickly whereas another expert may let it heal on its own.

We also generated a third set of decision labels from evidence-based clinical guidelines, which provided objective decisions, independent of our two experts. The middle and right confusion matrices in Fig. 1 show the agreement between clinical guideline decisions and the two wound experts.

To establish ground truth when the wound experts disagreed on decisions, we investigated four policies to assign a final decision ($\text{dec}_{\text{final}}$) based on our three decision results (dec_{exp1} = Expert 1's, dec_{exp2} = Expert 2's, $\text{dec}_{\text{clinical}}$ = decision from clinical guidelines rules). Each decision assigned is given a corresponding numerical value of 1, 2, or 3 (see Table 1).

Policy 1, Cautious Decision: Select the more cautious decision assigned by the two wound experts Equation (1). Specifically, decision 3 is more cautious than decision 2, which is more cautious than decision 1. Decision 3 is the most cautious because the patient would eventually see a wound specialist in person. Equation (1) uses the max function so that the highest numerical value (1, 2, or 3) is assigned as the final decision.

$$dec_{final} = \max(dec_{\exp 1}, dec_{\exp 2}) \quad (1)$$

Policy 2, Surgical Decision: Select the decision assigned by the plastic surgeon (Expert 1) as more severe wounds were typically referred for surgical treatment Equation (2).

$$dec_{final} = dec_{\exp 1} \quad (2)$$

Policy 3, Holistic Decision: Select the decision assigned by the dually credentialed podiatric surgeon/vascular Nurse Practitioner (DPM/NP) (Expert 2) due to the DPM/NP's daily interaction with a wide variety of severity of wounds, and experience with limb salvage as a podiatric surgeon, and preventative treatments Equation (3).

$$dec_{final} = dec_{\exp 2} \quad (3)$$

Policy 4, Majority Decision: Select the majority decision among three decisions (Expert 1, Expert 2, clinical guidelines). In cases when there is a 3-way tie (e.g., each of the three decisions were assigned and, thus, there are three distinct numerical values), there is no majority decision so the most cautious decision is selected Equation (4). For example, if both experts assign a value of 2 (i.e., they chose decision 2 for a wound), but the clinical guidelines recommend a value of 3 (i.e., decision 3), then the majority decision is assigned which would be 2. In another example, there may be three different decisions assigned (e.g., a 1, 2, and 3). In this case, since there is no consensus, we assign the most cautious decision which is the max value. Thus, the final decision assigned in a 3-way tie would be decision 3.

$$set_{dec} = \{dec_{\exp 1}, dec_{\exp 2}, dec_{clinical}\} \quad (4)$$

The decision disagreement policies resulted in some class imbalance. Policy 2 had the most balance among classes, followed by Policy 3, then Policy 4, and finally Policy 1. We hypothesize that Policy 4 may be the best representation of the truth since it accounts for both experts' opinions as well as clinical guidelines. Additionally, it prioritizes consensus among these three decision-makers, but in cases of disagreement, the final decision assigned is the most cautious. In this way, Policy 4 balances decision-maker consensus with caution for wound cases that may be more difficult to determine a decision.

3.3. Visual feature extraction

In envisioned use scenario 1 in which there is no expert involvement, a nurse would rely on the wound features that are visually discernible in-person. In this ideal use case, the smartphone wound app (our CDSS) would automatically extract visual features (via image analysis) and the nurse would provide supplemental information such as odor. Thus, the wound app would have a set of visual features as input to the ML classifiers that would generate an actionable decision. These visual features are based on clinically-validated important wound attributes as specified by the PWAT wound grading rubric (See Appendix A). In our experiments, we manually extracted the visual features and scored each wound to generate these PWAT sub-scores used as input to the ML classifiers. However, longer term, we are researching and developing deep learning methods to automatically analyze the wound image and extract these clinically-validated visual features (PWAT sub-scores) as well as wound type (diabetic foot, venous, arterial, pressure, or surgical wounds). Wounds could be labelled with one or more wound types (mixed wounds). PWAT sub-scores and total score were calculated as the average of three independent investigator scores. Table 2 show a comprehensive overview of the visual features that we used in our experiments and the values accepted.

3.4. Feature extraction from expert comments/notes

Comments collected from experts simulated observations a clinician might record in the EHR. For each wound, experts' comments were merged and split into sentences. Text cleaning, labelling of negated terms, tokenization, stop word removal and stemming were performed, examples of which are in Table 3.

The Term Frequency Inverse Document Frequency (TF-IDF) approach (Afzal et al., 2018; Castro et al., 2015; Zhang et al., 2018; Zheng et al., 2016) was used to vectorize each comment, generating text features. Similar to bag-of-words, TF-IDF weights how frequently a term occurs in the entire corpus. Data was scaled to unit variance so that each feature had a mean value of zero, ensuring unit independence. Lastly, Principal Components Analysis (PCA) was used for dimensionality reduction to potentially improve classification accuracy.

3.5. Visual classification tasks and experimental datasets

Our two classification tasks generated four datasets.

1. Visual Classification Task: Classify a wound as one of three actionable decisions using visual features as input.
   - VIS (Visual): Dataset was generated by processing only visual features – 8 PWAT sub-scores, 1 PWAT total score, and 5 wound types (14 features total).
2. Visual and Text Classification Task: Classify a wound as one of three actionable decisions using visual features and textual EHR features as input.
   - B (Basic): Dataset was generated by processing text using TF-IDF then combining with visual features (638 features resulting).
   - NEG (Negated terms marked): Dataset was generated by processing text using TF-IDF and labeling of negated tokens (capturing some token context), then combining with visual features (722 features resulting).
   - PCA (Principal Components Analysis): Dataset was generated by processing text using TF-IDF then combining with visual features but transformed using PCA (159 features resulting), representing the original sparse text data more succinctly.

3.6. Machine learning classification

The SMOTE (Synthetic Minority Over-sampling Technique) (Castro et al., 2015) was applied to balance the dataset, reducing bias. Single classifiers, specifically Decision Tree (DT), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP), were investigated (Eom et al., 2008; Ohlsson, 2004). We also implemented bagged (bootstrap aggregated) DT and SVM classifiers by training multiple single classifiers on separate training sets (generated through oversampling and bootstrapping), and then aggregating predictions (Breiman, 1996). Aggregation occurred using two voting methods adapted from (Eom et al., 2008): 1) Majority voting (most frequent class was predicted), and 2) Weighted majority voting (classifiers weighted in proportion to accuracy on training set).

We also explored (1) a Random Forest classifier, with automatic feature importance ranking and selection, to handle the sparsity of the text features and (2) gradient boosted trees, using extreme gradient boosting (XGBoost), which iteratively creates multiple weak learners to learn the error from the previous learner thus improving learning performance.

We transformed text input into word embeddings, numerical one-dimensional vectors that capture semantic meaning. We used Word2Vec, a neural network model that predicts the next word (context) given the current word. Given our small text corpus, we used pre-trained word embeddings from the Google News corpus containing 3 million words and trained on ~100 billion words.

To capture greater semantic meaning and context from textual features (i.e., the simulated EHR data) with the word embeddings as input, we evaluated a Hierarchical Attention Network (HAN) model that contained two layers of bidirectional LSTMs (Long Short Term memory, bi-LSTM) or bidirectional GRUs (Gated Recurrent Units, bi-GRU) (Gao et al., 2018). With two hierarchies in the HAN, the word embeddings were fed to the lower hierarchy, then weighted with an attention mechanism to create a sentence embedding. This was then weighted with another attention mechanism to create a document embedding. Finally, the document embedding was concatenated with the visual features and fed into a Dense network layer to produce the final decision. We experimented with different hyperparameters, such as the type of bidirectional cells (bi-LSTM or bi-GRU), the number of nodes in each layer, the number of nodes in the hidden attention layer, and dropout rate.

All single and ensemble classifiers were trained using nested 10-fold Cross Validation (CV) to find optimal hyperparameters and to determine generalizability of classifiers. We used sequential-model based optimization, comparing two types of surrogate functions: Gaussian process and gradient boosted regression trees. We experimented with the hyperparameter values in Table 4 (see Appendix E for further explanation).

4. Results

Our results below are averages of five iterations of 10-fold cross validation. Each classifier's performance was evaluated for each of the four decision policies using weighted F-score Equations (5)–(7), which was weighted using the number of true instances of each class.

$$precision = \frac{T_{pos}}{T_{pos} + F_{neg}} \tag{5}$$

$$recall = \frac{T_{pos}}{T_{pos} + F_{pos}} \tag{6}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

The baseline classification method was a majority classifier, OneR (or One Rule), which predicts the most frequently occurring class label across the dataset as its output. OneR demonstrated that our classifiers performed better than random guessing. For our data, OneR predicted Decision 3 as its output.

Table 3
Text mining process with example comment.

Text Mining Step	Example Comment
Original comment	"I don't see tendon. This is about 25% necrotic tissue. Thin, white, grey necrotic."
(Step 1) Mark negation	"I don't_NEG see_NEG tendon_NEG. This is about 25% necrotic tissue. Thin, white, grey necrotic."
(Step 2) Remove non-alphanumeric characters	"I don't_NEG see_NEG tendon_NEG This is about 25 necrotic tissue Thin white grey necrotic"
(Step 3) Lowercase	"i don't_NEG see_NEG tendon_NEG this is about 25 necrotic tissue thin white grey necrotic"
(Step 4) Tokenize	["i", "don't_NEG", "see_NEG", "tendon_NEG", "this", "is", "about", "25", "necrotic", "tissue", "thin", "white", "grey", "necrotic"]
(Step 5) Remove stop words	["see_NEG", "tendon_NEG", "25", "necrotic", "tissue", "thin", "white", "grey", "necrotic"]
(Step 6) Stemming	["see_NEG", "tendon_NEG", "25", "necrot", "tissu", "thin", "white", "grey", "necrot"]

4.1. Visual Classification Task

As illustrated in Table 5, all classifiers performed better than the baseline classifier (random guessing) using the VIS dataset as input. Across all policies, the DT ensemble (with weighted majority voting, noted as ^{w} in Table 5 with ^{m} being majority voting) had the worst performance and the single SVM had the best performance. The ensembles outperformed the single classifiers only in Policy 2. Ensembles are designed to generate a stronger prediction by utilizing single classifiers that learn differently. However, with such a small feature set, the single classifiers compiling the ensemble may have been highly correlated (i.e., producing very similar predictions) and, thus, less useful in determining the final prediction. We hypothesized that SVM would perform well on such a small dataset given its ability to handle misclassified instances, or the support vectors.

4.2. Visual and Text Classification Task

As illustrated in Table 6, the ensembles generally performed better than single classifiers on the datasets including both visual and textual features. For Policy 1 and 2, Random Forest performed best while Policy 3 and 4 had XGBoost as the optimal classifier. This is probably due to the sparse textual feature matrix, since feature and instance selection (inherent to the classifier) is useful when there are many more features than instances. XGBoost had the best performance achieved across all experiments for Policy 4.

We also analyzed performance between the B dataset (combination of textual and visual features with no further preprocessing aside from scaling to unit variance), the NEG dataset (negated terms in text were marked), and the PCA dataset (Principal Components Analysis applied). The NEG dataset produced the best performance for all but Policy 3, which had the highest performance with the PCA dataset. Preprocessing the text by marking negation was an important step in distinguishing wound experts' comments. The Visual and Text Classification Task produced better performance compared to the Visual Classification Task (see Appendix F for details), which answers one of the research questions.

Table 4
Hyperparameters tuned.

	Hyperparameter	Value
Decision Trees	Criterion	entropy, gini
	Max depth	5–20
	Min leaf samples	2–10
	Min samples split	2–10
SVM	Kernel	linear, polynomial, rbf
	C	0.001–1000
	Gamma	0.0001–100
MLP	Number of layers	3–7
	Number of neurons	50–500
	Dropout	0.0–1.0
	Activation	relu, selu
	Weight initialization	he_normal, random_uniform
Random Forest	Number of estimators	10–100
	Criterion	entropy, gini
	Max Depth	2–20
	Min leaf samples	2–10
	Min split samples	2–10
XGBoost	Number of estimators	20–100
	Eta (learn rate)	0.01–0.5
	Max delta step	1–10
	Min child weight	1–10
	Max depth	5–20
	Gamma	1.1–1.0
	Subsample	0.5–10
	Colsample by tree	0.5–1.0
	Lambda (L2 regularization)	1–5
HAN	Type of RNN unit	bi-LSTM, bi-GRU
	Number of RNN units in each layer	10–500
	Number of neurons in attention layer	50–300
	Number of neurons in dense layer	10–300
	Dropout	0.0, 0.5

Table 5
Visual Classification Task results (grouped by decision policy).

Classifier	Policy 1	Policy 2	Policy 3	Policy 4	Average
Baseline	0.548	0.409	0.304	0.517	0.445
Single Classifiers					
DT	0.647	0.556	0.506	0.659	0.592
SVM	**0.743**	0.612	**0.584**	**0.766**	**0.676**
MLP	0.699	0.622	0.569	0.725	0.654
Ensemble Classifiers					
DT^m	0.652	0.631	0.581	0.714	0.645
SVM^m	0.720	0.626	0.552	0.736	0.659
DT^w	0.603	0.514	0.477	0.656	0.563
SVM^w	0.654	0.578	0.520	0.703	0.614
Random Forest	0.707	**0.660**	0.576	0.750	0.673
XGBoost	0.713	0.619	0.545	0.729	0.652

4.3. Hierarchical Attention Network

Based on results from hyperparameter optimization, we used an architecture of 117 bi-LSTM cells with an attention context of 178 nodes, and 49 dense nodes to process the concatenated document embedding and visual features. During the optimization process, bi-LSTM was chosen more often than a bi-GRU cell. In order to ensure that the model generalized, dropout of 0.5 was used.

We hypothesized that using pre-trained word embeddings would aid in leveraging semantic information found from the free text and enhance the performance of the HAN model. We took advantage of SMOTE to mitigate the small sample size and balance our augmented dataset. Using SMOTE we generated training sets with 336, 279, 231 and 324 samples for Policies 1 through 4, respectively, with an equal testing set of 41 data samples for all the policies. As shown in Table 6 above, our HAN model achieved an average F-score of 0.601 across all four policies compared to the baseline (F-score = 0.445). The highest F-score obtained was 0.657 for Policy 4. This was expected as we hypothesized in section 3.2 that Policy 4 represents the truth best since it integrates both experts' opinions and clinical guidelines. Compared to the prior study (Gao et al., 2018) that utilized HAN for only clinical text classification to predict the primary site of cancer given the content of the 942 cancer pathology reports, our HAN model produced an overall average F-score of 0.601 across all policies whereas the other study achieved an average F-score of 0.594. However, for their histological grade classification task (predict the histological grade of a cancer given the cancer pathology report) their model achieved a higher average F-score of 0.822. This suggests that in addition to the number of training samples the complexity of the classification task and the content and length of the textual features may also affect the performance of HAN model. We argue that higher performance for the classification of complex chronic wound decisions can be achieved using our HAN with sufficient training data (comments and visual features), which will be investigated in future work using this technique with additional training data.

4.4. Feature importance

Feature analysis for the best performing classifiers, Random Forest and XGBoost, revealed the most important words for each Policy, separated by Decision class. Random Forest and XGBoost demonstrated that across classifiers and policies, visual features had

Table 6
Visual and Text Classification Task results (grouped by decision policy).

Classifier	Policy 1			Policy 2			Policy 3			Policy 4		
	B	NEG	PCA	B	NEG	PCA	B	NEG	PCA	B	NEG	PCA
Baseline	0.548			0.409			0.304			0.517		
Single Classifiers												
DT	0.627	0.606	0.605	0.560	0.576	0.497	0.578	0.574	0.577	0.710	0.719	0.628
SVM	0.758	0.764	0.752	0.660	0.671	0.634	0.706	0.688	0.683	0.782	0.785	0.770
MLP	0.761	0.754	0.729	0.628	0.616	0.590	0.644	0.636	0.632	0.735	0.753	0.737
Ensemble Classifiers												
DT^m	0.703	0.712	0.707	0.642	0.646	0.618	0.660	0.687	0.717	0.761	0.772	0.716
SVM^m	0.456	0.780	0.722	0.678	0.674	0.598	0.712	0.712	0.670	0.787	0.781	0.747
DT^w	0.514	0.602	0.597	0.562	0.561	0.486	0.581	0.580	0.558	0.655	0.687	0.635
SVM^w	0.760	0.756	0.724	0.662	0.675	0.601	0.703	0.697	0.675	0.775	0.777	0.740
Random Forest	0.778	**0.781**	0.769	0.689	**0.717**	0.645	0.706	0.706	0.675	0.797	0.802	0.782
XGBoost	0.743	0.732	0.768	0.671	0.659	0.648	0.705	0.702	**0.727**	0.805	**0.810**	0.764
Deep Learning Classifier (Average across policies = 0.601)												
HAN (SMOTE)	0.646			0.560			0.541			**0.657**		

greater importance in determining a decision (Fig. 2). While the total PWAT score was the most important feature, particular textual features played a large role, such as “offload” and “clean”. Extrapolating from medical knowledge, the term “offload” can indicate a Decision 2 (particularly for a diabetic foot ulcer or a pressure ulcer) and “clean” may indicate a healing wound (i.e., a Decision 1). This suggests that while visual features are most important, textual features derived from expert comments are able to improve the accuracy of decisions.

4.5. Error analysis

We examined the confusion matrices from Policy 4 from our top ensemble models. Both Random Forest and XGBoost had very similar confusion matrices regarding decision predictions. The success, as compared to other ensemble methods, can be attributed to inherent feature and instance selection.

Both Random Forest and XGBoost misclassified a wound image with a total PWAT score of 15, with scores of 3 or higher for the four sub-scores relating to necrotic and granulation tissue. Despite unknown depth and presence of necrosis, Decision 1 was assigned by wound expert and clinical guidelines. This suggests that additional visual features may need to be provided (such as specific wound location and specific wound size) as well as other similar cases to learn from. Additionally, having access to wound healing progression over time may provide essential information; if wound size decreases significantly compared to a previous wound measurement then it indicates that the wound is healing (Snyder et al., 2010).

4.6. Decision disagreement resolution policy analysis

Policy 4 produced the best performance (average weighted F-score) across all policies and datasets, followed by Policy 1, Policy 3 and Policy 2. Additionally, the best performance for each classification task was with Policy 4 (0.766 for the Visual Classification Task with a Support Vector Machine classifier, and 0.810 for the Visual and Text Classification Task with an XGBoost classifier). This suggests that adding the generation of ground truth decisions using clinical guidelines (Policy 4) mitigated the subjective opinions of the wound experts (Policies 1–3).

5. Discussion

5.1. Findings

The Visual Classification Task demonstrated that it is possible to achieve reasonable accuracy (weighted F-score = 0.766) with a SVM classifier. Additional visual features, such as specific wound location that is important for wound care decisions, could be added in future work to improve the accuracy of classifying wounds solely based on an image.

Including textual EHR features in addition to visual features generally increased the accuracy of wound decision classifiers.

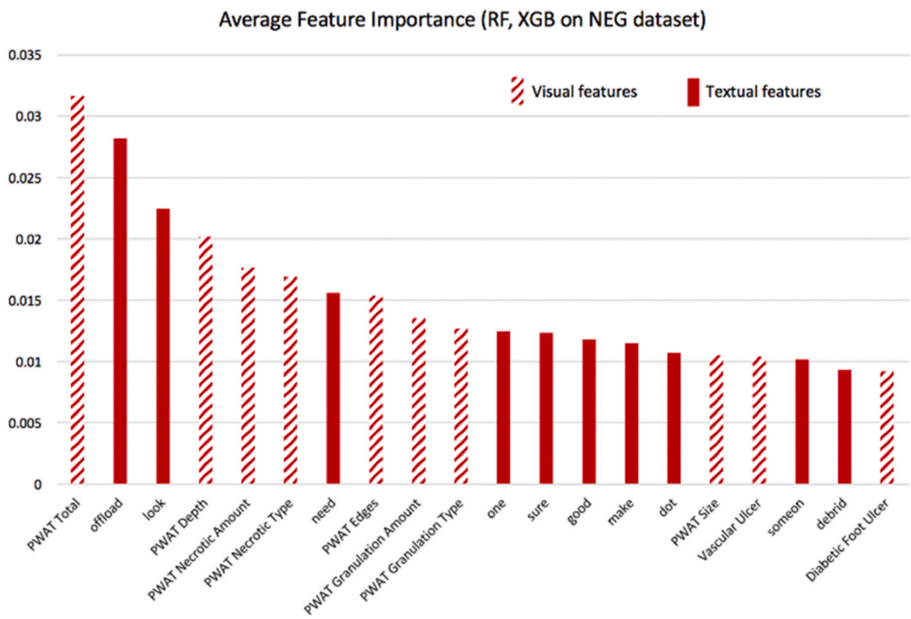

Fig. 2. Average feature importance from Random Forest and XGBoost.

Classification of wounds with visual and textual features achieved up to 81% accuracy with an XGBoost classifier, about 6% higher than the performance achieved with only visual features. Feature importance analysis also revealed that visual and textual features were both important. Marking negated terms during preprocessing improved performance. Comparing performance on the three experimental datasets which included textual features, the majority (6/8) of best performances (for each policy) were from training on the dataset with negated terms marked.

5.2. Machine learning classification

Evaluation of all machine learning classifiers across decision policies and datasets demonstrated the following outcomes:

- SVM had the best overall performance (highest F-score) when using only visual features as input.
- Ensemble classifiers outperformed single classifiers (based on weighted F-score) when using visual and textual input.
- XGBoost had the highest accuracy for all combinations of models, experimental datasets and decision policies.

The HAN did not perform as well as the ensemble classifiers (and single classifiers in some cases), probably due to the complex content of medical notes and lack of comments from wound experts for some wound images, as well as the generally small size of the dataset.

5.3. Decision disagreement resolution policy discussion

Establishing ground truth was an important and challenging step. Policy 4 performed best (weighted F-score) across all classifier models and classification tasks. This could be attributed to the addition of labels generated from evidence-based clinical guidelines, demonstrating that objective measures from clinical guidelines improves the accuracy of decision-making. In the future, ground truth should be generated not only on wound experts' opinions but also the most current evidence-based clinical guidelines.

6. Conclusion

This research demonstrated that actionable wound care decisions could be classified given a combination of visual and text features with 81% accuracy. The envisioned smartphone wound assessment system has the potential to be used as a CDSS to aid a registered nurse in deciding what treatment a chronic wound requires, thereby standardizing wound care.

CRediT authorship contribution statement

Holly Nguyen: Methodology, Data curation, Software, Investigation, Writing - original draft. **Emmanuel Agu:** Supervision, Conceptualization, Data curation, Writing - review & editing. **Bengisu Tulu:** Conceptualization, Data curation, Writing - review & editing. **Diane Strong:** Conceptualization, Writing - review & editing. **Haadi Mombini:** Methodology, Data curation, Investigation, Writing - original draft. **Peder Pedersen:** Conceptualization, Writing - review & editing. **Clifford Lindsay:** Conceptualization, Writing - review & editing. **Raymond Dunn:** Conceptualization, Writing - review & editing. **Lorraine Loretz:** Conceptualization, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025801. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

A. Photographic Wound Assessment Tool

Photographic Wound Assessment Tool
PWAT – Revised

Item	Assessment	Score
1. Size	0 = wound is closed (skin intact) or nearly closed (<0.3cm^2) 1 = 0.5 – 2.0 cm^2 2 = 2.0 - 10.0 cm^2 3 = 10.0 – 20.0 cm^2 4 > 20.0 cm^2	
2. Depth	0. wound is healed (skin intact) or nearly closed (<0.3cm^2) 1. full thickness 2. unable to judge because majority of wound base is covered by yellow/black eschar 3. full thickness involving underlying tissue layers 4. tendon, joint capsule, bone, visible/ present in wound base	
3. Necrotic Tissue Type	0 = None visible or wound is closed (skin intact) or nearly closed (<0. 3cm^2) 1 = majority of necrotic tissue is thin White/grey or yellow slough 2 = majority of necrotic tissue is thick, adherent white yellow slough or fibrin 3 – majority of necrotic tissue is white/grey devitalized tissue or eschar 4 = majority of necrotic tissue is hard grey to black eschar	
4. Total Amount of Necrotic Tissue	0 = None visible in open wound or wound is closed (skin intact) or nearly closed(0.3cm^2) 1 = < 25% of wound bed covered 2 = 25% to 50% of wound covered 3 = > 50% and < 75% of wound covered 4 = 75% to 100% of wound covered	
5. Granulation Tissue type	0 = Wound is closed (skin intact) or nearly closed (<0.3cm^2) 1 = majority (>50%) of granulation tissue is healthy looking (even bright red appearance) 2 = majority of granulation tissue is unhealthy (eg. pale, dull, dusky, hypergranulation) 3 = majority of granulation tissue is damaged, friable, degrading 4 = there is no granulation tissue present in the base of the open wound (all necrotic)	
6. Total Amount of Granulation Tissue	0 = Wound is closed (skin intact) or nearly closed (<0.3cm^2) 1 =75% to 100% of open wound is covered with granulation tissue 2 = >50% and <75% of open wound is covered with granulation tissue 3 = 25% to 50% of wound bed is covered with granulation tissue 4 = <25% of wound bed is covered with granulation tissue	
7. Edges (directly touching and within 0.5cm of wound edge)	0 = Wound is closed (skin intact) or nearly closed (<0.3cm^2) or edges are indistinct, diffuse, not clearly visible because of re-epithelialization 1 = majority (>50%) of edges are attached with an advancing border of epithelium 2 = majority of edges are attached even with wound base (not advancing) 3 = majority of edges are unattached and/or undermined 4 = majority of edges are rolled, thickened or fibrotic (do not include callus formation)	
8. Periulcer Skin Viability (consider skin visible in photo or within 10 cm of wound edge)	Number of factors affected 0 = None 1 = One only 2 = Two or Three 3 = Four or Five 4 = six or more - callus - dermatitis - maceration - desiccation or cracking - bright red, erythemic - edema - excoriation - skin tearing/irritation r/t wound dressing or tape - hypo/hyper pigmentation - other: ____________	
TOTAL SCORE		

© Hodgkinson, Bowles, Gordy, Parslow, Houghton, 2010

Fig. A1 Rubric from the revised Photographic Wound Assessment Tool (PWAT)

B. Assumptions about Care Provider Core Medical Competencies

Table B.1
Assumptions about Care Provider Core Medical Competencies

Care Provider	Core Medical Competencies
Registered Nurse	Standard wound care (dressing change)
	General lavage (i.e., cleaning of the wound surface and periwound area)
	Application of compression dressing
	Application of vacuum assisted closure (VAC) dressing
	Ability to administer oral antibiotic
	Aid for patient in reducing pressure on wound area (i.e., offloading or moving patient)
Wound Specialist	Debridement
	Surgery (for debridement, amputation, skin graft)
	Infection assessment, diagnosis and design of treatment plan
	Vascular evaluation

C. Wound Decision Examples

Table C.1
Wound Decision Examples (with an associated treatment and reason for classification)

Decision 1 Example

Example Wound	Type of Treatment /Indicator	Reason for Classification
	Continue applying a standard wound dressing	• Clean wound (no necrotic tissue), not too moist or dry • Wound edges are blending with granulation tissue • Small enough wound size

Decision 2 Examples

Example Wound	Type of Treatment /Indicator	Reason for Classification
	Offloading (foot put in a cast to reduce pressure)	• Etiology: diabetic foot ulcer • Wound located over bony prominence
	Compression	• Etiology: venous ulcer

(continued on next page)

D. Wound Exclusion Examples

Table C.1 (*continued*)

Decision 1 Example		
	Change to vacuum assisted closure (VAC) or a moist occlusive dressing	• Clean wound bed • Wound bed is dry or needs closure assistance
	Change to a dry dressing	• Moist wound bed with macerated islands of granulation tissue • Indicates healing with epithelium (new skin)
	Antibiotic	• Clean tissue but looks inflamed surrounding the wound
Decision 3 Examples		
Example Wound	Type of Treatment /Indicator	Reason for Classification
	Debridement	• Almost no granulation tissue visible in the wound bed

(*continued on next page*)

Table C.1 (*continued*)

Decision 1 Example		
	Ischemia	• Toe turning black indicates critical limb ischemia
	Wet gangrene	• Presence of wet gangrene indicates immediate attention
	Surgery (bone or tendon revision)	• Tendon visible, which requires surgical revision
	Amputation	• Toe looks ischemic and red indicates vascular issues
	Skin graft	• Wound has clean, beefy, red granulation tissue • Size of wound is large • Complications could ensue later based on wound location over Achilles tendon if wound is not properly treated

Table D.1
Wound Exclusion Examples

Example Wound	Exclusion Criteria
	Poor lighting conditions (i.e., image was too dark or too light to see the wound clearly)
	Poor camera angle which prevented the full wound being in view
	Detail of wound image was limited due to the distance at which the image was captured (either the camera was positioned too close or too far away)
	Wound was covered (e.g., a bandage obstructed view of the wound)

(*continued on next page*)

Table D.1 (*continued*)

Example Wound	Exclusion Criteria
 	Duplicate images of the same wound captured at approximately the same time

E. Hyperparameter optimization process

SMBO (including Bayesian optimization) is typically viewed as more efficient than a grid search because the next set of hyperparameters are chosen based on approximating improvement from the previous set of hyperparameters. A grid search, as the name suggests, will experiment with all combinations of hyperparameters even if the search in a certain hyperparameter space is unnecessary. SMBO consists of the following components: an objective function (the weighted F-score we want to minimize as validation error), domain space (possible range of hyperparameter values), optimization algorithm (also known as the surrogate, or the specific method/model used to choose the next hyperparameter values), and the results (results from hyperparameters and the objective function scores). The surrogate model is used because the objective function is very expensive to evaluate. We used two versions of SMBO, namely the sci-kit optimize package implementation of gp-minimize and gbrt-minimize. These offer two options for the surrogate: Gaussian process (gp-minimize) and gradient boosted regression trees (gbrt-minimize). Both were tested for each classifier, comparing the convergence (i.e., how many iterations it took to find an optimal score) and the final objective score to determine which was a better fit for this dataset and features.

We used nested 10-fold cross validation to accomplish this. For the neural network models, MLP and HAN, 10-fold cross-validation and 5-fold cross validation was used, respectively with an 80-10-10 train/validation/test split due to the length of time to accomplish hyperparameter tuning and cross-validation especially for HAN model. No hyperparameters were chosen based on evaluation on the test set (only a validation set). Process:

1. Perform 10-fold cross validation split (i.e., 90/10 training/test split).
2. Conduct hyperparameter optimization on the training set by performing another 10-fold cross validation split (use both Gaussian process and gradient boosted regression trees) with 50 iterations and 10 random restarts.
3. Save best configuration of hyperparameters (based on performance on inner loop test set).
4. Train a classifier on the full training set with the hyperparameters from Step 3.

F. Classification task evaluation

We conducted confidence interval analysis between the results from the two classification tasks. We used the weighted F-score from the Visual Classification Task (Task 1) and the highest weighted F-score (from one of the three datasets, either B, NEG, or PCA) for the Visual and Text Classification Task (Task 2). We constructed a 95% confidence interval (with $n = 205$, $Z_n = 1.96$) as shown in Equations F.1 and F.2 (the subscript indicates the number of the classification task).

If the confidence interval contains zero, this indicates that the difference in error could be zero, and thus, the models are not statistically significantly different. In Table F.1 we present these results with bolded values indicating an interval containing zero. Notably, all intervals where models were not statistically significantly different can be found with DT classifiers (either the single classifier or weighted ensemble). This indicates that DT may perform better with the subset of visual features, which is much smaller than the sparse text matrix produced from adding textual EHR features.

$$d = error_2 - error_1 \quad \text{(F.1)}$$

$$d \pm Z_n \sqrt{\frac{error_2(1 - error_2)}{n} + \frac{error_1(1 - error_1)}{n}} \quad \text{(F.2)}$$

Table F.1
Classification Task evaluation

Classifier	Policy 1	Policy 2	Policy 3	Policy 4
Single Classifiers				
DT	**(-0.073, 0.113)**	**(-0.117, 0.075)**	**(-0.168, 0.024)**	**(-0.150, 0.029)**
SVM	(-0.591, −0.424)	(-0.376, −0.190)	(-0.382, −0.198)	(-0.632, −0.470)
MLP	(-0.546, −0.375)	(-0.344, −0.156)	(-0.307, −0.119)	(-0.562, −0.392)
Ensemble Classifiers				
DTm	(-0.454, −0.273)	(-0.370, −0.184)	(-0.390, −0.207)	(-0.571, −0.402)
SVMm	(-0.584, −0.416)	(-0.396, −0.212)	(-0.356, −0.172)	(-0.605, −0.440)
DTw	(-0.300, −0.110)	**(-0.173, 0.020)**	**(-0.155, 0.037)**	(-0.403, −0.219)
SVMw	(-0.501, −0.326)	(-0.346, −0.160)	(-0.316, −0131)	(-0.565, −0.395)

References

Afzal, N., Mallipeddi, V. P., Sohn, S., Liu, H., Chaudhry, R., Scott, C. G., Kullo, I. J., & Arruda-Olson, A. M. (2018). Natural language processing of clinical notes for identification of critical limb ischemia. *International Journal of Medical Informatics, 111*, 83–89.

Bates-Jensen, B. M., Vredevoe, D. L., & Brecht, M. (1992). Validity and reliability of the pressure sore status tool. *Decubitus, 5*(6), 20–28.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018). Multi-label classification of patient notes: Case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Brown, A. D., & Marotta, T. R. (2017). Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. *Jamia, 25*(5), 568–571. https://doi.org/10.1093/jamia/ocx125. Available:.

Castro, V., Shen, Y., Yu, S., Finan, S., Pau, C. T., Gainer, V., Keefe, C. C., Savova, G., Murphy, S. N., Cai, T., & Welt, C. K. (2015). Identification of subjects with polycystic ovary syndrome using electronic health records. *Reproductive Biology and Endocrinology, 13*(1), 116. https://doi.org/10.1186/s12958-015-0115-z

Eom, J., Kim, S., & Zhang, B. (2008). AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Systems with Applications, 34*(4), 2465–2479.

Flanagan, M. (2013). *Wound healing and skin integrity: Principles and practice*. John Wiley & Sons.

Frykberg, R. G., & Banks, J. (2015). *Advances in Wound Care, 4*(9), 560–582. Available: https://www.liebertpub.com/doi/pdfplus/10.1089/wound.2015.0635.

Gao, S., Young, M. T., Qiu, J. X., Yoon, H., Christian, J. B., Fearn, P. A., Tourassi, G. D., & Ramanthan, A. (2018). Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association, 25*(3), 321–330. Available: https://www.ncbi.nlm.nih.gov/pubmed/29155996.

Gillespie, D. L. (2010). Venous ulcer diagnosis, treatment, and prevention of recurrences. *Journal of Vascular Surgery, 52*(5), 8S–14S.

Greatrex-White, S., & Moxey, H. (2015). Wound assessment tools and nurses' needs: An evaluation study. *International Wound Journal, 12*(3), 293–301.

Guest, J. F., Ayoub, N., McIlwraith, T., Uchegbu, I., Gerrish, A., Weidlich, D., Vowden, K., & Vowden, P. (2015). Health economic burden that wounds impose on the National Health Service in the UK. *BMJ Open, 5*(12), Article e009283. Available: http://bmjopen.bmj.com/content/5/12/e009283.abstract.

Hon, J., Lagden, K., McLaren, A., O'Sullivan, D., Orr, L., Houghton, P. E., & Woodbury, M. G. (2010). A prospective, multicenter study to validate use of the PUSH in patients with diabetic, venous, and pressure ulcers. *Ostomy. Wound., 56*(2), 26–36.

Houghton, P. E., Kincaid, C. B., Campbell, K. E., Woodbury, M. G., & Keast, D. H. (2000). Photographic assessment of the appearance of chronic pressure and leg ulcers. *Ostomy/Wound Management, 46*(4), 20–35.

Järbrink, K., Ni, G., Sönnergren, H., Schmidtchen, A., Pang, C., Bajpai, R., & Car, J. (2017). The humanistic and economic burden of chronic wounds: A protocol for a systematic review. *Systematic Reviews, 6*(15). https://doi.org/10.1186/s13643-016-0400-8. Available:.

Kirsner, R. S., & Vivas, A. C. (2015). Lower-extremity ulcers: Diagnosis and management.(report)(disease/disorder overview). *British Journal of Dermatology, 173*(2), 379.

Ohlsson, M. (2004). WeAidU—a decision support system for myocardial perfusion images using artificial neural networks. *Artificial intelligence in medicine* (Vol. 30,(1), 49–60.

Sen, C. K., Gordillo, G. M., Roy, S., Kirsner, R., Lambert, L., Hunt, T. K., & Longaker, M. T. (2009). Human skin wounds: A major and snowballing threat to public Health and the economy," wound repair and regeneration. *Official Publication of the Wound Healing Society [and] the European Tissue Repair Society* (Vol. 17,(6), 763–771. https://doi.org/10.1111/j.1524-475X.2009.00543.x. Available:.

Snyder, R. J., Cardinal, M., Dauphinée, D. M., & Stavosky, J. (2010). A post-hoc analysis of reduction in diabetic foot ulcer size at 4 weeks as a predictor of healing by 12 weeks. *Ostomy. Wound., 56*(3), 44.

Strong, D., Tulu, B., Agu, E., & He, S. Q. (2014). *Design of the feedback engine for a diabetes self-care smartphone app*.

Thompson, N., Gordey, L., Bowles, H., Parslow, N., & Houghton, P. (2013). Reliability and validity of the revised photographic wound assessment tool on digital images taken of various types of chronic wounds. *Advances in Skin & Wound Care, 26*(8), 360–373.

Wang, L., Pedersen, P. C., Agu, E., Strong, D. M., & Tulu, B. (2017). Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification. *IEEE Transactions on Biomedical Engineering, 64*(9), 2098–2109.

Wang, L., Pedersen, P. C., Strong, D., Tulu, B., & Agu, E. (2013). Wound image analysis system for diabetics. In *Medical imaging 2013: Image processing* (p. 866924).

Wang, L., Pedersen, P. C., Strong, D. M., Tulu, B., Agu, E., & Ignotz, R. (2015). Smartphone-based wound assessment system for patients with diabetes. *IEEE Transactions on Biomedical Engineering, 62*(2), 477–488.

Wang, L., Pedersen, P. C., Strong, D. M., Tulu, B., Agu, E., Ignotz, R., & He, Q. (2016). An automatic assessment system of diabetic foot ulcers based on wound area determination, color segmentation, and healing score evaluation. *Journal of Diabetes Science and Technology, 10*(2), 421–428.

Woodbury, M. G., Houghton, P. E., Campbell, K. E., & Keast, D. H. (2004). Development, validity, reliability, and responsiveness of a new leg ulcer measurement tool. *Advances in Skin & Wound Care, 17*(4), 187–196.
Zarchi, K., Haugaard, V. B., Dufour, D. N., & Jemec, G. B. E. (2015). Expert advice provided through telemedicine improves healing of chronic wounds: Prospective cluster controlled study. *Journal of Investigative Dermatology, 135*(3), 895–900.
Zarchi, K., Latif, S., HAugAARD, V. B., HjALAgER, I. R., & Jemec, G. B. (2014). Significant differences in nurses' knowledge of basic wound management–implications for treatment. *Acta Dermato-Venereologica, 94*(4), 403–407.
Zhang, Y., Trepp, R., Wang, W., Luna, J., Vawdrey, D. K., & Tiase, V. (2018). Developing and maintaining clinical decision support using clinical knowledge and machine learning: The case of order sets. *Jamia, 25*(11), 1547–1551. https://doi.org/10.1093/jamia/ocy099. Available:.
Zheng, L., Wang, Y., Hao, S., Shin, A. Y., Jin, B., Ngo, A. D., Jackson-Browne, M., Feller, D. J., Fu, T., Zhang, K., Zhou, X., Zhu, C., Dai, D., Yu, Y., Zheng, G., Li, Y., McElhinney, D. B., Culver, D. S., Alfreds, S. T., … Ling, X. B. (2016). Web-based real-time case finding for the population health management of patients with diabetes Mellitus: A prospective validation of the Natural Language processing–based algorithm with statewide electronic medical records. *JMIR Med Inform, 4*(4), e37. https://doi.org/10.2196/medinform.6328. http://medinform.jmir.org/2016/4/e37/. http://www.ncbi.nlm.nih.gov/pubmed/27836816