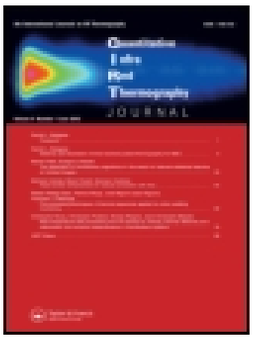

Quantitative InfraRed Thermography Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tqrt20

Medical infrared thermal image based fatty liver classification using machine and deep learning

Ahmet Özdil & Bülent Yilmaz

To cite this article: Ahmet Özdil & Bülent Yilmaz (2023): Medical infrared thermal image based fatty liver classification using machine and deep learning, Quantitative InfraRed Thermography Journal, DOI: 10.1080/17686733.2022.2158678

To link to this article: https://doi.org/10.1080/17686733.2022.2158678

QUANTITATIVE INFRARED THERMOGRAPHY JOURNAL
https://doi.org/10.1080/17686733.2022.2158678

Taylor & Francis
Taylor & Francis Group

Medical infrared thermal image based fatty liver classification using machine and deep learning

Ahmet Özdil [a,b] and Bülent Yilmaz [b,c]

[a]Faculty of Engineering and Architecture, Computer Engineering Department, Kırşehir Ahi Evran University, Kırşehir, Türkiye; [b]College of Engineering and Architecture, Electrical Engineering Department, Gulf University for Science and Technology, Mishref, Kuwait; [c]School of Engineering, Electrical and Electronics Engineering Department, Abdullah Gül University, Kayseri, Türkiye

ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) causes accumulation of excess fat in the liver affecting people who drink little to no alcohol. Non-alcoholic steatohepatitis (NASH) is an aggressive form of fatty liver disease (inflammation in the liver), may progress to cirrhosis and liver failure. Liver function tests, ultrasound (US) and magnetic resonance imaging (MRI) are used to help diagnose and monitor liver disease or damage. In this study, the feasibility of medical infrared thermal imaging (MITI) in automatic detection of NAFLD was investigated, and 167 MITI images (44 positive) from 32 patients (7 positive) were evaluated using image processing and classification methods. Convolutional neural network (CNN) architectures and texture analysis methods were used in the feature selection phase. After feature selection and binary classification, the highest values from different setups for recall, f-score, specificity, accuracy, and area-under-curve (AUC) were 1.00, 1.00, 0.83, 1.0, 0.94, and 0.92, respectively. The highest values were achieved by CNN based methods on different datasets, however, texture analysis method performed lower. Here, it is shown that some of the CNN architectures have high potential on extracting features from thermal images. Finally, machine and deep learning approaches can be combined in detecting NAFLD using infrared thermal images.

ARTICLE HISTORY
Received 18 July 2022
Accepted 10 December 2022

KEYWORDS
Non-alcoholic fatty liver disease; medical infrared thermal imaging; machine learning; convolutional neural networks

Introduction

Non-alcoholic fatty liver disease (NAFLD) is a common disorder where there is accumulation of excess fat in the liver affecting people who drink little to no alcohol [1]. NAFLD includes different stages of metabolic liver disease from simple steatosis to steatohepatitis, advanced fibrosis, cirrhosis, and later hepatocellular carcinoma (HCC) [1]. Due to the substantial change in lifestyle and diet of many people in different countries NAFLD is gaining spread as 15% in 2005 to 25% in 2010 [2,3]. In the USA approximately 100 million people are estimated to have NAFLD, and especially among children it is the most common form of liver disease.

CONTACT Ahmet Özdil ahmet.ozdil@ahievran.edu.tr Faculty of Engineering and Architecture, Computer Engineering Department, Kırşehir Ahi Evran University, Bağbaşı Campus, Kırşehir, Türkiye

Liver function tests are used to help diagnose and monitor liver disease or damage. In addition, medical imaging and rarely liver biopsy is performed during the evaluation process. Imaging approaches like ultrasonography (US), transient elastography, computed tomography (CT), and magnetic resonance imaging (MRI) are currently being employed in the clinical routine, however, liver biopsy is still perceived as the golden standard to assess progression in the disease [4]. After the biopsy, patients may encounter complications like bleeding, infection and puncture in biliary tract along with temporary pain. Early diagnosis of NAFLD can be ensured by regular follow-up of a patient, however, applying an invasive method like biopsy many times may cause other health risks.

To accompany blood tests and medical imaging as popular non-invasive diagnosis and monitoring approaches, a new pre-diagnosis/scanning modality based on medical infrared thermal imaging (MITI) may be developed as a non-invasive, harmless, user-friendly and cost-effective method. However, MITI is a relatively new and not a standard/golden method compared to other medical imaging approaches.

Since MITI captures the infrared energy that is naturally emitted from body [5], it should be studied extensively to discover new application areas. MITI investigates the thermal differences throughout the body surface [5] and minor temperature differences on the body surface may be significant indicators for various health problems beneath that area, all small temperature changes should be evaluated during thermal examination of patients.

Vast amount of MITI studies focus on breast screening [6–11]. There are other anomalies such as the inflammatory arthritis, osteoarthritis, soft tissue rheumatism, tennis elbow, fibromyalgia, complex regional pain syndrome, peripheral circulation, fever, burn diagnosis [12–15] that MITI has proposed as the diagnosis/monitoring tool. Moreover, in recent years the field of quantitative analysis of thermal images emerged and has the potential to yield objective interpretation rather than physician-dependent (subjective) analysis [16]. To be more specific, in [10], different sets of features were generated from breast infrared (IR) thermal images and their classification performances were evaluated using support vector machines (SVM). Neonatal monitoring is also another field that MITI is employed as in [17]. The study in [18] includes the performance comparison of MITI on screening cardiovascular disease with traditional techniques and found promising results. In [19], MITI was used for identifying diabetic foot ulcers. The researchers calculated the mean temperature of regions of interest by considering thermal asymmetry and applied different machine learning (ML) algorithms on the calculated features. In [20], IR images of both iris were used for diagnosing type-2 diabetes. A decision support system was offered for predicting haemodynamic shock in paediatric intensive care unit and promising performance was achieved in [21]. In [22], burn wounded patients were monitored and a treatment decision system was developed to offer the patient most effective treatment method. A review of studies performed on different diseases can be found in [23].

Furthermore, attaching an infrared camera into a monitoring system is another beneficial use of MITI in medicine. In [24], a smartphone camera – based photoplethysmography (PPG) and a low-cost thermal camera were combined in an instant stress detection model. A facial expression recognition system, using thermal images and SVM, was proposed in [25].

Disorders, related to blood circulation system and skin surface can be monitored with MITI due to their effect on skin surface temperature. A tumour in breast disturbs the blood

circulation within the breast. Imbalanced blood circulation beneath the skin affects the thermal distribution on the surface. The structure of breast made it easy to apply MITI techniques. Therefore, vast amount of the MITI studies were done on the breast tumour screening. Fat accumulated within the liver also disturbs the blood circulation of liver, and therefore, MITI can be effective for screening liver over the skin surface. There is a recent study in [26] which was conducted for classifying the fatty liver disease among mice. They used MITI for capturing the skin temperature map over the liver and achieved 100% binary classification success.

Non-destructive methods used in medicine should be encouraged to reduce the harm on patients. It is well known that many screening and diagnosing techniques are invasive and cause some kind of harm such as frequently repeated mammograms, CT scans or X-ray imaging causing a risk of tumour development. Therefore, we believe that the use of MITI should be expanded since it has no negative physiological effects on patients. Recently many studies were conducted on employing MITI in medicine [9,14,19,21,22,25,26]. These kinds of efforts will increase the acceptance of MITI as a pre-diagnosis and treatment follow-up method.

Especially in the context of NAFLD, in [26] infrared thermal images were collected during experiments on mice and a fatty liver classification method was developed using machine learning techniques. The researchers extracted 9 texture features from every image and clustered them into 2 clusters as normal and steatosis. This study was inspirational for us to devise this current research.

There is a thermal symmetry between right and left sides of the body. When there is a health problem, this symmetry is broken. If there is a problem in liver like NAFLD etc. which causes hyperthermia, it also causes hyperthermic reflection on the skin area which covers the liver. There are two subjects below, a healthy subject on the left side and an NAFLD subject on the right side. The hyperthermic reflection of liver is pointed out with a red circle in the NAFLD subject thermogram in Figure 1.

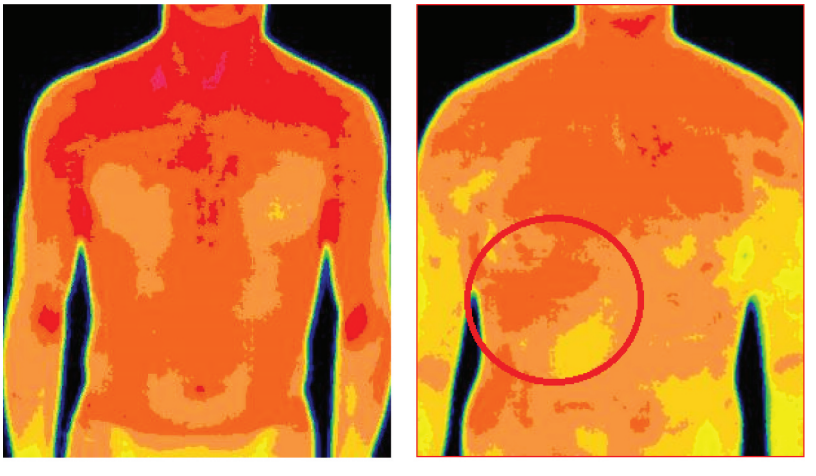

Figure 1. Thermal images of healthy (left side) and NAFLD diagnosed (right side) subjects.

In this study, our aim was to investigate the feasibility of MITI in automatic detection of NAFLD. Thermal images were evaluated using deep learning- and machine learning-based feature extraction, feature selection and classification methods. Performance comparison of convolutional neural networks (CNN)- and texture analysis-based feature extraction was the basis for the subsequent analysis aiming at differentiating healthy individuals and the ones with NAFLD.

Materials and methods

Acquisition of thermal images

Thermal images were captured during routine examination of patients of Dr. YILMAZ clinic in Kayseri/Türkiye. This clinic focuses on traditional and complementary medicine, and the responsible physician has been using medical thermal images as the initial step of pre-diagnosis of certain diseases (NAFLD is one of them) for the last 10 years. The diagnosis for NAFLD was confirmed with conventional methods such as enzyme tests and US imaging performed in public or private hospitals. The collaboration between Abdullah Gül University and Dr. YILMAZ clinic has been in place since 2019. The research was conducted with the permission of Erciyes University Ethical Council of Clinical Studies, Kayseri, Türkiye (permission number: 2019/524). Informed consent was obtained from all individuals (or their parents if their ages are below 18).

IRIS-XP Infrared Thermography Device (Medicore Co. Ltd., Seoul, South Korea) is used in the clinic for acquiring the images. Camera resolution is 384 × 288 pixels (pixel size 17um), focal length is 8.1 mm, FOV is 49° x 36°, spatial resolution (IFOV) is 1.1mRad, spectral range is 8 ~ 14 microns, detector type is uncooled micro-bolometer, and frame rate is 9 or 30 Hz. In the clinic, the thermal images are acquired in a controlled environment with stable thermal conditions. The imaging room is about 10 m^2, a thick curtain for the window is used to keep the sunlight away, and temperature (about 22 degrees Celsius) and humidity are kept stable using an air conditioner. Patients acclimate 15 minutes and stay only with their underwear approximately 1.5 metres away from the camera and the camera is positioned according to his/her height by the physician. First, the patient stands still facing the camera, then turns 90 and 180 degrees around his/her own axis, and then raises the arms. Thus, during the image acquisition procedure thermal images from the upper, lower, frontal and posterior parts of the body were captured. The room where thermal images are acquired can be seen in Figure 2.

Images from adult women were not included because the setup allows only to capture images while the patient is standing, and adult women breasts covered up the region corresponding to the liver. However, a young female with age 11 was kept for further analysis. Only the images with upper front view were selected, and each region-of-interest (ROI) corresponding to the anatomical location of the liver was extracted manually. Figure 3 demonstrates one sample from the database with the upper front view of grey-level and coloured images and the corresponding liver regions selected manually. During evaluations the images from 32 subjects were included with 7 positives, and their characteristics are presented in Table 1 (all variables were distributed non-normally).

In this study, the dataset of images consisted of 123 negative (no NAFLD) and 44 positive (with NAFLD) infrared thermal images taken from 32 subjects (25 negative and 7

Figure 2. Thermal image acquisition room.

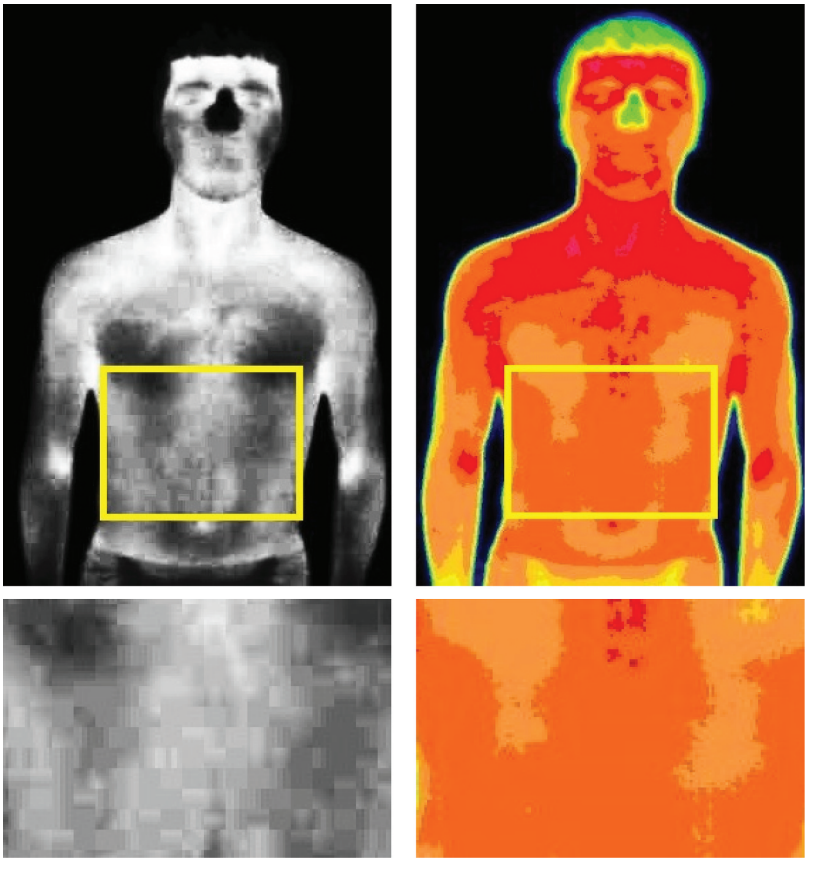

Figure 3. Upper front view of thermal grey-level and colored images and the corresponding liver regions selected manually.

Table 1. Subject characteristics.

Subject Characteristics	NAFLD (n=7) Median [IQR]	Control (n=25) Median [IQR]	p-value
Age (years)	54.00 [22.50]	24.00 [24.00]	0.0358
Weight (kg)	88.00 [20.50]	70.00 [32.00]	0.0114
Height (cm)	175.00 [6.00]	167.00 [19.00]	0.0129
Body Mass Index (kg/m2)	30.45 [7.705]	23.66 [10.42]	0.0449

IQR: Inter Quartile Range

positive subjects). Only one subject was female (age of 11). The ages of the subjects ranged from 9 to 68, and 9 subjects were below 18 years old. The mean ± standard deviation of the body mass index (BMI) of patients who are below and above 18 years old are 19.37 ± 4.71 and 28.51 ± 5.62 respectively. The average BMIs of positive and negative groups are 30.58 and 24.82 respectively. The statistical comparison between two groups was performed using the Mann-Whitney U test, and the corresponding p-values are presented in the table.

The control group patients were selected among other patients who were seeking treatment for diseases other than NAFLD. The disorders the control group suffered from are the arm shoulder pain, hypothyroidism, scoliosis, depression, cerebrovascular disease, epilepsy, mental motor retardation, coronary artery disease, migraine, kyphosis, frequent infection, motor neuron disease, spina bifida, chest pain, backache, headache, dyspnoea, sleep apnoea, restless legs syndrome, medulloblastoma (operated), rheumatoid arthritis, leg varicose veins, rhabdomyosarcoma, autism, psoriasis, and spasticity. The NAFLD group members also suffer from nodules in the lung, urinary tract infection, lumbar HNP, angioedema, urticaria, psoriasis, knee pain, ulcerative colitis, and tinnitus.

Pre-processing of thermal images

Thermal imaging captures the temperature values of the scene as pixel values using infrared sensor arrays. These pixel (temperature) values are used to colour the image either as RGB or grey level. In this study, only the grey level images (raw images) were used. In the beginning foreground extraction was applied to images using Otsu's approach [27] for eliminating the background (uninformative temperature values) and letting us to focus on the trunk images of the patients. Several segmentation approaches like active contours region growing and local first-order statistics were investigated, and Otsu's thresholding method was easy and more successful in distinguishing foreground and background than the other two approaches. Figure 4 describes the anatomical landmarks for liver ROI in which the upper border was taken as the nipple line, the lower border as the superior iliac crest, and lateral borders as the mid/anterior axillary lines.

Histogram equalisation (HE) method [29] was applied to the outcomes of the Otsu's segmentation approach to boost slight differences in the temperature over the body. Thus, HE method was modified to equalise the histograms only within the body, i.e. Otsu's segmentation produced a mask and HE was applied on the pixels corresponding to that mask [30]. When HE process was over, the anatomical location corresponding to the liver on

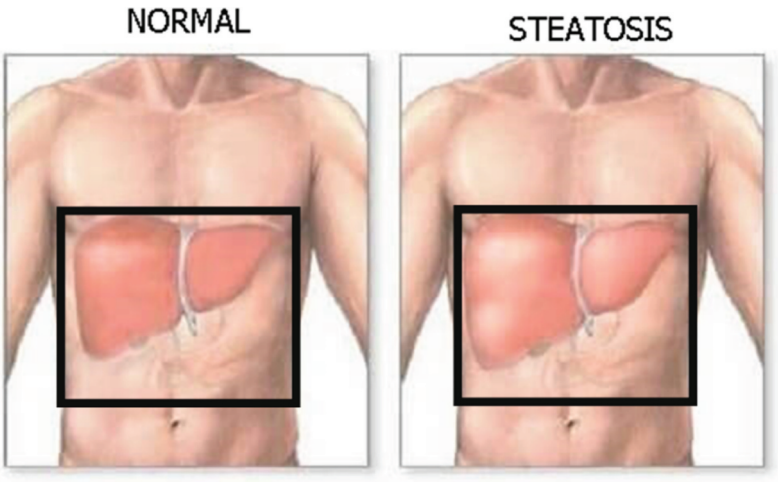

Figure 4. The anatomical landmarks for liver ROI [28].

each image was extracted from the image manually. The pre-processing steps can be seen in Figure 5, where the original image, the mask of foreground/body using Otsu's approach, HE applied on masked image only within the body, and manually extracted liver region can be seen on columns a, b, c, and d respectively. In the last step of the pre-processing phase, selection of ROI corresponding to the liver of the patient was done manually since postures of the patients were not uniform.

Image datasets

Thermal images were used to generate three different datasets for evaluating different methods. Dataset 1 contained all thermal images (123 negative and 44 positive). Dataset 2 was formed by eliminating 50% of negative subjects' images by keeping the subject number constant (61 negative and 44 positive) for evaluating if the imbalance between negative and positive subjects has negative affect on classification performance. Dataset 3 was formed to improve the performance by reducing the ROI to the upper triangle of the thermal images and no images were removed while generating this dataset (123 negative and 44 positive). Our datasets were summarised in Table 2.

In Figure 6, the formation of an upper triangular-shaped ROI is explained on a positive sample. The left column shows the original image, and the right column depicts the ROI with upper triangular shape. Upper triangular ROI was considered to be well suited to the shape of the liver and more uninformative pixels (temperature pixels that do not correspond to the liver area) would be eliminated.

Feature extraction and selection of thermal images

Two different feature extraction approaches were used in this study; texture analysis and CNN architectures. Texture analysis approach was similar to the one used on mice [26] such as mean temperature, variance, skewness, kurtosis, and entropy features were extracted from thermal images, and contrast, correlation, energy, and homogeneity were calculated from the

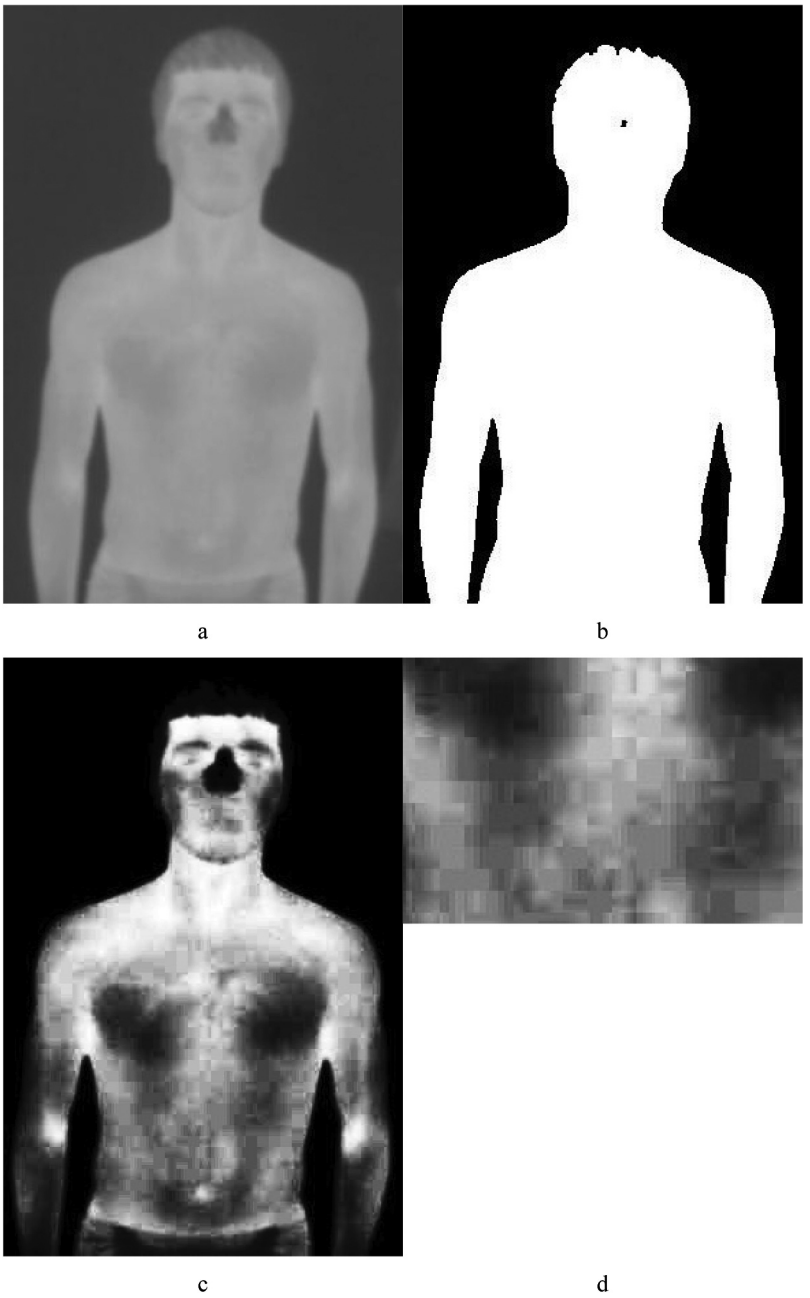

Figure 5. Pre-processing of images a: original image, b: foreground mask, c: HE applied only within the body, and d: ROI corresponding to the liver of the patient.

grey level co-occurrence matrix (GLCM). GLCM contains the number of co-occurrence of 2 grey level values within a neighbourhood throughout the image. In this study, 1-neighbour GLCM was calculated both over the whole image and only within the upper triangular ROI of the image to compare if there would be an improvement on classification performance. By

Table 2. Details of dataset used in the experiments.

	Positive Subjects	Negative Subjects	Dataset specialty
Dataset 1	7 (44 images)	25 (123 images)	Original
Dataset 2	7 (44 images)	25 (61 images)	Down sampled
Dataset 3	7 (44 images)	25 (123 images)	Triangle ROI

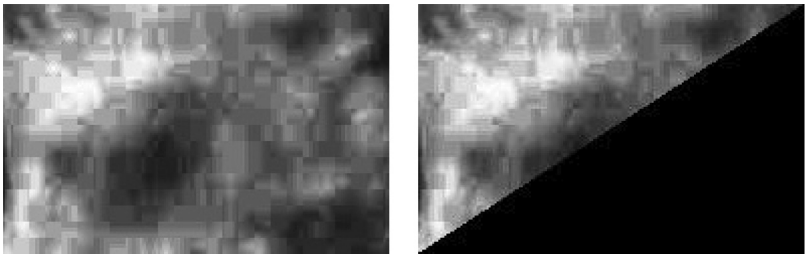

Figure 6. Upper triangular ROI used in dataset 2.

default, the scaling of the grey levels of the images was performed into 1-to-8 integer interval including boundaries while calculating GLCM. In our study, GLCM calculation was modified and was added to performance comparison to calculate GLCM in 1-to-256 integer interval. This modification was applied not to lose the details on thermal images while scaling into 1-to-8 integer interval. In medicine, even 0.5-degree change may be an important symptom of a disease.

Convolutional Neural Networks (CNNs) have proven themselves as a powerful tool for image segmentation, feature extraction and binary classification. In this study, different pre-trained CNN architectures [31] were investigated for feature extraction from the thermal images. The number of channels in the grey level images was increased from 1 to 3 by copying pixel values to new channels for evaluating in CNN phase. The CNNs used here are listed as DarkNet-19, DarkNet-53, ResNet-18, ResNet-50, ResNet-101, DenseNet-201, VGG-16, GoogleNet, SqueezeNet, AlexNet, and Inception-ResNet-v2. The details of these architectures can be found in this review article [32]. These CNN architectures produced 1000 to 4096 features from different datasets.

Two different feature selection methods were evaluated for feature selection step: principal component analysis (PCA) [33] and t-distributed stochastic neighbour embedding (t-SNE) [34] approaches. PCA is computationally low-cost approach and finds the most varying dimensions in the dataset [33]. Therefore, PCA was selected for feature selection in this study, and t-SNE was eliminated due to performance failure on time consumption and binary classification accuracy during preliminaries.

Binary classification

These selected feature sets were used to generate train and test files to be used in The Waikato Environment for Knowledge Analysis (Weka) [35] tool. The binary classification phase was implemented in a patient-wise manner. Leave-one-subject-out cross-validation method was used (n-fold cross-validation, in this case 32-fold), and 81,078 and 54,052

Table 3. Abbreviations of classification methods.

Method Name	Abbreviation
Bayesian Logistic Regression	BLR
Classification Via Clustering	CVC
Ensembles of Balanced Nested Dichotomies	END
Filtered Classifier	FC
Multi Class Classifier	MCC
Multinomial Logistic Regression	Logistic
Primal Estimated Sub-GrAdient SOlver for SVM	SPegasos
Repeated Incremental Pruning	JRip
Sequential Minimal Optimization	SMO
Voted Perceptron	VP
Voting Feature Intervals	VFI

train-test set combinations were generated during CNN and texture analysis steps respectively for evaluating classification methods implemented in Weka, which is a well-known tool for evaluating classification methods and contains numerous methods to investigate on. In this study, up to 52 different classification methods were employed to determine the one with the best performance. The best performing methods and their abbreviations are listed in Table 3. For this reason, about 5 million different binary classification attempts have been made by developing an interface using MATLAB and through this interface Weka tool was used for automatic classification attempts. The experimental setup was prepared for $(81078 + 54052) \times 52 = 6021184$ classification attempts, but some methods were eliminated during evaluations due to low classification accuracy and high computational cost.

The binary classification phase was implemented with 32-folds since there are 32 subjects in this study. Multiple images were included for 25 subjects and only one image was included for the rest of the subjects. During training and test phases the images from 31 subjects were included in the training phase and the images from one subject were used for test phase in a fold, and this process was repeated for 32 folds. Majority voting method was employed while the test subject was labelled at the end of the training and test phase. After a fold is completed the test results of the test subject's images were inspected and if majority of the images were labelled positive the subject was labelled as positive otherwise the s/he was labelled as negative.

All classification results were analysed using CCI (correctly classified instances) values of 32 folds of each setup, which were computed during the analysis using Weka tool. These 32-fold CCI values were processed with majority voting method (if a CCI value of test set of a fold was greater than 50% then that subject got 1 vote else 0 vote), TP (true positive) and TN (true negative) values were computed by summation of these vote values. Then, performance metrics were computed using TP and TN values. Classification performances were compared on recall, F-score, specificity, accuracy, and area-under-curve (AUC) of the results after majority voting phase.

Results and discussion

In this study, thermal imaging was investigated if it is useful for monitoring problematic liver area. However, in Figure 7 some of the images used in this study were presented in RGB form. It can be seen in the positive images that the area over the liver produces

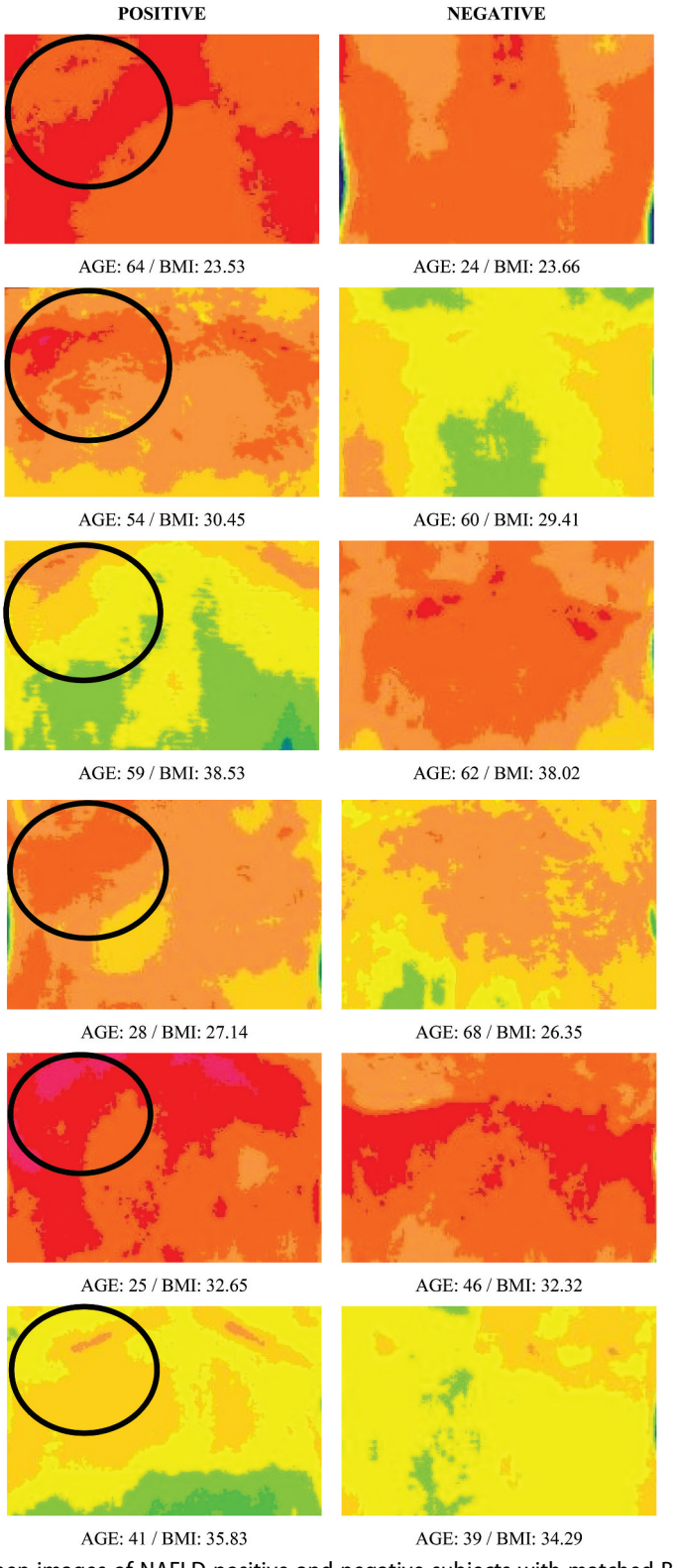

Figure 7. Abdomen images of NAFLD positive and negative subjects with matched BMIs.

a hyperthermic field, and it can be seen when RGB image's colour map is adjusted properly by an expert physician. Due to thermal conduction, there is a heat transfer between the organs in the body. Body loses heat by different ways. The lost heat does not only belong to skin and the vessels close to skin surface, but also belong to the organs close to the skin like the liver, intestines, sinuses, etc. Therefore, thermal sign of the liver can be seen over the skin surface that covers the liver itself.

The preliminary visual observations supported our hypothesis that hyperthermic region over the liver can be interpreted as a sign of an abnormality in that organ. In Figure 7, the thermograms depicted as RGB images demonstrate a hyperthermic area originating from the liver. For fair comparison we included abdominal images of NAFLD positive and negative subjects with matched BMIs. On the right top corner of the abdomen (ROIs) the hyperthermic regions are highlighted using a black circle.

The feature exploration step was conducted to find out the most informative features. The importance of this step can be seen in Figure 8 where classification performance results from different feature sets (number of selected features ranged from 1 to 148 in dataset 1) were displayed. If all cases were not explored, the system would converge to local optimum and miss the global optimum.

Table 4 summarises the best classification performances of the evaluated binary classification attempts on different setups. Each line of the table represents one classification phase. PCA method performed better than the t-SNE method did nearly in all classification evaluations, therefore, the second feature selection method was not included in Table 4 for simplicity.

The experiments were repeated for the listed setups in Table 4 for coloured images. The classification performance on RGB images decreased, because of the fact that the image colouring is an area of expertise. An experienced physician should adjust the colour map to uncover the information hidden in the thermogram. In Figure 7, the coloured images were generated manually by an experienced physician and hence the hyperthermic skin area covering the liver can be seen clearly. However, we used same colour map

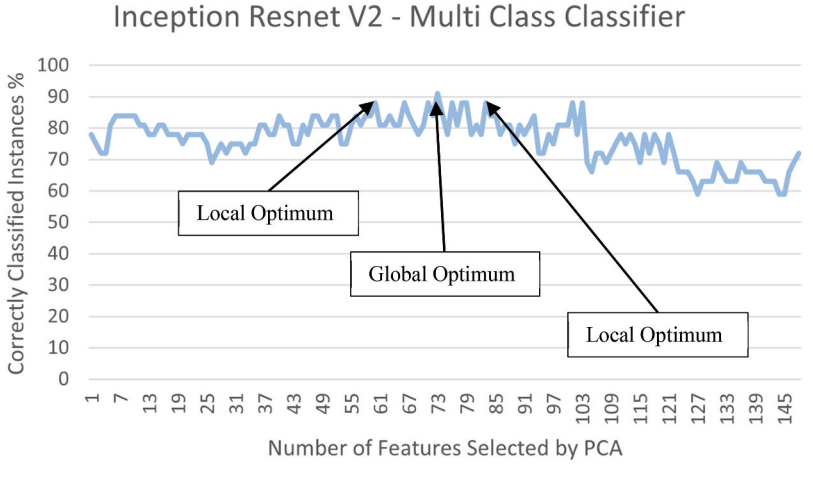

Figure 8. Classification accuracy of MCC classifier in which the features were extracted using inception resnet V2 extractor and selected by PCA (1 to 148 features).

Table 4. Summary of the classification attempts those with the highest performances.

Setup	Dataset	Feature Extractor	# of PCA Features	Classifier	Recall	Specificity	F-Score	Accuracy	AUC
4	1	Inception-ResNet-v2	5	FC	0.71	1	0.83	0.94	0.86
14	3	AlexNet	94	SMO	0.71	0.96	0.77	0.91	0.84
3	1	Inception-ResNet-v2	79	MCC	1	0.84	0.78	0.88	0.92
6	1	Texture-256 level	80	CVC	0.43	1	0.6	0.88	0.71
9	2	Inception-ResNet-v2	82	MCC	0.86	0.71	0.84	0.84	0.85
20	3	Texture-256 level	134	JRip	0.43	0.96	0.55	0.84	0.69
8	2	Inception-ResNet-v2	73	MCC	1	0.7	0.76	0.81	0.88
11	2	DenseNet-201	78	JRip	1	0.67	0.72	0.78	0.86
15	3	AlexNet	112	Logistic	1	0.72	0.67	0.78	0.86
16	3	DenseNet-201	65	VP	0.86	0.76	0.63	0.78	0.81
10	2	AlexNet	42	BLR	0.71	0.59	0.8	0.78	0.76
5	1	Texture-8 level	122	END	0.71	0.8	0.59	0.78	0.76
17	3	Inception-ResNet-v2	76	VP	0.71	0.8	0.59	0.78	0.76
12	2	Texture-8 level	55	SMO	0.71	0.55	0.76	0.75	0.74
1	1	AlexNet	6	VFI	1	0.6	0.58	0.69	0.8
2	1	DenseNet-201	11	VP	1	0.56	0.56	0.66	0.78
13	2	Texture-256 level	6	SPegasos	0.86	0.5	0.56	0.63	0.71
7	1	Texture-256 level	3	SPegasos	0.86	0.52	0.48	0.59	0.69
18	3	Inception-ResNet-v2	1	CVC	1	0.44	0.5	0.56	0.72
19	3	Texture-8 level	11	VP	0.71	0.52	0.41	0.56	0.62
21	3	Texture-256 level	11	VP	0.71	0.52	0.41	0.56	0.62

for every image during automatic pre-processing phase and this caused the loss of information. We did not include the RGB image results in the manuscript in order not to make the manuscript long.

The highest recall, F-score, specificity, accuracy, and AUC values were achieved by different setups, which means there was no overall winning setup across all performance metrics. Setup 4 has the highest results on specificity, F-score, and accuracy with 1.00, 0.83, 0.94, respectively. However, the highest AUC score was achieved by setup 3 and followed by setup 9 with 0.92, and 0.86, respectively. Both texture- and CNN-based feature extraction methods performed well, however, CNN-based methods outperformed texture-based methods. The 256-grey level GLCM calculation produced better accuracy over traditional 8-grey level GLCM computation on dataset 1 (original dataset), but the AUC was decreased due to recall failure. The modified GLCM computation method (computation of the GLCM only within upper triangular ROI) was applied with both 8- and 256-grey levels on dataset 2 and depicted in Table 4. Texture-256 level feature extraction method with JRip classifier outperformed texture-8 level except for the recall metric.

Dataset 1 is the original set of images on which no modifications were made. This dataset is imbalanced between positive and negative instances. Therefore, dataset 2 was built to help us investigate the effect of the imbalanced nature of the original dataset. The highest classification accuracy results of dataset 1 and dataset 2 were tested with Kruskal-Wallis test [36]. The result of Kruskal-Wallis test show that the results of dataset 1 and dataset 2 were not normally distributed, therefore the non-parametric t-test was applied, and no significant difference was reported between the results of dataset 1 and dataset 2 (p-value >0.05). It is revealed in our study that the imbalance between the classes did not have significant or any negative effect on the classification accuracy. Dataset 3 was built to evaluate if the pixels that do not belong to liver on the abdomen have negative effect on the classification performance, and a second ROI extraction approach was applied by selecting an upper triangular ROI for evaluation. The highest classification accuracy results

of dataset 1 and dataset 3 were tested with non-parametric t-test since they were not normally distributed. The results revealed that there was no significant difference between the results (p-value >0.05). Majority voting method was applied to classification accuracy result of each subject as such, if accuracy is greater than 50%, it is assumed that the subject was classified correctly. Thus, reducing the ROI to upper triangle improved the success for the feature extraction methods AlexNet and DenseNet-201, however the performance decreased for feature extraction methods Inception-ResNet-v2 and the two texture methods.

In the literature many recent studies include machine learning methods for automatic processing of thermal images [23]. However, deep learning methodologies have not been widely explored for MITI. Deep learning CNN architectures proved themselves as a powerful image processing tool. In this study, the performances of the texture analysis and 11 different CNN architectures were cooperatively evaluated, and remarkable results were reported by using different combinations of deep learning or machine learning techniques.

The main research question of this study was to find out if deep learning methods as the feature extraction approach would perform well on thermal images. It was proven that pre-trained CNN architectures were powerful tools in MITI since these architectures achieved the highest classification performance over texture-based methods. The pre-trained networks were trained using more than a million images to recognise 1000 different objects which improves their learning capacity deeply. It was revealed in [26] that steatosis and inflammation affects infrared radiation release of the liver, therefore they analysed texture features from the thermal images. In our study it was proven that deep learning architectures are better on differentiating the texture properties of thermal images in the NAFLD context.

Another question investigated in this study was the GLCM computation on the thermal images. The default behaviour in GLCM computation is to scale the grey levels from 0-to-255 into 1-to-8 integer interval. Since the pixel values are scaled into a smaller range and the new values are rounded to integer values, many details may be lost during this scaling. Therefore, in this study GLCM computation was modified to use the images' original grey level interval. Another default behaviour of GLCM computation is to cover all pixels in the image. However, in this study the region of the liver was the focus of our research and including non-liver area pixels into the calculations will decrease the steatosis information coming from liver area. Therefore, computation of the GLCM of the pixels that belong to liver area was proposed and applied to thermal images, which directs to another finding, combining 256-grey level and including only the pixels that belong to the liver area for GLCM calculation increased performance for AlexNet and DenseNet-201 methods as described in Results section.

Another aim in this study was to explore the performance of all feature sets of a feature selection method. In Figure 8, it can be clearly seen that not exploring all possible feature combinations may converge to a local optimum. However, generally the first few principal components were included in binary classification phase which contains about 95% of variation information of the data, but later component may include meaningful information from the data as depicted in Table 4.

PCA has proven itself as a powerful feature selector throughout all evaluations against t-SNE method. The t-SNE method also produced good results that were close to PCA, but not better. In addition, t-SNE method's computation cost was high.

The research conducted in this study was performed in collaboration with a traditional and complementary medicine clinic in Kayseri, Türkiye. Due to working conditions of the clinic the thermal images are not standardised. Another limitation in the image collection phase was the non-standardised postures of patients who were mostly elderly or very young. Adult woman subjects were not included in this study, because the breasts cover up the liver and infrared radiation becomes insensible. For future studies, women subjects may be asked to raise their breasts to uncover the liver area, so that they would be included in such research tuned for women. Another improvement may include the capturing of the lateral view of subjects as in Figure 9, it can be seen that the areas covering the liver have different shapes both in healthy and positive subjects.

We note that multivariate analysis should be conducted on a larger cohort to address the effect of co-morbidities on thermal images of subjects. Existing co-morbidities may affect the thermal image if they disturb the circulatory system. In addition, an inflammation may increase or decrease the tissue blood supply which may directly affect the thermal image.

The dataset will be improved by obtaining new thermal images. New methods for capturing images are being offered to clinic to improve the standardisation among thermal images.

We believe that the findings in this research lead to new directions on MITI. New feature selection methods should be evaluated for different deep learning and machine learning methods. The main result of this study is to use different combinations of methods to find out the highest performing combination. It is important to find the best combination of methods that show the best harmony with each other. This should guide researchers to develop more effective methodologies.

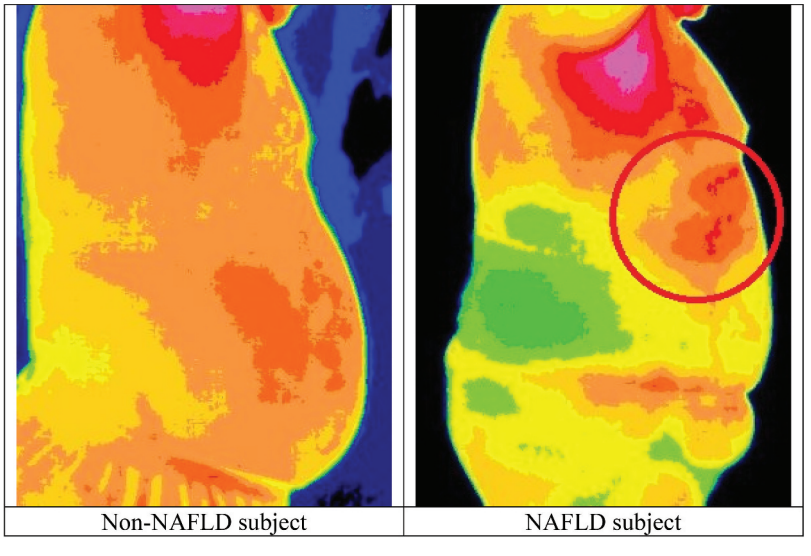

Figure 9. Non-NAFLD vs. NAFLD subjects from lateral view.

Conclusions

The main contribution of this study is to show the potential of evaluating human abdomen thermal images using machine and deep learning techniques for NAFLD classification. To our best knowledge, no studies have been conducted in this context on human subjects so far. Different contributions were also made during evaluations of experiments. It was revealed that along with grey level co-occurrence matrix (GLCM) based texture analysis, which was previously shown to be feasible in NAFLD on mice, CNN architectures are better to be used for feature extraction in this context. Another contribution was to modify GLCM calculation methodology to find pixel co-occurrences only within the upper triangular region-of-interest (ROI) similar to the shape of the liver instead of a rectangular ROI which is more common and includes more unrelated information about the temperature distribution over the liver and nearby regions, thus some methods' performances increase (AlexNet and DenseNet-201).

The results in this study reveal different directions for further studies. Improvements and modifications can be made on new pre-processing techniques on thermal images, which will improve revealing deeper features in the thermal images using CNN architectures. For example, different supervised or unsupervised feature selection methods can be evaluated. Another improvement can be done in training the CNN architecture with one channel grey level images, i.e. original raw images. In addition, in this study only one offset was calculated for GLCM computation, for future studies different offsets should be evaluated. Another research direction can be to develop a new CNN architecture by combining the evaluated CNN architectures that performed well during this research endeavour.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Ahmet Özdil http://orcid.org/0000-0002-6651-1968
Bülent Yilmaz http://orcid.org/0000-0003-2954-1217

References

[1] Sevastianos VA, Hadziyannis SJ. Nonalcoholic fatty liver disease: from clinical recognition to treatment. Expert Rev Gastroenterol Hepatol. 2008;2(1):59–79.
[2] Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2018;15(1):11–20. DOI:10.1038/nrgastro.2017.109
[3] Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet. Oct 2020;396(10258):1204–1222.
[4] Schwenzer NF, Springer F, Schraml C, et al. Non-invasive assessment and quantification of liver steatosis by ultrasound, computed tomography and magnetic resonance. J Hepatol. 2009;51(3):433–445.
[5] Ring EFJ, Ammer K. Infrared thermal imaging in medicine. Physiol Meas. 2012 Mar;33(3):R33–46.

[6] Gogoi UR, Bhowmik MK, Ghosh AK, et al. Discriminative feature selection for breast abnormality detection and accurate classification of thermograms. *2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*, Shillong, India; 2017 Apr. p. 39–44.
[7] Ghayoumi Zadeh H, Haddadnia J, Rahmani Seryasat O, et al. Segmenting breast cancerous regions in thermal images using fuzzy active contours. EXCLI J. 2016;15:532–550.
[8] Gogoi UR, Majumdar G, Bhowmik MK, et al. Breast abnormality detection through statistical feature analysis using infrared thermograms. *2015 International Symposium on Advanced Computing and Communication (ISACC)*, Silchar, India; 2015 Sep. p. 258–265.
[9] Morales-Cervantes A, Kolosovas-Machuca ES, Guevara E, et al. Evaluation of breast cancer by infrared thermography. Res Comput Sci. 2020;149(5):137–149.
[10] Gogoi UR, Majumdar G, Bhowmik MK, et al. Evaluating the efficiency of infrared breast thermography for early breast cancer risk prediction in asymptomatic population. Infrared Phys Technol. 2019;99(September 2018):201–211.
[11] Ng EY-K, Fok SC, Peh YC, et al. Computerized detection of breast cancer with artificial intelligence and thermograms. J Med Eng Technol. 2002 Jan;26(4):152–157.
[12] Ring E. The historical development of thermometry and thermal imaging in medicine. J Med Eng Technol. 2006;30(4):192–198.
[13] Anbar M. Clinical thermal imaging today. IEEE Eng Med Biol Mag. 1998;17(4):25–33.
[14] Hooshmand H, Hashmi M, Phillips EM. Infrared thermal imaging as a tool in pain management - an 11 year study, part II: clinical applications. Thermol Int. 2001;11(3):1–13.
[15] Nowakowski A, Kaczmarek M. Active dynamic thermography - problems of implementation in medical diagnostics. Quant Infrared Thermogr J. 2011;8(1):89–106.
[16] Balageas DL. A brief history of QIRT concept, structures and community. Quant Infrared Thermogr J. 2016;13(1):109–125.
[17] Ervural S, Ceylan M. Thermogram classification using deep siamese network for neonatal disease detection with limited data. Quant Infrared Thermogr J. 2021 Dec;19(5):1–19.
[18] Thiruvengadam J, Mariamichael A. Model-based computer-aided method for diagnosis of cardiovascular disease using IR thermogram. Biomed Res. 2019;30(1). DOI:10.35841/biomedicalresearch.30-19-004
[19] Vardasca R, Magalhaes C, Seixas A, et al. Diabetic foot monitoring using dynamic thermography and AI classifiers. *Proceedings of QIRT Asia 2019*, Tokyo, Japan; 2019. p. 0–4.
[20] Samant P, Agarwal R. Machine learning techniques for medical diagnosis of diabetes using iris images. Comput Methods Programs Biomed. 2018;157:121–128.
[21] Nagori A, Dhingra LS, Bhatnagar A, et al. Predicting hemodynamic shock from thermal images using machine learning. Sci Rep. 2019;9(1):1–9.
[22] Martínez-Jiménez MA, Ramirez-GarciaLuna JL, Kolosovas-Machuca ES, et al. Development and validation of an algorithm to predict the treatment modality of burn wounds using thermographic scans: prospective cohort study. PLoS ONE. 2018;13(11):1–16.
[23] Vardasca R, Magalhaes C, Mendes J. Biomedical applications of infrared thermal imaging: current state of machine learning classification. Proceedings. 2019;27(1):46.
[24] Cho Y, Julier SJ, Bianchi-Berthouze N. Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. JMIR Ment Heal. 2019;6(4):e10140.
[25] Wang S, Pan B, Chen H, et al. Thermal augmented expression recognition. IEEE Trans Cybern. 2018;48(7):2203–2214.
[26] Brzezinski RY, Levin-Kotler L, Rabin N, et al. Automated thermal imaging for the detection of fatty liver disease. Sci Rep. 2020;10(1):1–24. DOI:10.1038/s41598-020-72433-5
[27] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9(1):62–66.
[28] Karaciğer İltihabı: Semptomlar ve Doğru Beslenme. https://sagligabiradim.com/karaciger-iltihabi-semptomlar-uygun-diyet/ .
[29] Gonzalez RC, Woods RE. Digital image processing(second edition). 2nd ed. New Jersey, USA: Prentice-Hall; 2002.

[30] Özdil A, Yılmaz B. Automatic body part and pose detection in medical infrared thermal images. Quant Infrared Thermogr J. 2021 Jun;19(4):223–238.
[31] Matlab, MATLAB. 2021.
[32] Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8(1):53. DOI:10.1186/s40537-021-00444-8
[33] Karamizadeh S, Abdullah SM, Manaf AA, et al. An overview of principal component analysis. J Signal Inf Process. 2013;04(03):173–175.
[34] García-Alonso CR, Pérez-Naranjo LM, Fernández-Caballero JC. Multiobjective evolutionary algorithms to identify highly autocorrelated areas: the case of spatial distribution in financially compromised farms. Ann Oper Res. 2014;219(1):187–202.
[35] Hall M, Frank E, Holmes G, et al. The weka data mining software. ACM SIGKDD Explor Newsl. 2009 Nov;11(1):10–18.
[36] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952 Dec;47(260):583–621.