Artificial intelligence in the management and treatment of burns

Taib, Bilal; Karwath, Andreas; Wensley, K; Minku, Leandro; Gkoutos, Georgios; Moiemen, N. S.

DOI:
10.1016/j.bjps.2022.11.049

Document Version
Peer reviewed version

Citation for published version (Harvard):
Taib, B, Karwath, A, Wensley, K, Minku, L, Gkoutos, G & Moiemen, NS 2023, 'Artificial intelligence in the management and treatment of burns: a systematic review and meta-analyses', *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 77, pp. 133-161. https://doi.org/10.1016/j.bjps.2022.11.049

Artificial Intelligence in the Management and Treatment of Burns: A Systematic Review and Meta-analyses

Bilal Gani Taib MRCS [1] ,A Karwath PhD [2,3,4], K Wensley MRCS [1], L Minku PhD[5], G.V.Gkoutos PhD[2,3,4,8], N Moiemen FRCS (Plast)[6,7,8]

[1]Burns and Plastic Surgery Department, Queen Elizabeth Hospital, Birmingham, UK.
[2]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK
[3]Health Data Research UK Midlands Site, Birmingham, UK
[4]University Hospitals Birmingham NHS Foundation Trust, Edgbaston, Birmingham, UK
[5]School of Computer Science, University of Birmingham, Birmingham, UK
[6]College of Medical and Dental Sciences, University of Birmingham, UK
[7]Centre for Conflict Wound Research, Scar Free Foundation, Birmingham, UK
[8]NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham, UK

Corresponding Author
Bilal G Taib
bilal.taib@nhs.net

Department of Burns & Plastic Surgery
Queen Elizabeth Hospital,
Mindelsohn Way,
Birmingham, B15 2TH

+441216272000

*Part of this research has been presented at the International Society for Burns Injury 20th International Congress June 2021

Abstract

Introduction and Aim
Artificial Intelligence (AI) is already being successfully employed to aid the interpretation of multiple facets of burns care. In light of the growing influence of AI this systematic review and diagnostic test accuracy meta-analyses aims to appraise and summarise the current direction of research in this field.

Method
A systematic literature review was conducted of relevant studies published between 1990 to 2021 yielding 35 studies. 12 studies were suitable for a Diagnostic Test Meta-Analyses.

Results
The studies generally focussed on burn depth (Accuracy 68.9%-95.4%, Sensitivity 90.8% Specificity 84.4%), burn segmentation (Accuracy 76.0%-99.4%, Sensitivity 97.9% and specificity 97.6%) and burn related mortality (Accuracy >90%-97.5% Sensitivity 92.9% and specificity 93.4%). Neural networks were the most common machine learning algorithm utilised in 69% of the studies. The QUADAS-2 tool identified significant heterogeneity between studies.

Discussion

The potential application of AI in the management of burns patients is promising, especially given its propitious results across a spectrum of dimensions, including burn depth, size, mortality, related sepsis, and acute kidney injuries. The accuracy of the results analysed within this study are comparable to current practices in burns care.

Conclusion

The application of AI in the treatment and management of burns patients, as a series of point of care diagnostic adjuncts, is promising. Whilst AI is a potentially valuable tool a full evaluation of its current utility and potential is limited by significant variations in research methodology and reporting.

Key Words
Artificial Intelligence
Machine Learning
Burns
Diagnostic Test Meta Analyses
Systematic Review

Introduction

The World Health organisation estimates that 300,000 patients die from burns each year (1). Significant morbidity is associated with both acute and chronic sequelae of large burns. Inaccurate diagnosis when treating burns patients can result in fatal consequences or a plethora of long-term complications; often the first healthcare professional to assess and manage a burns patient is not part of the Burns Team.

Machine Learning (ML) is a subset of Artificial Intelligence (AI), pertains algorithms able to detect patterns directly from data, a task, typically, beyond the capabilities of humans or traditional statistical methods (2). ML algorithms are generally classified into supervised and unsupervised approaches. The former class is trained on datasets with known results, sometimes termed ground truth, and aims to predict results on previously unseen datasets while the latter attempts to identify patterns, clusters or associations within unlabelled data (3). Advances in image analysis have meant that ML algorithms can accurately segment images and identify regions of interest such as burnt tissue to allow estimates of burn size/ segmentation and or focus on these areas to estimate a burn depth(4).

Healthcare professionals are starting to adopt ML-based approaches in the field of burns. These algorithms have been shown to accurately detect burn depths, survivability of burns and anticipate patients being at risk of sepsis or acute kidney injury faster and more accurately than current guidelines and technologies (5–8).

Despite the growing influence of AI in the field of burns, to the best of our knowledge, there is no overview of the research to date. This is vital if healthcare professionals are to interpret and apply ML principles in clinical practice. This systematic review and diagnostic test accuracy meta-analyses aims to appraise the current evidence base for the use of AI in the management and treatment of burns.

Methodology

In this systematic review all studies that compared the performance of machine learning algorithms to a reference standard were selected for inclusion across any original study in the field of burns:

- The population received one or multiple index tests compared to a single reference standard.
- Diagnostic case control studies that selectively recruit patients according to disease status
- Retrospective and prospective studies. This includes studies where previously acquired images are assessed and then prospectively reviewed.

Participants and Target Conditions

Due to the anticipated paucity of evidence all human studies (adult and paediatric) were included regardless of sample size (Table 1).

Study target conditions included, but were not limited to, assessment of burn size including segmentation, burn depth and burn mortality.

Inclusion	Exclusion
Human studies (adult and paediatric)	Animal studies or In Vitro studies
Application of ML in any burns related treatment pathway	Target conditions beyond the scope of burns related care
Original studies	No full text available
	Non- english language studies
	Non-ML studies
	Review articles

Table 1: Inclusion and Exclusion criterion of studies.

Reference Standard and Index Tests

There is great variability in the reference standard for assessing burn depth which includes clinician analysis, laser doppler imaging and time to heal for the burn wounds. Hence the inclusion of variable reference standards is permitted. Similarly, there is no single and generally accepted ML algorithm able to solve the various aspects of burns care. Hence, all ML algorithms were considered in this systematic review.

Search Methodology

All studies published between 1990 to the present date were sought. A comprehensive search of MEDLINE (OVID and SCOPUS) and EMBASE (SCOPUS) was performed using a combination of Medical Subject Headings (MeSH). The search strategy in is shown below.

Burn* OR Plastic Surgery OR Reconstructive Surgery
AND
Artificial Intelligence OR Machine Learning OR Big Data

Burn* OR Plastic Surgery OR Reconstructive Surgery
AND
Neural Network OR Support Vector Machine OR Decision Tree OR Nearest Neighbour

Burn* OR Plastic Surgery OR Reconstructive Surgery
AND
Supervised Learning OR Unsupervised Learning

Burn* OR Plastic Surgery OR Reconstructive Surgery
AND
Computer Aided Diagnosis OR Image Segmentation

Following the search and analysis of the abstracts, studies were populated into Zotero (George Mason University, USA) for full text review. The final list was cross-checked against

a previous systematic review on Artificial Intelligence in Plastic Surgery as well as reference lists of all included studies in the Systematic Review (9).

Data Collection and Analysis

Selection of studies

Both a clinical reviewer and methodological reviewer applied the inclusion criteria to full text articles. These articles formed the corpus of the systematic review. To allow the calculation of sensitivity and specificity, only studies where the true positive (TP), true Negative (TN), false positive (FP) and false negative (FN) could be obtained were included in the diagnostic test meta-analysis (DTAm) part of this study. Study acquisition is illustrated in the PRISMA Flow Chart (Figure 1).

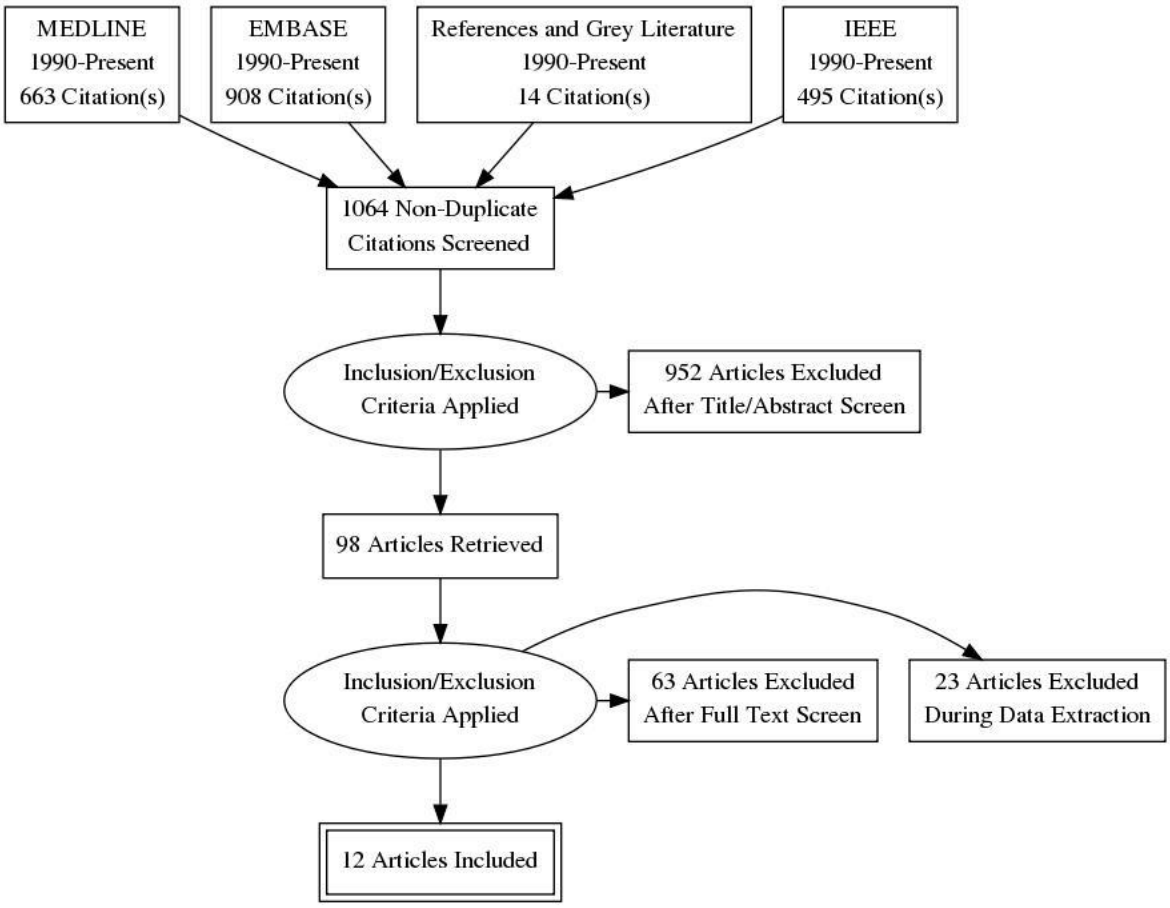


Figure 1: PRISMA Flowchart for the Diagnostic Test Accuracy Meta-Analyses for Artificial Intelligence in the Management and Treatment of Burns. *
*Please note the 12 DTAm articles and the 23 articles excluded during the data extraction make up the 35 articles for the systematic review component. The excluded 23 articles did

not contain enough information for meta-analytical purposes but are useful for a narrative review.

Data Extraction and Management

All studies eligible for the study were reviewed by our team and relevant features were input into our data extraction spreadsheet. Variables collated include the aim of the study, population, the ground truth, the type of ML algorithm employed, input features of the algorithm and the validation method or train and test split of the data as well as the performance results. Multiclass problems, such as depth of burns, were categorised into superficial partial thickness equivalent versus deep dermal and full thickness burns due to the variable nature of the nomenclature used. The mapping of the categories is shown in Table 2. In studies comparing multiple ML algorithms or parameter settings for a single data set, the best performing algorithm and parameters quoted by the paper' authors were chosen. Proof of concept studies where only the validation results are exhibited were included in the systematic review for descriptive analysis.

For five studies where the original confusion matrices are not included in the publication, the 2x2 matrices were reverse engineered based on the data provided Table 2 (6,10–13).

Author	Analysis	Comment	2 Class Mapping	
			Positive (severe)	Negative (non-severe)
Serrano (2005) (14)	Depth		DD , FT	SPT
Abubakar (2020) (15)	Depth		2nd/3rd Degree	1st Degree
Acha (2003) (11)	Depth	Reverse Analysis based on recall result in paper	DD to FT	SPT
Suvarna (2013) (5)	Depth		2nd/3rd Degree	1st Degree
Yadav (2019) (16)	Depth		Skin Graft	Non-grafting
Patil (2009) (17)	Survival		Death	Alive
Abubakar (2019) (18)	Segment		Burn	No burn
Cirillo (2019) (19)	Depth		DPT+FT	SPT+IPT
Dubey (2019) (10)	Segment	Reverse analysis estimated on 40% test set	Burn	No burn
Stylianou (2015) (6)	Survival	Reverse analysis estimated on averages results ANN (Sensitivity=Specificity)	Death	Alive
Kuan (2018) (12)	Depth	Reverse analysis estimated on 30% test set	DD , FT	SPT
Wang (2020) (13)	Depth	Reverse analysis estimated on 15% test set	Moderate, Deep	Shallow

Table 2: The mapping of discriminators for the 12 studies included in the meta-analysis and identification of those studies from which the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were reverse engineered based on data from the paper.

* Deep Dermal (DD), Full Thickness (FT), Superficial Partial Thickness (SPT), Deep Partial Thickness (DPT), Intermediate Partial Thickness (IPT)

Assessment of Methodological Quality

The risk of bias and applicability of included studies in the meta-analysis was assessed using the COCHRANE group recommended QUADAS-2 (Quality Assessment of Diagnostic accuracy Studies-2) tool (19).

Statistical Analysis

Due to the scarcity of the evidence only 8/12 studies, on burn depth, underwent a full DTAm. All analyses were performed using MetaDTA, an online application which uses a random effects bivariate binomial model where sensitivity and specificity are jointly modelled (20,21). The hierarchical summary receiver operating characteristic (HSROC) parameters were estimated using the bivariate model parameters. Estimates of I^2 statistics were not calculated as they do not account for the heterogeneity of the different positivity thresholds (22). Hence heterogeneity was interpreted through how close the studies lay near the HSROC curve line of best fit and comparisons between the 95% prediction and confidence interval ellipse around the summary point (summary sensitivity and specificity modelled on the bivariate model at a notional threshold). The 95% predictive region is useful when estimating where future studies would lie on the HSROC curve graph. Global test measures fail to distinguish between false positive and false negative results hence ranges were provided where applicable (23). Since sensitivity and specificity for each study are interpreted over different thresholds comparing different data sets single summary analysis should be viewed with caution. Hence, coupled Forest Plots demonstrating the sensitivity and specificity for each study with 95% confidence intervals were depicted.

Co-variates, such as ML algorithm type, image or patient selection, QUADAS-2 score and whether a reverse analysis was conducted, were used for the sensitivity analysis in this DTAm.

For illustrative purposes the burn survival (n= 2) and burn segmentation (n=2) studies' sensitivity and specificity estimates were also plotted on coupled forest plots.

Results

A total of 35 studies examining the use of AI/ML in the management and treatment of burns in humans from 1996 till 2021 were considered (Table 3). The studies generally focus on burn depth (n=15), mortality (n=6), segmentation (n=5), infection (n=3) and miscellaneous (kidney injury, burns vs bruises, burns vs pressure ulcers, scarring and open wounds n=6). Only one study focussed solely on paediatric patients (19).

Accuracy of burn depth prediction ranges from 68.9% to 95.4%. Similarly, burn segmentation accuracies vary 76.0%-99.4%. Ache et al. demonstrated that burn depth predictions improves when the multiclass problem is mapped into a two-class discriminator (11). Burn mortality prediction is more conclusive with accuracies >90%-97.5%.

69% (24/35) of studies use some form of artificial neural network either as a standalone algorithm, feature selector for input into the rest of the algorithm or in multi-ensemble algorithms. Different types of neural networks exist with fuzzy logic neural networks used to solve burn depth analysis or for limited datasets 'off the shelf' neural networks typically of the ResNet or VGG variety. Other popular machine learning algorithms include k-Nearest Neighbour and Support Vector Machines.

Three studies compare machine learning algorithms to logistic regression with comparable results for burn depth and mortality. However, when predicting the plasma concentration of antibiotics in burns patients, ML algorithms were found to exhibit a superior performance (6,12,24). Tran et al. and Jimenez et al., used automated and evolutionary algorithms which demonstrated superior diagnostic test accuracies and less features were selected when compared to more traditional machine learning algorithms). It is also worth noting that Yeong et al. and Dubey et al. combine existing medical imaging technologies (Optical Coherence Tomography and Reflectance Spectrometry) with ML algorithms to improve feature selection and accuracy of burn depth and segmentation (10,25).

Author	Aim	Study Type	Population	Ground Truth	Machine Algorithm +/- Comparators	Input	Output	Train/Validation / Test Split	Summary Statistics	Conclusion
Chauhan 2020 India(26)	To create a customised body part-specific artificial intelligence model for burn severity assessment from colour images	Retrospective	Non Burn Body Part Images- Datasets Burn Images- Google Images and iStock Images Total Augmented Burn Images Training 396 from 141 and Test 63 images	Medical Expert Review of Images	CNN (ResNet 50) to identify body part then SVM to categorise depth	Image of body part defined by CNN	Burn Depth- Mild, moderate and severe	70% Training and Validation 30% Test Random Sampling	Sensitivity= 77.60% PPV= 78.0% Average Accuracy= 84.9% F1 Score77.8%	A highly customised body part specific approach that could be used to deal burn region segmentation issues.
Chauhan 2020 India (27)	To demonstrate the use of advanced AI techniques for colour images-based burn region segmentation	Retrospective	434 Images from Google Images and iStock photos	Medical Expert Review of Images	ResNet 101 Atrous Convolution CNN	Image	Burn Segmentation	Training / Validation= 316/76 Test= 42	Sensitivity= 83.4% Specificity= 95.7% Accuracy= 93.4%	The paper demonstrates an atrous convolution neural network for efficient automated burn region segmentation.
Yadav 2019 India (16)	To develop an automated classification system for burn identification as graft and non-graft	Retrospective	74 Images (increased to pick out areas of interest see test/train) Derived from Virgen del Rocío Hospital database	Assupmtion that it is similar to other Spanish papers from the region 1 week Clinician Follow Up	SVM	Image Channels (Hog Featured Image, Hue, Kurotsis, Chroma, Skewness)	Class I- Superficial Dermal No Graft, Class II- Deep Dermal or Full Thickness Requiring Graft	Training=20 Test= 74	Sensitivity= 87.8% Specificity= 83.3% Accuracy= 82.4%	The novelty of the work lies within the automatic burn classification based on different features.
Khan 2020 Pakistan (28)	The main objectives of this research work are to segment the burn wounds and classification of burn depths into 1st, 2nd and 3rd degrees respectively	Retrospective	450 Images 3-52 years old Burn Centre of Allied Hospital Faisalabad and Internet images	Medical Expert Review of Images	CNN	Image Segmentation and then feature extraction based on colour, texture and shape	Burn Depth 1st-3rd degree	Training= 65% Test= 35%	Overall Accuracy79.4% 1st Degree 87.7% Second Degree 84.5% Third Degree 66%	From the obtained results of this research work, non- expert doctors will be able to apply the correct first treatment for the quality evaluation of burn depths.
Kuan 2018 Malaysia (12)	The objective of this paper is to conduct a comparative study of different types of classification algorithms on the classification	Retrospective	164 Images Hospital	Medical Expert Collated Images	Logistic or Multi-Class Classifier	Image Colour and Texture	Burn Depth (Superficial, Superficial Partial Thickness, Deep Dermal, Full Thickness)	Training = 70% Testing = 30% 10 fold Cross Validation	Logistic or Multi-Class Classifier Average Accuracy = 68.9% Superficial Partial Thickness= 62.2% Deep Partial Thickness= 68.9% Full Thickness= 75.6%	Simple logistic regression analysis provided comparable results to logistic regression and the multiclass classifier.

	of different burn depths by using an image mining approach									
Suvarna 2013 India (5)	To develop an automatic skin burn wound analyser to aid the diagnosis of burn victims	Retrospective	120 Images Hospital and Internet	Medical Expert Review of Images	Template Matching vs k-NN and SVM	Image	Burn Depth Grades 1 (Superficial Dermal), Grade 2 (Partial Thickness) and Grade 3 (Full Thickness)	3 fold Cross Validation	SVM Overall Accuracy 90% Grade 1 85% Grade 2 87.5% Grade 3 92.5%	The novelty of the work reported here is that it is simple, can be built using readily available devices such as camera and laptop and can therefore be used as some form of telemedicine.
Wantanajittikul 2011 Thailand(29)	To develop an automatic system with the ability of providing the first assessment to burn injury from burn colour images	Retrospective	5 patient images from Department of Health from which there are 34 derived segmented mages	Medical Expert Review of Images	SVM	Burn Images	Burn Depth - 2nd and 3rd Degree	Training 80% (4 images) Test 20% (1 image) Cross Validation 4-fold	Accuracy= 75.3%	SVM yielded the best results over k-NN and Bayes classifiers.
Jiao 2019 China (30)	To develop a deep learning technology to diagnose burn size and depth via segmentation of images	Retrospective	1150 Images of 'fresh' burn wounds from Wuhan Hospital No 3	Medical Expert Review of Images	CNN (ResNet 101) used to train SVM	Burn Images	Burn Depth (Superficial, SPT, DD, FT) Burn Size (<5%, 5% < %TBSA< 20%, %TBSA> 20%)	Training= 1000 Pictures Test= 150 pictures	Average Dice's Co-Efficient (Burn Size) = 83.7% Average Dice's Co-Efficient (Burn Depth)= 85.14% Paper surmises 84.51% total accuracy	This article proposes a new segmentation framework for burn images based on deep learning technology.
Abubakar 2019 United Kingdom (18)	To discriminate between unburnt and burnt skin	Retrospective	680 burnt skin and 680 of healthy skin Internet Image Acquisition	Author Review	CNN ResNet 101 with SVM	Images	Healthy or Burnt skin	Pre trained model	Sensitivity = 99.4% Specificity = 99.6% Accuracy=99.5%	This study has shown that machine learning can be used to discriminate between skin burn injury and normal skin with high accuracy
Abubakar 2019 (Transfer Learning) United Kingdom (31)	To distinguish between burnt skin and bruises	Retrospective	32 Burns Images from Bradford Teaching Hospital scaled to 700 regions of interest 100 Skin bruise images scaled to 900 images	Medical Expert Review of Images	CNN ResNet 152 with SVM	CNN feature extraction	Burn or Bruise	Pre trained model- Transfer learning	Sensitivity= 100% Specificity= 100% Accuracy= 100%	SVM achieved perfect identification accuracy with ImageNet model features.

Abubakar (Burns Depth) 2020 United Kingdom (15)	To develop a native burn depth evaluation using deep learning features to objectively predict those burns that that require surgical intervention and those that do not	Retrospective	Images from the internet (1st Degree) and Bradford Teaching Hospitals (2nd and 3rd Degree) Pre-augmented 743 Post augmented 1560	Laser Doppler Imaging (2nd and 3rd Degree Burns)	CNN ResNet 50 used to train SVM	CNN feature extraction	Burn Depth (Normal Skin, 1st Degree-3rd Degree Burns)	10 fold Cross Validation	Accuracy= 95.4%	State of the art prediction accuracy with a decision made in less than a minute whether the burn injury required surgical intervention such as skin grafting or not.
Abubakar (Caucasian vs African skin burns) 2020 United Kingdom (32)	Comparative analysis of healthy skin versus burnt skin in Caucausian and Africans	Retrospective	32 Caucasian burn patients (1360 augmentation) from Bradford Teaching Hospitals 60 African burn patients (700 augmentation) from Federal Teaching Hospital Gombe	Author and Medical Experts	CNN VGG-16 used to train SVM	CNN feature extraction	Burnt Skin or Healthy Skin	10 fold Cross Validation	Caucasian Skin Sensitivity= 98.9% Specificity= 99.6% Accuracy= 99.3% African Skin Accuracy Sensitivity= 98.5% Specificity= 99.3% Accuracy= 98.9% Combined Accuracy = 98.8%	Training local models for each ethnic group (or race) tends to be more robust than a single global model for all skin colours.
Abubakar (Pressure Ulcer vs Burn) 2020 United Kingdom (33)	ML to differentiate between a burn and a pressure ulcer	Retrospective	29 Pressure Ulcer Images from the Internet augmented to 990 31 burn images from a local teaching hospital augmented to 990 990 images of healthy skin	Author and Medical Experts	CNN ResNet 152 with SVM classifier	CNN feature extraction	Healthy skin, Burn or Pressure Ulcer	10 fold Cross Validation	Sensitivity= 99.1% Specificity= 100% Accuracy= 99.6%	Machine learning can be used to diagnose skin abnormalities.
Acha 2003 Spain (11)	To separate burn wounds from healthy skin, and the different types of burns (burn depths) from each other	Prospective Component	62 patient images taken on admission from Virgen del Rocío Hospital	Medical Expert Follow Up (Time to Heal)	Fuzzy-ARTMAP Neural Network	Segmented Burn Images	Burn Depth (Superficial Dermal, Deep Dermal and Full Thickness)	Training=250 49x49 images Test = 62 images <24 hours 5 fold Cross Validation	Average Accuracy= 82.3%	The study has a very good method for segmenting and classifying images into their burn depths.
Acha 2013 Spain (34)	A psychophysical experiment and a multidimensional	Prospective Component	94 patients images taken on admission from Virgen	Medical Expert Follow Up (Time to Heal)	k-NN	Feature Extractions from prior	Burn Depth grafts vs non grafts and 3 depths	Train = 20 Test= 74	2 Depth Results Sensitivity = 71% Specificity= 94% PPV= 93%	Mathematical features extracted from the psychophysical experiments can help classify burns depths.

	scaling (MDS) analysis are undergone to determine the physical characteristics that physicians employ to diagnose a burn depth. Subsequently, these characteristics are translated into mathematical features to classify burns .		del Rocío Hospital			experiment (psychosocial)	(Superficial Dermal, Deep Dermal and Full Thickness)	Leave on out validation	NPV=75% Accuracy= 83.8%	
Serrano 2005 Spain (14)	To create a computer aided diagnosis tool for the classification of burns	Prospective Component	35 burn images <24 hours from injury from Virgen del Rocío Hospital	Clinician 1 week follow up	Fuzzy-ARTMAP Neural Network	Manually segmented burns images then 6 input descriptors; Lightness, hue, standard deviation of the hue component, u* chrominance component, standard deviation of the v component and skewness of lightness	Burn Depth (Superficial Dermal, Deep Dermal and Full Thickness)	/	Accuracy Superficial Dermal= 100% Deep Dermal= 84.6% Full Thickness= 77.8% Overall 88.6%	An automatic system to classify burns into their depths is proposed using digital photographs taken by clinicians using a pre-approved protocol.
Badea 2016 Romania (35)	To describe a convolutional neural network approach to the identification of burn areas from color image patches	Retrospective	53 patients (611 Images and 200494 patches) 1 day-62 days old burns 8 months-17 years old	Medical Expert Review of Images	CNN	Burn Images	Burn Segmentation	Training= 37% (74763) Test =63% (125731)	Accuracy 75.9%	The proposed approach achieves an overall performance comparable to the literature-reported average performance of a specialist surgeon.
Yeong 2005 Taiwan (25)	To develop a non-invasive objective system for the prediction of burn healing time, based on optical information using reflectance spectormetry	Retrospective	35 Burn wounds in 35 patients 6-79 years old Assessment 3-4 days after initial burn Mechanism includes Scald, contact and flame Male: Female 1.2:1	Indirect healing within or more than 14 days	Radial Basis Function Neural Network	Spectroscopic Analysis	Burn Healing Time <14 days	Leave one out cross validation	Accuracy= 86% Sensitivity= 75% Specificity=97%	The authors have developed a non-invasive burn depth assessment tool that can be used by inexperienced clinicians.

Cirillo 2019 Sweden (19)	To predict time independent Burn Depth using Artificial Intelligence	Not clearly stated	23 burn images used to extract 676 different regions of interest Images were obtained <2 days, 2-4 days, 5-7 days Age <16 years old Linköping University Hospital	Time to Heal superficial partial thickness healed within 7 days, superficial to intermediate partial thickness healed between 8 and 13 days, intermediate to deep partial thickness healed within 14 to 20 days, and deep or full-thickness burn healed after ≫21 days or underwent surgery. Perfusion images if available were also utilised.	CNN ResNet 101	Burn Image	Intermediate partial thickness, intermediate to deep par tial thickness, deep partial and full thickness, normal skin, and background	Pre-Trained Models 10 fold Cross Validation	Accuracy= 90.5% Sensitivity= 74.4% Specificity=94.3% Augmented values were lower	The application of AI with state-of-the-art CNNs is a useful tool in guiding initial treatment of burn wounds. The next step forward is semantic segmentation so clinicians can also obtain the burn size as well as the depth.
Acha Pinero 2005 Spain (36)	The aim of the system is to separate burn wounds from healthy skin, and to distinguish among the different types of burns (burn depths)	Prospective Component	62 patient images taken on admission from Virgen del Rocío Hospital	Medical Expert Follow Up (Time to Heal)	Fuzzy Artmap Neural Network	lightness, hue, standard deviation of the hue component, u* chrominance component, standard deviation of the v* component, and skewness of lightness	Burn depth (Superficial dermal, Deep Dermal and Full Thickness)	Training= 250 images obtained prior Testing= 62 images 5 fold cross validation	Accuracy Superficial Dermal = 86.36% Deep Dermal= 83.33% Full Thickness= 77.27%	This paper describes an effective burn colour image segmentation and machine learning burn depth classification system.
Wang 2020 China (13)	To improve early judgement of burn depth	Retrospectiv e	484 early wound photos segmented to 5637 images <80 years old Images taken within 48 hours Across 5 hospitals	Physician assessed time to heal (Shallow 0-10 days, Moderate 11-20 days, >20days deep or those requiring skin grafts)	CNN ResNet 50	Burn Images	Burn Depth	Training 70%, Validation 15%, Test 15%	Sensitivity= Shallow 73%, Moderate 81%, Deep 93% PPV= Shallow 84%, Moderate 81%, Deep 82% Overall AUC= 95%	This study describes a new method to diagnose burns using artificial intelligence and burn images.

Dubey 2018 India (10)	To analyse data from Optical coherence tomography (OCT) of burnt skin using machine learning	Prospective	68 human skin tissue (34 Normal Skin and 34 Burnt Skin) Chemical, electrical or fire	Author Analysis	Decision Tree, Neural Network, Random Forest, Extreme Learning Machine (ELM), Average Neural Network (avNNet) and SVM via a 3 tier Multi-Ensemble model	Features extracted from OCT analysis using Principal Component Analysis	Burn Segmentation	Training 60% Test 40% 10 fold Cross Validation	Sensitivity= 92.8% Specificity= 92.3% Accuracy= 92.5%	The paper demonstrates that a multi-ensemble classifier is able to detect abnormality burnt human skin in vivo using data from OCT.
Estahbanati 2002 Iran (37)	To create an Artificial Neural Network to predict survival of burns patients	Retrospective	1082 patients admitted to a burns centre in Tehran <10 to >90 years old 60% female and 40% male TBSA 5-100%	Hospital records	NN	age, sex, TBSA, data of admission (month and season of year the burn in-jury has been incorporated, since burn injury is more common and often more extensive in the cold season), lapse time (time from burn to admission to hospital), refereed or non-refereed status, inhalation injury, haematology and bio-chemistry lab values, medical outcome, number and type of surgical episodes (debridement or skin graft) were obtained	Burn Mortality	Training 75% Test 25%	Sensitivity= 80% Accuracy= >90%	This study describes an Artificial NN used for prediction of mortality in burn patients.
Cobb 2018 USA (38)	To identify predictors of survival for burn patients at the	Retrospective	31350 patients 670 hospitals Mean Age	Hospital Records	Stochastic gradient boosting (SGB)	Top 5 for SBG model were younger age, absence of	Burns Mortality	Training 66% Test 34%	SBG Specificty= 74% RF Specificity= 71%	Patient and hospital factors are predictive of survival in burn patients. It is difficult to control patient factors, but hospital factors, better

	patient and hospital level using machine learning technique		40.5 years old Burn Mortality California 2006-2011, Florida 2009-2013 and New York 2009-2013 registries		decision tree and random forest (RF)	electrolyte imbalance or coagulopathy, admission on a weekend, and absence of renal failure Top 5 for RF Model were absence of electrolyte imbalance or coagulopathy, younger age, absence of congestive heart failure, and presence of weight loss.			SBG AUC= 93% RF AUC= 90%	predicted by RF, can inform decisions about where burn patients should be treated.
Frye 1996 USA (39)	Whether artificial intelligence could predict the burn outcome and length of stay (weeks) for patients	Retrospective	1585 patients 1988-1995 South Alalabama Medical Centre	Hospital Records	NN	Age, Sex, TBSA, Transport to Hospital, Location at time of Injury, Inhalation +/- Length of Stay	Burns Mortality and Length of Stay	Training 90%= 1420 Testing 10%= 158	Accuracy Length of Stay 72% Mortality 95%	Increasing age, TBSA and an inhalational component had the greatest impact on mortality whereas area of burn impacted on length of stay.
Jimenez 2014 Spain (40)	To describe a novel rule-based fuzzy classification methodology for survival/mortality prediction in severely burnt patients	Retrospective	99 ICU Patients 1999-2002	Hospital Records	Traditional ML algorithms (Fuzzy classifier, DT, NB, ANN) vs Evolutionary algorithm for diversity reinforcement (ENORA) and the non-dominated sorting genetic algorithm (NSGA-II)	Burn Size, Burn Depth, Infection, Age, Weight, Sex and Co-Morbidities	Burns Mortality and Length of Stay	10 fold Cross Validation	ENORA Accuracy= 92.3% Sensitivity= 93.6% Specificity= 93.9%	Evolutionary algorithms improve the accuracy and interpretability of the classifiers, compared with other non-evolutionary techniques.
Stylianou 2015 United Kingdom (6)	A comparison of logistic regression and machine learning for predicting burns mortality	Retrospective	65764 Burns Patients (Flame, flash, scald, contact, chemical and other)	iBiD Hospital Database	Artificial neural network, support vector machine, random forests and naive Bayes vs Logistic Regression	Classifier tuning with age, age squared and TBSA	Burns Mortality	Train 70%= 46626 Test 30% = 19985 Data resampled 10 times	Sensitvity=Specificity Sensitivity= ANN- 92.2% LR-91.9% Specificity= ANN- 93.4% LR- 92.3% AUC ANN-97.4% LR- 97.1%	An established logistic regression model performs as well as more complex machine learning methods.

Patil 2009 India (17)	To predict the survivability of the burn	Retrospective	180 Patients 2002-2006 Swami Ramanand Tirth Hospital	Hospital Records	Decision Tree	Age, sex and size of burn for 8 different body parts	Burns Mortality	Train 58%= 104 Test 42%= 76 10 fold Cross Validation	Sensitivity = 97.5% Specificity= 97.2% Accuracy= 97.4%	The results may be further improved by depth of burn, heat source of burn and pre-existing diseases.
Tran 2019 USA (41)	Evaluate the clinical utility for ML in augmenting the predictive power of both traditional and novel indicators of acute kidney injury (AKI) within 24 hours.	Retrospective	n=50 >18 years old >=20% TBSA Burns ICU	Kidney Disease: Improving Global Outcomes Criteria	k-NN	Neutrophil Gelatinase associated Lipocalin (NGAL), N-terminal B-type natriuretic peptide , creatinine or urine output <24 hours	Predication of AKI within 24 hours of admission	60-80% Train 20-40% Validation Up to 10 Fold Cross Validation	NGAL, creatinine, UOP, and NT-proBNP average cross validation accuracy 98%	Performance of urine output and creatinine for predicting AKI can be enhanced with a ML algorithm using a k-NN approach.
Rashidi 2020 USA(8)	To determine if a burn-trained ML algorithm could be generalized to a non-burned population and evaluate the value of including novel renal injury biomarker combinations to enhance AKI prediction <24 hours	Retrospective Cohort and Prospective Cohort	Cohort A n=50 >18 years old >=20% TBSA Burns ICU Cohort B n=51 with >=20% TBSA or non burn related trauma	Kidney Disease: Improving Global Outcomes Criteria	5 ML Algorithms trialled Logistic regression (LR), k-NN, RF, SVM and a multi-layer perceptron (MLP) deep neural network (DNN)	Neutrophil Gelatinase associated Lipocalin (NGAL), N-terminal B-type natriuretic peptide , creatinine and urine output	Predication of AKI within 24 hours of admission	Training= Cohort A n=50 Test= Cohort B n=51 Cross Validation	Accuracy DNN (NGAL, urine output and creatinine)= 100%	The AI algorithm helped predict AKI 61.8 (32.5) hours faster than the Kidney Disease and Improving Global Disease Outcomes (KDIGO) criteria for burn and non-burned trauma patients.
Tran 2020 USA (7)	Traditional indicators of sepsis exhibit poor predictive performance. To address this challenge, we developed the Machine intelligence Learning optimizer (MILO)	Retrospective	211 (92 Septic) >18 years old >=20% TBSA Across 5 Hospital ICUs	Sepsis as defined by the 2007 American Burn Association guidelines	Multivariate Logistic Regression vs Traditional Machine Learning Algorithms vs Automated ML (MILO)- A combination of supervised and unsupervised algorithms.	MILO= 5 variables (Heart Rate, body temperature, haemoglobin, blood urea nitrogen (BUN), and total CO2) Traditional ML Logistic Regression= 16 variables (MAP, RR, body temperature, GCS, WBCs, Hb, HCT, platelets, Na+, K+, BUN,, creatinine, BUN/creatinine , glucose, total CO_2, and MOD score Multivariate Logistic Regression= 7	Sepsis	80% Training 20% Validation Cross Validation	Multivariate Logistic Regression AUC= 88% Accuracy = 86% AUC = 96% Senstivity= 98% Specificity = 82% MILO k-NN Accuracy = 90% AUC = 96% Senstivity = 96% Specificity = 88%	The deployment of MILO not only accelerates the development of ML models, but quickly helps identify optimal features and algorithms for burn sepsis prediction.

						variables (WBC, Hb, HCT, Sodium, and Platelets as predictors of sepsis)				
Yamamura 2004 Japan (42)	NN to determine the non linear relationship between aminoglycoside concentration and burn severity	Retrospective	30 patients 23-85 years old 1993-2003	Concentration of arbekacin in plasma via immunoassay	NN	Dose, body mass index, parenteral fluid and creatinine concentration provides information about peak plasma concentration in addition to one for burn severity (burn area except for area of skin graft)	Plasma concentration of arbekacin	Leave one out cross validation	Rate of Corrective Prediction=86.6% Sensitivity=66.7% Specificity=95.2%	The artificial neural network model showed superior predictive performance for plasma concentration prediction of arbekacin based on patents' physiological parameters compared to predictions made using a linear model.
Yamamura 2008 Japan (24)	To predict the response of aminoglycoside antibiotics (arbekacin: ABK) against methicillin-resistant Staphylococcus aureus (MRSA) infection in burn patients	Retrospective	25 patients Age 23-85 TBSA 53.7 +/- 19.8 (11.5-80.1) Nippon University Medical Hospital	Assessment of laboratory parameters white blood cells, C reactive protein and number of bacteria	NN vs Logistic Regression	Maximum concentration of Arbekacin Sulfate, serum creatinine, duration of dose, blood sugar, burn area after skin graft operation	Response of Arbekacin Sulfate against MRSA	Leave one out cross validation	Accuracy = 88% (note logistic regression was 60%)	Based on the patients' physiologic data, ANN modelling would be useful for the prediction of the ABK response in burn patients with MRSA infection
Berchialla 2014 Italy (43)	To study the factors associated with an increased risk for developing pathological scarring after burns	Retrospective	752	Medical Experts from Standard Reporting Form	Bayesian Network	Gender, Age, TBSA, Full Thickness TBSA, Burn Mechanism, Anatomical area, Surgical or Medical Treatment, Number of Surgical Procedures, Type of Surgical Approach, Type of Skin Graft (Mesh), Wound Healing Time, Excision and Coverage Time, Scar Type	Normal vs Pathological Scar Type Probability	Training 703 patients=2440 Anatomical Burn Sites Validation 49 patients = 162 anatomical sites	Type of Surgical Approach, Number of Surgical Procedures and Burn Treatment impact the outcome node the most Error Rate for Validation sample 24.8%	The Bayesian Network output can support the physician in the prognosis of hypertrophic scars. It can vary the clinical scenario to provide detailed prognostic information.

Liu 2018 USA (44)	Predicting the ability of wounds to heal given any burn size and fluid volume	Retrospectiv e	121 patients >18 years old >20% TBSA	Surgeon Analysis	Decision Tree and Neural Network	Days since admission, Fluid (L), TBSA and Age	Open wound size ((Sum of TBSA + Surface Area of donor sites) - SA Healed))	10 fold Cross Validation	>90% Goodness of fit <4% abolsute error for combined Decision Tree and Neural Network Model	ML performed better than the traditional statistical methods employed using a four-variable analysis. Further the ML was able to differentiate between survivors and non survivors by their wound healing rates.

Table 3: Summary table of 35 articles that demonstrate the use of Artificial Intelligence in the Management and Treatment of Burns in humans.

* Artificial Intelligence (AI), Machine Learning (ML), Intensive Care Unit (ICU), k nearest neighbour (k-NN), Convolutional Neural Network (CNN), Support Vector Machine (SVM), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Logistic Regression (LR), Area Under Curve (AUC), Total Burn Surface Area (TBSA), Neural Network (NN), Random Forest (RF), Mean Arterial Pressure (MAP), Respiratory Rate (RR), Glasgow Coma Scale (GCS), White Blood Cells (WBC), Haemoglobin (Hb), Haematocrit (HCT), Blood Urea Nitrogen (BUN), Multiple Organ Dysfunction (MOD), LoS (Length of Stay)

Where applicable only the highest scoring machine learning algorithm is depicted in the results table.

Meta-Analysis

Burn Depth

A total of 8 studies were included in the diagnostic test accuracy meta-analysis. Table 4 illustrates the weighted sensitivities and specificities.

Author	Year	TP	FN	FP	TN	N	Sens	Spec	Weight_Spec	Weight_Sens
Serrano	2005	18	4	0	13	35	0.82	1.00	6.97	9.27
Abubakar	2020	1002	29	17	503	1551	0.97	0.97	15.70	14.35
Suvarna	2013	67	5	6	28	106	0.93	0.82	10.02	13.08
Yadav	2019	36	5	8	25	74	0.88	0.76	11.43	12.52
Acha	2003	19	3	5	35	62	0.86	0.88	11.61	9.87
Cirillo	2019	209	25	32	183	449	0.89	0.85	16.20	14.26
Kuan	2017	15	8	9	13	45	0.65	0.59	12.19	11.93
Wang	2020	549	36	70	190	845	0.94	0.73	15.88	14.73

Table 4: The weighted sensitivity and specificity of each of the burn depth studies included in the analysis.

The summary point of the 8 studies estimates a sensitivity of 90.8% (95% confidence intervals 84.6%-94.6%) and specificity 84.4% (73.6%-91.3%). The positive post-test likelihood ratio of 5.8 (3.2-10.5) and negative likelihood ratio of 0.11 (0.06-0.20) indicates diagnostic value in these tests. The negative likelihood ratio (for more superficial burns not requiring intervention) is stronger than the positive likelihood ratio. The suggested Cochrane Thresholds are used as a comparator (Positive Likelihood Ratio>10 and Negative Likelihood Ratio <0.1) (23).

Figure 2 allows an appreciation of the heterogeneity between the studies as depicted by the distance from the HSROC curve and the distance from the summary point. The 95% confidence interval doesn't include several studies such as Kuan et al. which has a much lower specificity and sensitivity as well as a large variance, as shown in the coupled Forest Plot (Figure 3), than its counterparts(12). This is reflected on the HSROC curve where it occupies a discrete position away from the rest of the data. The study by Abubakar et al. on the other hand depicts a very high sensitivity and specificity with small variance (Figure 3) suggesting its diagnostic accuracy of burn depth is very high, yet it also lies outside of the 95% confidence interval(15). The much larger predictive region further reinforces the notion that future studies can lie some distance away from the curve influenced perhaps by the heterogeneity in study design and application.

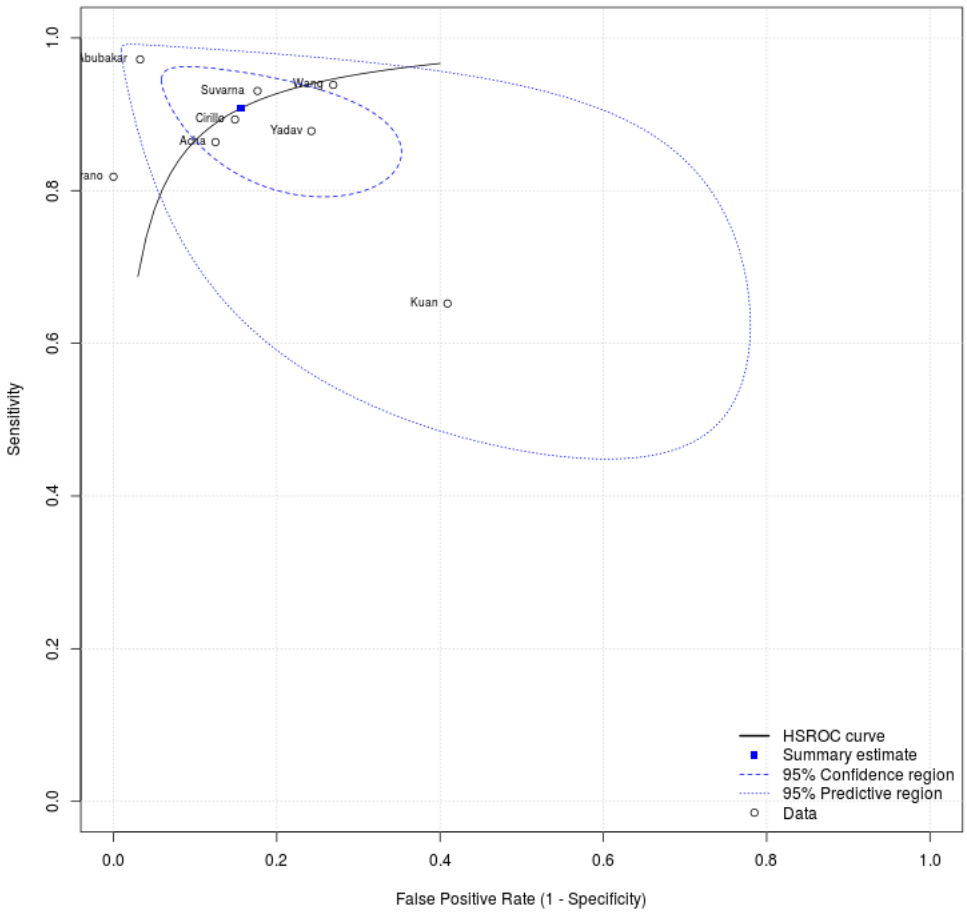


Figure 2: HSROC curve of burn depth (8 studies) the solid point represents the summary estimate surrounded by the larger 95% predictive region and the smaller 95% confidence region.

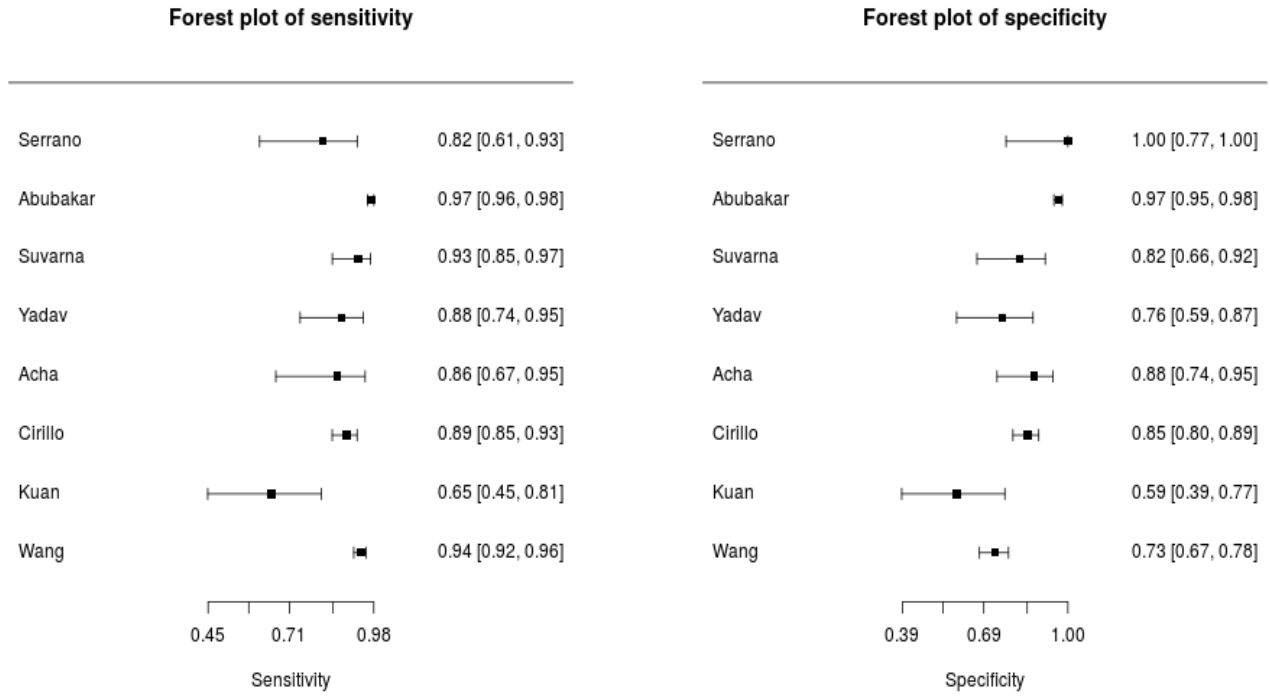


Figure 3: Coupled forest plot of sensitivities and specificities of the burn depth studies included in the Diagnostic Test Accuracy Meta-Analysis with 95% confidence intervals in brackets.

The QUADAS-2 results for each study ('Risk of Bias' and 'Applicability Concern' domains) are highlighted below (Table 5). The seven domains are superimposed onto a table. All studies demonstrated a significant element of uncertainty or a high risk of bias or applicability reflecting the lack of a standardised orthodox approach to reducing bias. The heterogeneity and variable reporting of studies limited the comparisons one can draw from them. A sensitivity analysis is shown in the Supplementary Material based on the QUADAS 2 outcomes and for those studies where we calculated the confusion matrices.

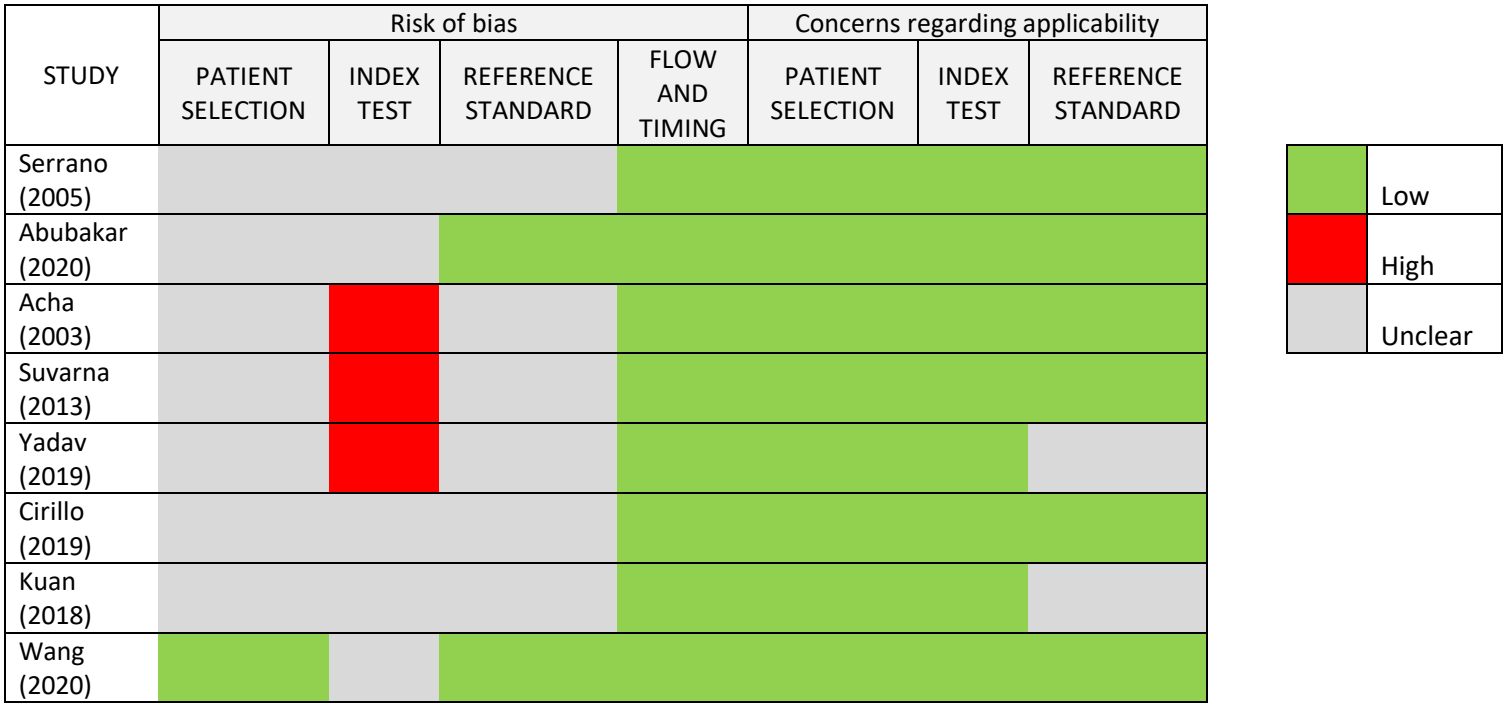


Table 5: QUADAS-2 scores for the Burn Depth studies. Red represents a high risk of bias, green low and grey unclear.

Burn Mortality/Survival

Author	Year	TP	FN	FP	TN	N	Sens	Spec	Weight_Spec	Weight_Sens
Patil	2009	39	1	1	35	76	0.98	0.97	0.18	13.61
Stylianou	2015	234	20	1300	18431	19985	0.92	0.93	99.82	86.40

Table 6: The raw data of each of the burn mortality studies included in the analysis.

The summary sensitivity of 92.9% (89.3%-95.3%) and specificity 93.4% (93.1%-93.8%) with a positive likelihood ratio of 14.1 (13.3-15.0) and negative likelihood ratio of 0.08 (0.05-0.12) suggests that the two studies are accurate at predicting burns mortalities. The raw data

(Table 6) and the forest plots reflect this. The forest plot (Figure 4) for Patil et al. (smaller sample size) shows greater variance of the two studies despite the Stylianou et al. study data being calculated through reverse engineering estimates(6,17).

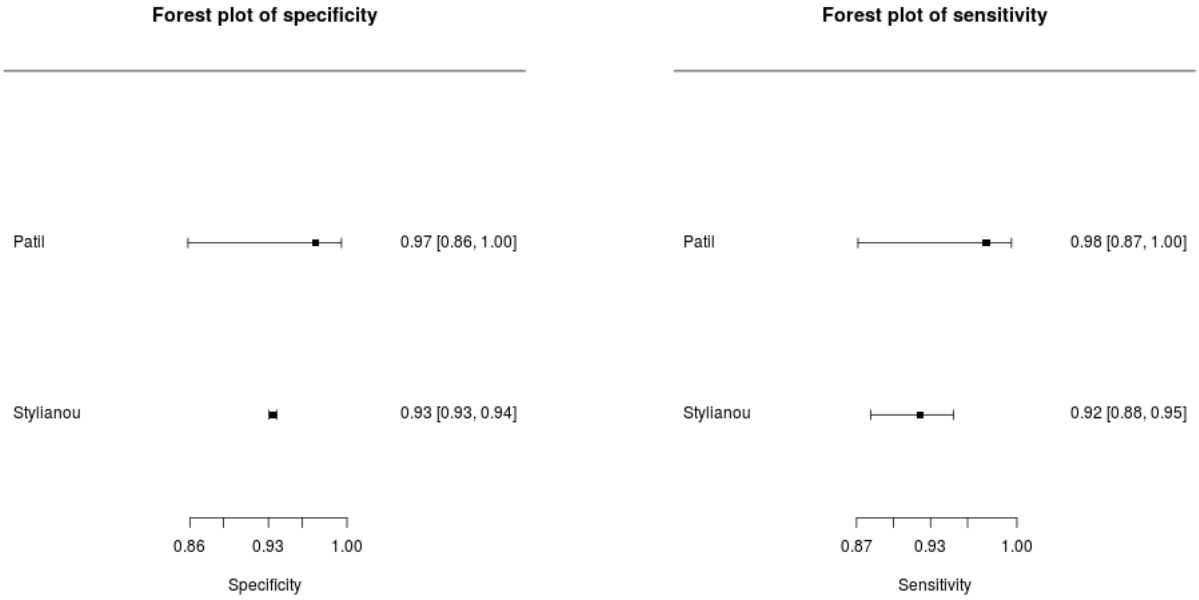


Figure 4: Coupled forest plot of sensitivities and specificities of the mortality studies included in the Diagnostic Test Accuracy Meta-Analysis with 95% confidence intervals in brackets.

Burn Segmentation

Author	Year	TP	FN	FP	TN	N	Sens	Spec	Weight_Spec	Weight_Sens
Abubakar	2019	676	3	4	377	1060	1.00	0.99	51.67	69.32
Dubey	2019	12	2	1	13	28	0.86	0.93	48.33	30.68

Table 7: The raw data of each of the burn segmentation studies included in the analysis.

The study by Abubakar et al shows an extremely high sensitivity and specificity highlighting a very accurate model (CNN ResNet 101 with SVM Classifier) on the test set (Table 7)(18). This is higher than the multi-ensemble algorithm employed by Dubey et al (10).

The summary sensitivity of 97.9% (78.9%-99.8%) and specificity 97.6% (88.4%-99.5%) with a positive likelihood ratio of 40.7 (7.7-215.9) and negative likelihood ratio of 0.02 (0.02-0.26) suggests that the studies are accurate at distinguishing between health and burnt skin.

The forest plot (Figure 5) for Dubey et al. like Patil et al. shows greater variance of the two studies, which is very likely caused by the smaller sample size (10,17). It is worth noting that for the Dubey et al. study data was obtained through reverse engineering the confusion matrix.

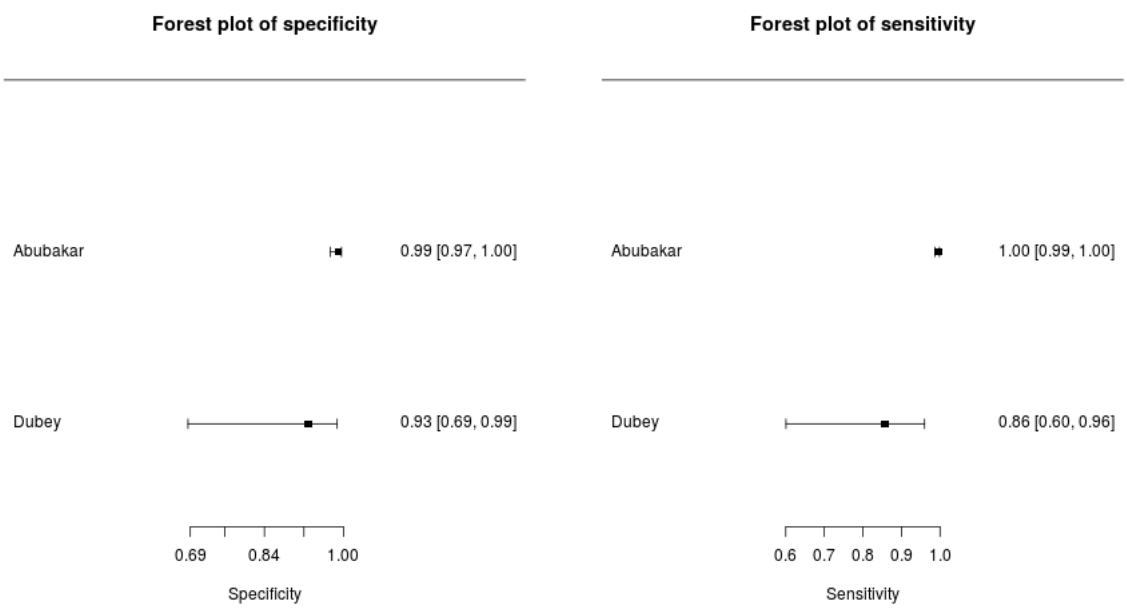


Figure 5: Coupled Forest plot of sensitivities and specificities of the segmentation studies included in the diagnostic test accuracy meta-analysis with 95% confidence intervals in brackets.

Discussion

Artificial Intelligence offers a promising method for the development of diagnostic and predictive tests at several key points in the burns management pathway. From estimating burn depth, size via segmentation, outcomes (including need for surgery, survival and pathological scarring), early prediction of sepsis and acute kidney injuries, machine learning is ubiquitously examining multiple facets of burn care.

The results collated within this study are comparable to current practices in burns care. Burn depth estimations are quoted as typically varying between 64%-84% for experienced clinicians (36,45,46). The 15 burn depth studies included in this review provide a comparable estimate of 68.9%-95.4%.

If only the ML studies that examine acute wounds within 24 hours are included the accuracy range is still between 81.32-88.6% (11,30,34,36) more than the 62% accuracy quoted by Hoeksema et al. for clinicians (47). Assuming that inexperienced physicians, who typically make the initial referrals, have an even lower accuracy figure this would further attenuate clinical decision making. Hence it is not unfathomable to infer that inexperienced clinicians would benefit from a predictive ML tool that can aid their decision making. The summary statistics of 90.8% sensitivity and specificity 84.4% of the 8 burn depth studies included in the meta-analysis are lower than for laser doppler imaging (sensitivity 94.5%, specificity 97.2%)(48). However, laser doppler imaging is typically used after 48 hours and has a high acquisition and maintenance cost. Further several factors can influence the accuracy namely

movement artefact, room temperature, tainting from wound dressings and user error(49). But these technologies can co-exist as there are examples of ML being combined with existing imaging technologies such as reflectance spectrometry and optical coherence tomography(10,25).

For burn mortality and burn segmentation artificial intelligence represents an extremely accurate diagnostic test with summary statistics of sensitivity and specificity of 92.9% and 93.4% the former and 97.9% and 97.6% for the latter. The burns mortality accuracies are particularly promising as the input features are centred around easily obtainable data; age and TBSA amongst others (6,17). The summary statistics presented in this study for predicting survival are higher than the Modified Baux Score, an aid used to ascertain survivability of burns patients (sensitivity: 59.8%; specificity: 82.9%) (50). Burn segmentation, a potential precursor to burn size estimation, is also useful not only as a prognostic factor for survival, but also for accurate estimation of fluid replacement in burns. Overall, burns mortality is reducing with improvements in burn care. However, sepsis remains one of the biggest causes of burn mortality and morbidity. It is difficult to differentiate between a septic response and the inflammatory cascade that burns patients exhibit. Tran et al. were able to exploit this difference with just 5 input features (90% accuracy) using an automated ML algorithm that identified its own patterns within the data which was superior to more traditional statistical methods (7). Acute kidney injuries are another important source of burn morbidity. Rashidi et al. demonstrated that earlier detection is possible, on average 61.8 hours earlier, compared to the application of Kidney Disease and Improving Global Disease Outcomes (KDIGO) criteria(8). This early window for intervention should improve burn patient outcomes, particularly for large resuscitation burns in which several treatments are required to restore fluid homeostasis and organ perfusion.

The use of AI as a diagnostic test has evolved over time. This is evident in the area of image analysis, e.g. for the task of burn segmentation and burn depth prediction. The advent of more powerful hardware and advances in ML algorithms renders AI sometimes superior to clinician-based analyses in certain scenarios (51–53). Before the era of deep learning and CNNs, analysis approaches required the extraction of image features that are subsequently used as input to machine learning algorithms. In contrast, CNNs are able to construct relevant features automatically during the training process. One drawback of neural network-based methods lies with their requirement for large image datasets enabling the learning of these relevant features. The number images required in the model training can be reduced when employing transfer learning, i.e. the use of pre-trained models able to detect relevant image features. Augmentation, re-use of rotated and scaled task-specific images has a similar effect. Both these techniques allow for models to exhibit comparable performances to novel networks trained from new(54).

Heterogeneity of Included Studies

Different studies examining the same topic use different datasets with different ML algorithms, different ground truths, different test/train/validation splits and different outcomes. It is therefore not surprising that significant heterogeneity exists as indicated by

the QUADAS-2 scores and the difference in size between the 95% confidence region and 95% predictive region around the summary points of the HSROC curves.

Table 8 categorises the different forms of methodological heterogeneity.

Input	Algorithm	Reporting
• Image sources range between those from clinical settings to internet searches each with different reliability of ground truths. • For burn depth heterogenous wounds may be more difficult to segment. • Several studies combine cropped images from the same patient in the training and testing cohorts hence this is not independent data and the risk of overfitting might lead to artificially high accuracies. • Some studies employ image augmentation whilst others do not • Variable terminology and thresholds for burn depth are used. • Timing of image acquisition varies between studies as do other features of the acquisition protocol and anatomic region burnt. • Various definitions of ground truth exist between studies. Even if the same definition is used between studies variation may exist due to differences in management protocols. Some studies take the requirement of a skin graft as a ground truth for burn depth analysis. Some units may manage burns more aggressively hence more	• Various different types of ML algorithms used either alone or in combination. • Variability of validation or sampling methods to train/validate/tes, more general cross validation or even external validation.	• Some papers depict their validation results as their test output. • Confusion matrices are not always provided. • Train/Validation/Test split is not always documented.

burns patients will undergo skin grafting in that cohort which will skew the ML output.		

Table 8: Methodological Sources of Heterogeneity

The Double Entendre

Part of the advantages of ML algorithms can be attributed to their generalisability as predictive tools or diagnostic tests. A hypothetical example of this is when a burn depth ML algorithm is applied to a low-resolution image of unknown origin and one taken professionally in a clinical setting. These pictures may represent completely different burns of varying mechanisms on different body parts, on different skin tones and taken on different quality cameras to name a few variables. The fact that the ML algorithms are potentially capable of reasonably analysing the images despite the various acquisition protocols, input features and ground truths is a testament to its versatility (15).

Limitations of Observed Studies

Some of the key limitations observed across the investigative studies include the lack of a publicly available burn-related image database for burn depth and segmentation analysis. Such a dataset could serve as validated ground truth, which could be used for example for any segmentation task. Furthermore, a strictly defined and published protocol, ideally including a defined chronological and normed image capturing of the wound with respective standardized annotations, would be beneficial for the data collection procedure definition. We anticipate that the images of infected burn wounds may cater for other research streams in the field of Burns ML.

A further limitation lies within the lack of adequate performed validation procedures, including but not limited to the proper separation of training and test patients in image segmentation tasks. Finally, a cost-analysis of the effectiveness of ML in the management and treatment of burns would be beneficial. This can be quantified through cost savings in early discharge, timing of interventions versus the cost of equipment, maintenance, and personnel (training and implementation).

QUADAS-2 is a versatile tool for judging the quality of diagnostic test papers. We used it to inform our sensitivity analysis, but its implementation was not straightforward since it is not tailored to AI diagnostic studies. Since their inception, the CONSORT-AI (Consolidated Standards of Reporting Trials) and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials) have published preferred reporting guidelines for trials using AI (55,56). We suggest, based on an integration of their recommendations with the ones made in this review, adapting the QUADAS-2 signalling questions to better appraise studies. Examples of these along the ML pathway are shown below in Figure 6.

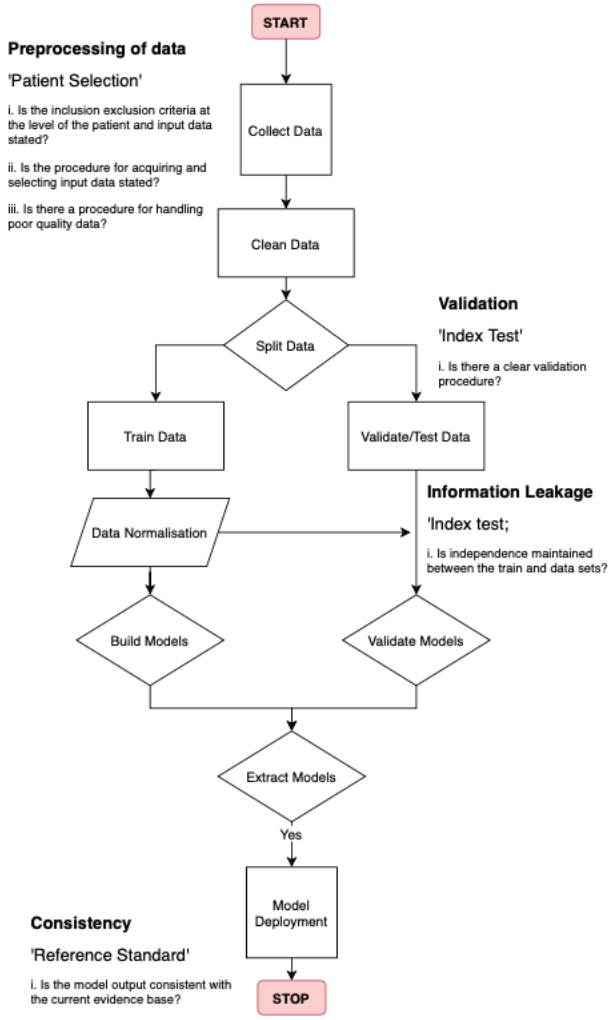


Figure 6: Potential ML Flowchart and signalling questions.

Study Limitations

There are several limitations to this study. Despite trying to categorise studies based on their outcomes significant intra-group heterogeneity exists between the small number of studies considered. Overall, this indicates that any results from the meta-analysis should be treated with caution. Ideally each study sub-type (burn depth, burn mortality and burn segmentation) should exhibit similar or even the same acquisition protocols, thresholds and ground truth definition. For burn depth, the gold standard ground truth definition would originate from the histopathological analysis, however, this is not often possible due to the risk of introducing infection, causing scarring (particularly for shallower burns) and delay wound healing. A burns community based concensus approach would be required on ground truths i.e. for burn depth and segmentation so that ML algorithms can be accurately scrutinised against a robust dataset.

The reverse calculation of values in the confusion matrices for some studies may have introduced bias through assumptions made as suggested by the sensitivity analysis for burn depth. Additionally, in order to compare studies, which depicted three burn depths with those employing two burn depths, certain labels may have altered. For example, 2nd degree burns encompass a vast array of burn depths, some of which may need a graft or require

longer to healing time. Labelling these injuries with being in the more severe category may not be a true reflection of the burns datasets analysed.

Conclusion

The application of AI in the treatment and management of burns patients, as a series of point of care diagnostic adjuncts, is promising, particularly in lower resource settings or outside the expertise of burns centres.

Whilst AI is a potentially valuable tool a full evaluation of its current utility and potential is limited by significant variations in research methodology and reporting. It is only by addressing these limitations, that clinicians will be able to drive forward the use of AI and incorporate it into the burns clinical repertoire.

Declarations

Contribution Statement

BGT, NM, GG and AK came up with the original study design. BGT and LM reviewed all studies that fulfilled the inclusion criteria. BGT and AK extracted the relevant information from each study before performing relevant analysis. BGT and KW assessed the risk of bias for each study. All authors were involved in revising the manuscript for publication.

Funding and Conflicts of Interest

No funding was received for this article and none of the authors have any conflicts of interest to declare.

Ethical Approval

No ethical approval was required for this study.

References

1. Grosu-Bularda A, Andrei MC, Mladin AD, Ionescu Sanda M, Dringa MM, Lunca DC, et al. Periorbital lesions in severely burned patients. Romanian J Ophthalmol. 2019;63(1):38–55.

2. Kanevsky J, Corban J, Gaster R, Kanevsky A, Lin S, Gilardino M. Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation. Plast Reconstr Surg. 2016;137(5):890e–7e.

3. Auger SD, Jacobs BM, Dobson R, Marshall CR, Noyce AJ. Big data, machine learning and artificial intelligence: a neurologist's guide. Pract Neurol. 2021 Feb 1;21(1):4–11.

4. Seo H, Khuzani MB, Vasudevan V, Huang C, Ren H, Xiao R, et al. Machine Learning Techniques for Biomedical Image Segmentation: An Overview of Technical Aspects and Introduction to State-of-Art Applications. Med Phys. 2020 Jun;47(5):e148–67.

5. Suvarna M, Kumar S, Niranjan U. Classification Methods of Skin Burn Images. Int J Comput Sci Inf Technol. 2013 Feb 28;5:109–18.

6. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn KW. Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. Burns. 2015 Aug;41(5):925–34.

7. Tran NK, Albahra S, Pham TN, Holmes JH, Greenhalgh D, Palmieri TL, et al. Novel application of an automated-machine learning development tool for predicting burn sepsis: proof of concept. Sci Rep. 2020 Jul 23;10(1):12354.

8. Rashidi HH, Sen S, Palmieri TL, Blackmon T, Wajda J, Tran NK. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. Sci Rep. 2020 Jan 14;10(1):205.

9. Liu NT, Salinas J. Machine learning in burn care and research: A systematic review of the literature. Burns. 2015;41(8):1636–41.

10. Dubey K, Srivastava V, Dalal K. In vivo automated quantification of thermally damaged human tissue using polarization sensitive optical coherence tomography. Comput Med Imaging Graph. 2018;64:22–8.

11. Acha B, Serrano C, Acha JI, Roa LM. CAD tool for burn diagnosis. Inf Process Med Imaging Proc Conf. 2003 Jul;18:294–305.

12. Kuan P, Chua S, Safawi E, Wang H, Tiong W. A Comparative Study of the Classification of Skin Burn Depth in Human. J Telecommun Electron Comput Eng. 2018;9(2–10):15–23.

13. Wang Y, Ke Z, He Z, Chen X, Zhang Y, Xie P, et al. Real-time burn depth assessment using artificial networks: a large-scale, multicentre study. Burns [Internet]. 2020 Jul 25 [cited 2020 Nov 23]; Available from: http://www.sciencedirect.com/science/article/pii/S0305417920304691

14. Serrano C, Acha B, Gómez-Cía T, Acha JI, Roa LM. A computer assisted diagnosis tool for the classification of burns by depth of injury. Burns. 2005 May 1;31(3):275–81.

15. Abubakar A, Ugail H, Smith KM, Bukar AM, Elmahmudi A. Burns Depth Assessment Using Deep Learning Features. J Med Biol Eng. 2020;40(6):923–33.

16. Yadav D, Sharma A, Singh M, Goyal A. Feature Extraction Based Machine Learning for Human Burn Diagnosis From Burn Images. IEEE J Transl Eng Health Med. 2019 Jul 18;PP:1–1.

17. Patil BM, Toshniwal D, Joshi RC. Predicting burn patient survivability using decision tree in WEKA environment. In 2009. p. 1353–6.

18. Abubakar A, Ugail H. Discrimination of Human Skin Burns Using Machine Learning. Adv Intell Syst Comput. 2019;997:641–7.

19. Cirillo MD, Mirdell R, Sjöberg F, Pham TD. Time-Independent Prediction of Burn Depth Using Deep Convolutional Neural Networks. J Burn Care Res. 2019;40(6):857–63.

20. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. BMC Med Res Methodol. 2019 Apr 18;19(1):81.

21. Patel A, Cooper N, Freeman S, Sutton A. Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. Res Synth Methods. 2021;12(1):34–44.

22. Deeks JJ, Higgins J, Altman D. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 62 (updated February 2021) Cochrane, 2021 [Internet]. [cited 2021 Mar 31]. Available from: www.training.cochrane.org/handbook.

23. Bossuyt P, Davenport C, Deeks J, Hyde C, Mariska L, Scholten R. Chapter 11 Interpreting results and drawing conclusions. In: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [Internet]. Available from: https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/DTA%20Handbook%20Chapter%2011%20201312.pdf

24. Yamamura S, Kawada K, Takehira R, Nishizawa K, Katayama S, Hirano M, et al. Prediction of aminoglycoside response against methicillin-resistant Staphylococcus aureus infection in burn patients by artificial neural network modeling. Biomed Pharmacother Biomedecine Pharmacother. 2008 Jan;62(1):53–8.

25. Yeong EK, Hsiao TC, Chiang HK, Lin CW. Prediction of burn healing time using artificial neural networks and reflectance spectrometer. Burns. 2005 Jun 1;31(4):415–20.

26. Chauhan J, Goyal P. BPBSAM: Body part-specific burn severity assessment model. Burns. 2020;46(6):1407–23.

27. Chauhan J, Goyal P. Convolution neural network for effective burn region segmentation of color images. Burns. 2020;

28. Khan FA, Butt AUR, Asif M, Ahmad W, Nawaz M, Jamjoom M, et al. Computer-aided diagnosis for burnt skin images using deep convolutional neural network. Multimed Tools Appl. 2020 Dec;79(45–46):34545–68.

29. Wantanajittikul K, Auephanwiriyakul S, Theera-Umpon N, Koanantakool T. Automatic segmentation and degree identification in burn color images. In: The 4th 2011 Biomedical Engineering International Conference. 2012. p. 169–73.

30. Jiao C, Su K, Xie W, Ye Z. Burn image segmentation based on Mask Regions with Convolutional Neural Network deep learning framework: more accurate and more convenient. Burns Trauma [Internet]. 2019 Feb 28 [cited 2020 Nov 24];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6394103/

31. Abubakar A, Ugail H, Bukar AM, Aminu AA, Musa A. Transfer learning based histopathologic image classification for burns recognition. In 2019.

32. Abubakar A, Ugail H, Bukar AM. Noninvasive assessment and classification of human skin burns using images of Caucasian and African patients. J Electron Imaging. 2020;29(4).

33. Abubakar A, Ugail H, Bukar AM. Can machine learning be used to discriminate between burns and pressure ulcer? Adv Intell Syst Comput. 2020;1038:870–80.

34. Acha B, Serrano C, Fondon I, Gomez-Cia T. Burn Depth Analysis Using Multidimensional Scaling Applied to Psychophysical Experiment Data. IEEE Trans Med Imaging. 2013 Jun;32(6):1111–20.

35. Badea M, Vertan C, Florea C, Florea L, Bădoiu S. Automatic burn area identification in color images. In: 2016 International Conference on Communications (COMM). 2016. p. 65–8.

36. Pinero BA, Serrano C, Acha JI, Roa LM. Segmentation and classification of burn images by color and texture information. J Biomed Opt. 2005 May;10(3):034014.

37. Estahbanati HK, Bouduhi N. Role of artificial neural networks in prediction of survival of burn patients—a new approach. Burns. 2002 Sep 1;28(6):579–86.

38. Cobb AN, Daungjaiboon W, Brownlee SA, Baldea AJ, Sanford AP, Mosier MM, et al. Seeing the forest beyond the trees: Predicting survival in burn patients with machine learning. Am J Surg. 2018 Mar;215(3):411–6.

39. Frye KE, Izenberg SD, Williams MD, Luterman A. Simulated biologic intelligence used to predict length of stay and survival of burns. J Burn Care Rehabil. 1996;17(6):540–6.

40. Jiménez F, Sánchez G, Juárez JM. Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. Artif Intell Med. 2014 Mar 1;60(3):197–219.

41. Tran NK, Sen S, Palmieri TL, Lima K, Falwell S, Wajda J, et al. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. Burns. 2019 Sep 1;45(6):1350–8.

42. Yamamura S, Kawada K, Takehira R, Nishizawa K, Katayama S, Hirano M, et al. Artificial neural network modeling to predict the plasma concentration of aminoglycosides in burn patients. Biomed Pharmacother. 2004 May 1;58(4):239–44.

43. Berchialla P, Gangemi EN, Foltran F, Haxhiaj A, Buja A, Lazzarato F, et al. Predicting severity of pathological scarring due to burn injuries: a clinical decision making tool using Bayesian networks. Int Wound J. 2014 Jun;11(3):246–52.

44. Liu NT, Rizzo JA, Shields BA, Serio-Melvin ML, Christy RJ, Salinas J. Predicting the Ability of Wounds to Heal Given Any Burn Size and Fluid Volume: An Analytical Approach. J Burn Care Res. 2018 Aug 17;39(5):661–9.

45. Brown RF, Rice P, Bennett NJ. The use of laser Doppler imaging as an aid in clinical management decision making in the treatment of vesicant burns. Burns J Int Soc Burn Inj. 1998 Dec;24(8):692–8.

46. Heimbach DM, Afromowitz MA, Engrav LH, Marvin JA, Perry B. Burn depth estimation--man or machine. J Trauma. 1984 May;24(5):373–8.

47. Hoeksema H, Van de Sijpe K, Tondu T, Hamdi M, Van Landuyt K, Blondeel P, et al. Accuracy of early burn depth assessment by laser Doppler imaging on different days post burn. Burns J Int Soc Burn Inj. 2009 Feb;35(1):36–45.

48. Monstrey SM, Hoeksema H, Baker RD, Jeng J, Spence RS, Wilson D, et al. Burn wound healing time assessed by laser Doppler imaging. Part 2: validation of a dedicated colour code for image interpretation. Burns J Int Soc Burn Inj. 2011 Mar;37(2):249–56.

49. Overview | moorLDI2-BI: a laser doppler blood flow imager for burn wound assessment | Guidance | NICE [Internet]. NICE; [cited 2021 Mar 29]. Available from: https://www.nice.org.uk/guidance/mtg2

50. Ward J, Phillips G, Radotra I, Smailes S, Dziewulski P, Zhang J, et al. Frailty: an independent predictor of burns mortality following in-patient admission. Burns J Int Soc Burn Inj. 2018 Dec;44(8):1895–902.

51. Varadarajan AV, Bavishi P, Ruamviboonsuk P, Chotcomwongse P, Venugopalan S, Narayanaswamy A, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat Commun. 2020 Jan 8;11(1):130.

52. Medical Image Analysis Using Deep Learning: A Systematic Literature Review | SpringerLink [Internet]. [cited 2021 May 23]. Available from: https://link.springer.com/chapter/10.1007%2F978-981-13-8300-7_8

53. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec 1;42:60–88.

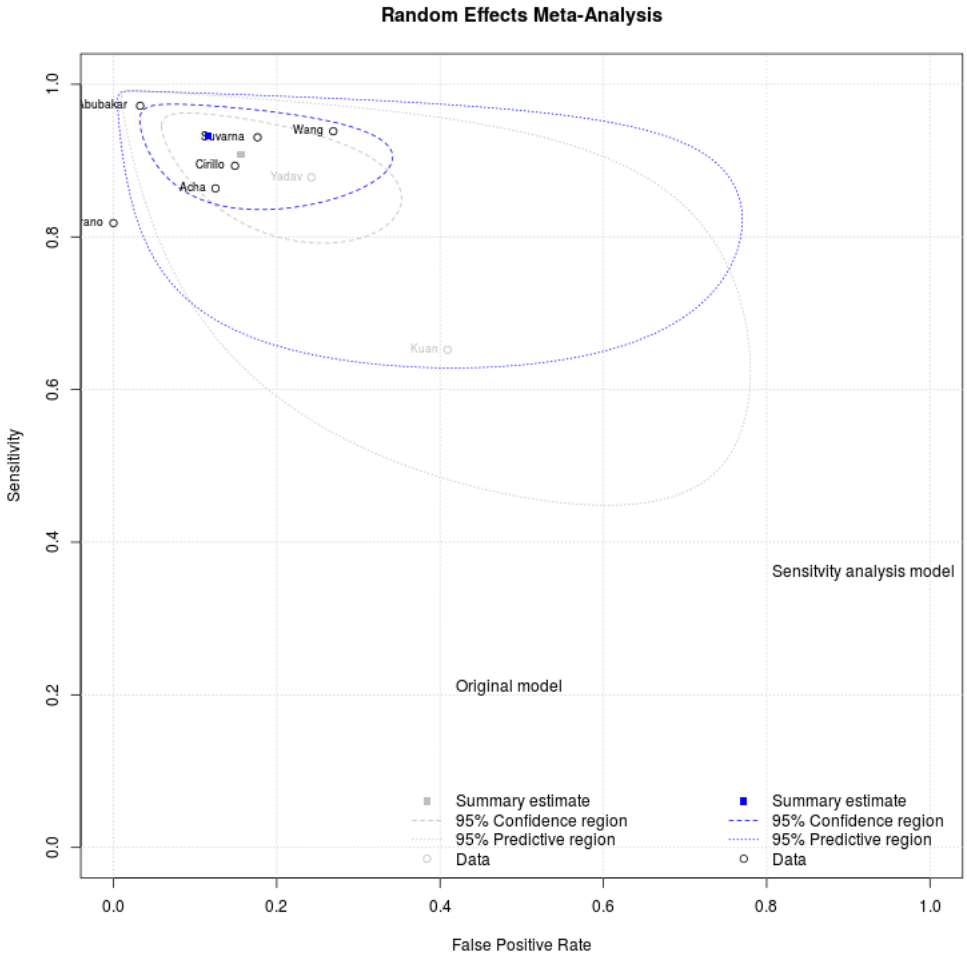
54. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data. 2019 Dec 17;6(1):113.

55. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ. 2020 Sep 9;370:m3210.

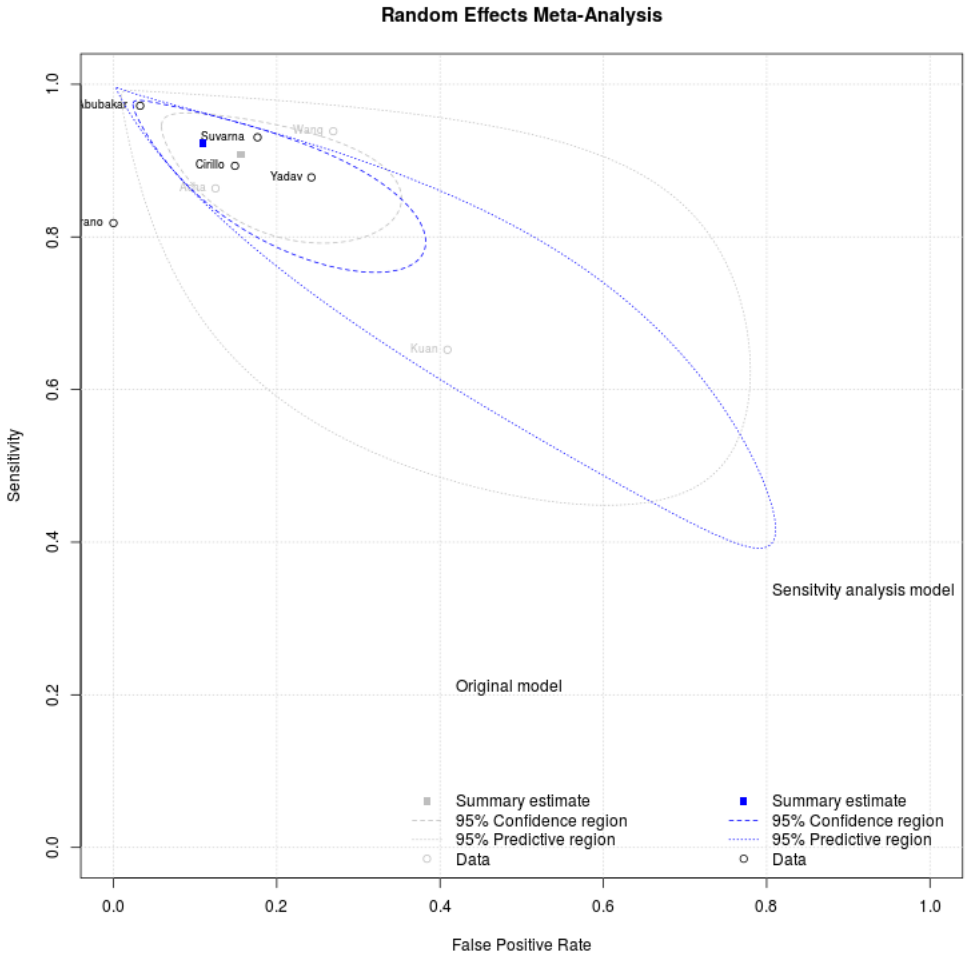
56. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension | Nature Medicine [Internet]. [cited 2021 May 20]. Available from: https://www.nature.com/articles/s41591-020-1034-x

Supplementary Material

Sensitivity Analyses



Supplementary Figure 1: Sensitivity analysis HSROC curve, 95% confidence and predictive region superimposed on the original HSROC curve after removal of the Kuan et al and Yadav et al. burn depth studies which showed the greatest risk of bias.



Supplementary Figure 2: Sensitivity analysis HSROC curve, 95% confidence and predictive region superimposed on the original HSROC curve after removal of the Kuan et al., Wang et al. and Acha et al. burn depth studies whose confusion matrices were reverse engineered from available data.