Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-019-02101-y

ORIGINAL ARTICLE

A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods

Hamza O. Ilhan[1] · I. Onur Sigirci[1] · Gorkem Serbes[2,3] · Nizamettin Aydin[1]

Received: 24 May 2019 / Accepted: 16 December 2019

Abstract

Sperm morphology, as an indicator of fertility, is a critical tool in semen analysis. In this study, a smartphone-based hybrid system that fully automates the sperm morphological analysis is introduced with the aim of eliminating unwanted human factors. Proposed hybrid system consists of two progressive steps: automatic segmentation of possible sperm shapes and classification of normal/ab-normal sperms. In the segmentation step, clustering techniques with/without group sparsity approach were tested to extract region of interests from the images. Subsequently, a novel publicly available morphological sperm image data set, whose labels were identified by experts as non-sperm, normal and abnormal sperm, was created as the ground truths of classification step. In the classification step, conventional and ensemble machine learning methods were applied to domain-specific features that were extracted by using wavelet transform and descriptors. Additionally, as an alternative to conventional features, three deep neural network architectures, which can extract high-level features from raw images after using statistical learning, were employed to increase the proposed method's performance. The results show that, for the conventional features, the highest classification accuracies were achieved as 80.5% and 83.8% by using the wavelet- and descriptor-based features that were fed to the Support Vector Machines respectively. On the other hand, the Mobile-Net, which is a very convenient network for smartphones, achieved 87% accuracy. In the light of obtained results, it is seen that a fully automatic hybrid system, which uses the group sparsity to enhance segmentation performance and the Mobile-Net to obtain high-level robust features, can be an effective mobile solution for the sperm morphology analysis problem.

Keywords Infertility · Sperm morphology · Sperm abnormality classification · Convolution neural networks · Group sparsity · Discrete wavelet transform · Support vector machines

1 Introduction

Infertility is diagnosed as the inability of having children after 1 year of regular sexual intercourse. Fifteen to 20% of all couples in the world are engaged in a form of infertility problems according to the World Health Organization (WHO) reports [25]. The reason of those problems are mainly categorised as follows: (i) male, (ii) female, (iii) unexplained and (iv) couple-based reasons [26].

Semen specimen analysis, also known as spermiogram, is the most popular test in the diagnosis of the infertility to observe the male factor-based problems. It is performed in two steps: (i) the given semen specimen is evaluated for the physical appearances such as viscosity, color and smell. (ii) an expert measures the sample characteristics in terms of the sperm morphology, concentration and motility parameters by employing either the computerised or manual evaluation techniques.

The sperm morphology analysis focuses on the dimensional evaluation for sperm head, mid-piece and tail. The abnormalities have a direct effect on the infertility. WHO reports the reference values to be used as a guide in the human sperm assessments. In the first edition published in 1980, the reference ratio for normal sperm density over all sperm concentration was defined at 80% [13] but later it was

✉ Hamza O. Ilhan
hoilhan@yildiz.edu.tr

[1] Department of Computer Engineering, Yildiz Technical University (YTU), 34220 Istanbul, Turkey

[2] Faculty of Engineering and Information Technologies, School of Biomedical Engineering, University of Sydney, Sydney, Australia

[3] Department of Biomedical Engineering, Yildiz Technical University (YTU), 34220, Istanbul, Turkey

Published online: 06 March 2020

updated as 4% in the 2010 edition [54]. These reports also indicate that the quality of the sperm morphology is rapidly decreasing.

In the preparation of the semen specimen prior to morphological analysis, sample is stained with chemicals to increase the visibility of sperm shapes. Then, sample is observed by an expert in the manual evaluation step, also known as visual assessment technique, which is cheaper and more practical compared with computer-based systems. Therefore, many laboratories still perform the tests with visual assessment technique. However, the results are strongly affected by the observer experience. This is also known as the observer variability problem in which the results of the analysis can vary by the analyser [20]. Alternatively, tests can be performed on computer-based systems that isolate the human factors in the analysis [73]. Such computer-based systems are known as Computer-Aided Sperm Analysis (CASA) systems. Comparing with the visual analysis, CASA system is more reliable, consistent and objective. However, the cost of installation and maintenance of the system is more expensive. Additionally, considering the morphological analysis, CASA systems require considerable improvements in many aspects such as denoising, border enhancing, occlusion detection and segmentation over full slice of images [6, 11]. Due to these reasons, CASA approaches are less preferred in the analysis. Therefore, low-cost/high-accurate automatic analysing systems are still needed for the sperm morphology analysis.

Computer-based analysis systems are mainly composed of two consecutive steps. Sperm detection in given semen specimen is a segmentation process and constitutes the first step of analysis. Subsequently, each segmented region of interest (ROI) patch needs to be assigned to a class (non-sperm, normal or abnormal sperm) by classification and machine learning techniques. There are different studies in the literature under these two main headings.

In the sperm detection step, Olalla et al. employed a combination of Otsu thresholding and wavelet transform [31]. Next, they used SVM in the classification step of the detected parts. They created a binary classification model that focuses on only acrosome defects labelled as normal and abnormal. Alegre et al. implemented the same approach in the sperm detection step but they formed a different classifier model using the Haralick and Counter features [2, 3] to evaluate the same data set as normal and abnormal. Khachane et al. presented a fuzzy rule-based classification technique for more detailed sperm abnormality classification instead of only acrosome-based abnormality [44]. Each sperm was individually segmented into detailed head, mid-piece and tail sub-segments. Then, the spatial features obtained by blob analysis (major and minor axes, regional area etc.) of each segmented sperm pieces were utilised in the fuzzy logical expressions based classification step [44]. Chang et al. proposed two-stage feature extraction schema to use in classifier design [22]. They implemented the clustering idea in the histogram-based color space analysis. In another paper, they aimed to introduce a gold standard procedure for the sperm morphological analysis [21]. They manually cropped and rotated each sperm patches from the stained microscope images. Then, in order to obtain patch features, they employed shape-based descriptors. Naive Bayes (NB), *k* Nearest Neighbor (*k*-NN), Decision Tree (DT) and Support Vector Machines (SVM) models were trained by the features for the classification process. The performance was measured as 58% for this data set. Similarly, Shaker et al. also manually cropped the sperm patches from the stained images and rotated each sperm patch to one specific direction [64]. Then, they used a dictionary-learning schema to classify the four abnormalities in the SVM model. They also tested their technique on the Chang et al. data set [21]. They reported that they achieved 92.2% accuracy for their data sets (HuSHem) and also increased the accuracy performance to 62% for Chang et al. data set (SCIAN-Morpho).

Normal sperm head can be modelled as an elliptical shape. In the detection step of the head, Nafisi et al. proposed a measurement criteria for the size and the existence of tail. They formed an elliptical mask to segment headpiece [52]. In the diagnosis, there are different acrosome-based sub-abnormalities. Nafisi et al. failed to classify the sub-abnormalities in their study due to inefficient head border extraction. Gonzales-Castro et al. modified conventional Otsu thresholding by using watershed segmentation for a detailed sperm head analysis [33]. This approach provided more visible sperm borders to define the head with more certain regions. However, microscope optical reflections, sperm occlusions and halos affect the detection step. In order to overcome this problem, Bijar et al. utilised a Bayesian classifier. They classified the pixels into three classes as background, nucleus and acrosome [14, 15]. Adaptive Mixture Models and Markov random field were used to obtain the class-based probability functions. Occlusion and halo effects were successfully eliminated in this approach, but the existence of non-sperm particles was still a problem because of having similar shape with the normal sperm. They focused on only the size information of extracted parts to classify them into sperm/non-sperm. Artifacts and debris having similar shapes to sperm structure affected the class probabilities.

In this study, we implemented a group-sparse signal denoising method as an initial process to maintain more enhanced sperm shape, and to suppress small particles. In this respect, the modified overlapping group shrinkage (MOGS) technique, which does not isolate the signal components having large magnitude (likely to form groups similar to sperms), was utilised [23]. Next, we extracted the ROI from the background by using Fuzzy C-Means clustering technique with MOGS. We also tested K-Means clustering technique with/without MOGS. As a preliminary study, this approach was performed on comparatively small data set and the results were presented in [40]. Following this, a range of spatial-based features, such as the size and eccentricity of segmented patches, was applied to raw microscopic images to automatically discriminate possible sperm forms (sperm vs non-sperm). Up to this point, all the applied processes were made automatically and the segmentation performance of the proposed method is presented in a similar way as described in [40]. As a prerequisite of the second step (classification step) of proposed cascaded system, the segmented patches ,which were identified by using spatial features with the aim of finding possible sperm/non-sperm patches, were manually labelled by experts into three classes as non-sperm, normal and abnormal sperms. This manual labelling process had created a data set which was later used to measure classification performance of proposed method and it is not a part of automatic hybrid system. After attaining the three-class sperm morphology data set, the classical machine learning workflow, which can be broken down into three steps as feature extraction, model learning and model evaluation process, was applied to measure the performance of conventional features. The classical dyadic discrete wavelet transform (DWT) [50], dual tree wavelet transform (DTWT) and dual tree complex wavelet transform (DTCWT) [61] were employed as the time-scale distribution-based feature extractors. On the other hand, Speed Up Robust Features (SURF), KAZE and Maximally Stable Extremal Regions (MSER) as interest point descriptors were applied to the created data set with the aim of capturing the sperm morphology. Following to this feature extraction step, the SVM, DT, k-NN and ensemble DTs models with different kernels and parameters were employed as the learners in the traditional classification step. In our preliminary studies, we have already tested several combination of this conventional techniques on a small size of our created dataset [39, 40]. DTCWT at 7 decomposition level with RBF kernel SVM achieved 82.33% accuracy in classification of the small data set [39]. In another preliminary study using the same data set, SURF descriptor resulted in 83.39% accuracy rate with SVM. In this study, we tested these and several more conventional techniques over our new extended dataset. Additionally, as an end-to-end method without feature extraction steps, three Convolutional Neural Networks (CNN) named as the InceptionV3 [71], VGG19 [67] and MobileNet [36] were employed as deep neural network (DNN) classifiers in this study. In these CNNs, the raw pixels of segmented patches are fed into the first level of a DNN and the outputs of that layer can be interpreted as representing the presence of different low-level features in the possible sperm images, such as lines and edges of sperms. At subsequent layers of these CNNs, low-level features are then combined into a measure of the likely presence of higher level features such as basic shapes (head, tail of a sperm or a noise component), which are further combined into sets of shapes (full sperm body). And finally, using all this information, the DNNs can provide a probability that these high-level features comprise a particular object (for example a sperm or non-sperm patch including multiple sperms or multiple optical microscope effects respectively) [69]. The obtained classification performances of all the conventional feature extraction plus traditional machine learning approaches and end-to-end DNN-based approaches were presented and the workflow having the highest accuracy accompanied by the lowest computational cost was chosen as the recommended hybrid model.

This paper is organised as follows; Section 2 covers data set and methods used in the study. The clustering results with the MOGS pre-processing step and the conventional, ensemble and deep learning classification results with descriptor and wavelet transform-based feature extraction scenarios are given in Section 3. Discussions and comparison with other studies are given in Section 4. Finally, conclusions and the future works that can be used in subsequent studies are presented in Section 5.

2 Materials and methods

In this study, we used a smartphone-based data acquisition approach to obtain the microscope ocular images. Following the smartphone-based data recording phase, the images are transmitted to a server in which the sperm segmentation and classification phases are applied. After testing all data processing steps, we have evaluated the optimum framework for the sperm morphology analysis problem in terms of the performance metrics and computational complexity. The flowcharts of the all employed techniques and the optimum hybrid framework are given in Figs. 1 and 2 respectively.

2.1 Data acquisition step

The semen specimen images are captured by a smartphone-based data acquisition approach from the ocular part of microscope. The data acquisition layout which was introduced and tested for the motile sperm detection process in [38] is illustrated in Fig. 3. The data acquisition software in the mobile phone includes a software-based image stabiliser that is employed to obtain more accurate images.

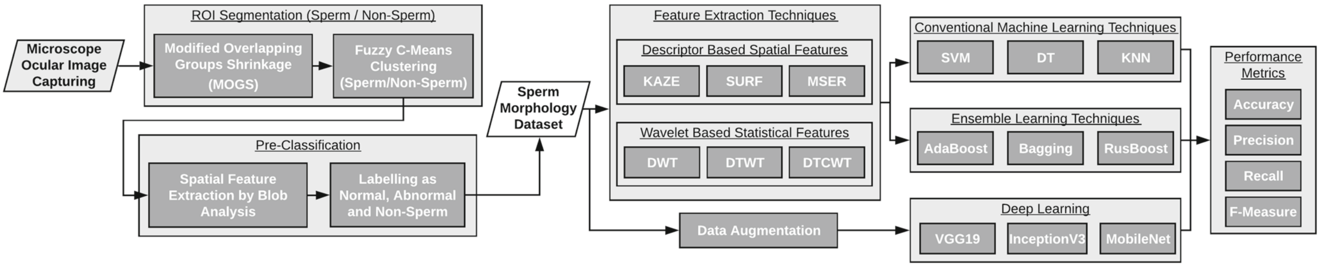

Microscope Ocular Image Capturing
ROI Segmentation (Sperm / Non-Sperm)
Modified Overlapping Groups Shrinkage (MOGS)
Fuzzy C-Means Clustering (Sperm/Non-Sperm)
Pre-Classification
Spatial Feature Extraction by Blob Analysis
Labelling as Normal, Abnormal and Non-Sperm
Sperm Morphology Dataset
Feature Extraction Techniques
Descriptor Based Spatial Features
KAZE
SURF
MSER
Wavelet Based Statistical Features
DWT
DTWT
DTCWT
Data Augmentation
Conventional Machine Learning Techniques
SVM
DT
KNN
Ensemble Learning Techniques
AdaBoost
Bagging
RusBoost
Deep Learning
VGG19
InceptionV3
MobileNet
Performance Metrics
Accuracy
Precision
Recall
F-Measure

Fig. 1 The flowchart of the tested methods in the sperm morphology analysis

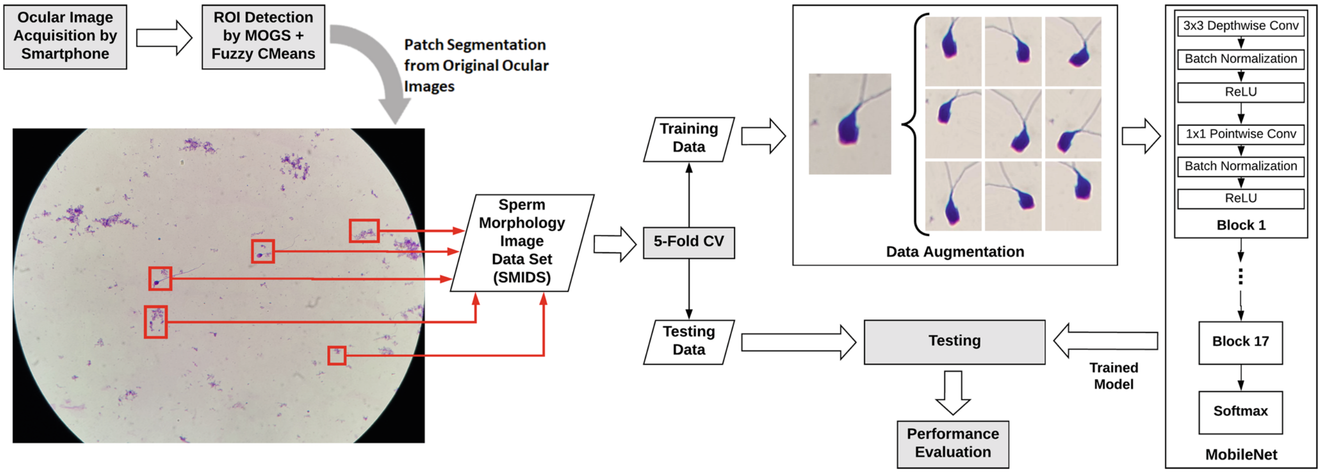
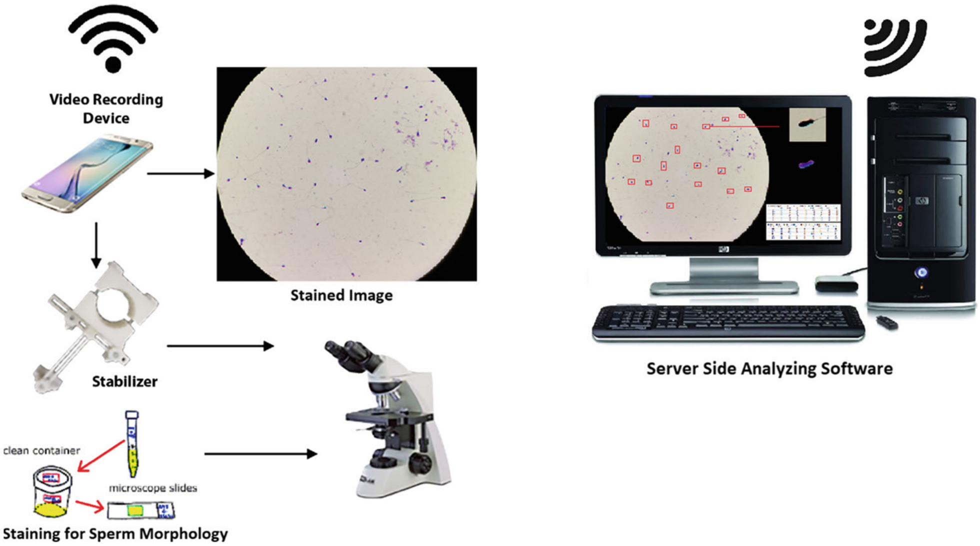


Fig. 2 Sperm Morphology Image Data Set (SMIDS) creation and MobileNet implementation for sperm abnormality classification

Fig. 3 Data acquisition approach



After capturing process of the ocular part, full-size images are sent to the server to be analysed for segmentation and classification steps.

In this study, 200 stained ocular images of 17 subjects, who visited the infertility centre of Istanbul University, were collected and investigated after obtaining informed consent from the couples. This study was approved by the ethics committee of Istanbul University. The age range who provided samples was 19–39 years. The laboratory checklist, which was published by Bjorndahl et al., was followed in the steps of preparation of samples for the analysis [16]. Semen specimen were collected from refrained males from any sexual activity (no ejaculation) for at least 2 days, but not more than 7 days. Patients have no missed early ejaculate fractions. Samples were requested from subjects as ejaculating into provided sterile sample cup by masturbation in the morning between 9 and 11 am. Samples were kept in 37° for 30 min after ejaculation process.

Due to the conventional morphological analysis procedure [54], semen samples were stained with a modified hematoxylin eosin assay before the data acquisition step. This process provides a better visualisation of sperm parts and makes the visual analysis more convenient for the experts. Subsequent to staining process, the ocular images of semen samples were captured by using the smartphone-based approach. The optical properties of the microscopy and the information of the recorded images are given in Table 1. Although the staining process maintains more enhanced morphological representation of possible sperm blobs, further image denoising approaches still needed to be applied for eliminating noise components and enhancing the visibility of ROIs.

2.2 Pre-processing

Pre-processing is one of the fundamental stages in any information retrieval applications based on image processing in order to suppress noise and/or undesired objects. Generally, a signal x contaminated with noise w can be formulated as in Eq. 1.

$$y = x + w \tag{1}$$

Table 1 The optical properties of the microscopy and the information of the recorded ocular images

Microscope	Olympus BX50
Magnification	20x
Illumination	12V/100W halogen lamp
Image resolution	3890 × 3000
Object space resolution	$1 \mu m /$ px

where x is the clean signal and w is the noise. In literature, the wavelet transform (WT) has been widely used in biomedical signals to de-noise and to enhance the ROIs [10, 34, 43, 62]. Traditionally, the WT-based denoising methods are carried out by employing the following steps; (i) decomposing the noisy biomedical one-/two-dimensional signals into several analysis levels, (ii) removing the noisy signal components according to thresholding techniques such as soft or hard thresholding and (iii) reconstructing the denoised biomedical signals. In traditional denoising methods, the signal components belonging the noise and ROI are both affected by the thresholding operation in the same level. However, it is well known that the signal components forming groups mostly belong to ROIs (sperm bodies in our study), while the noise components spread in a more coincidental manner. Therefore, in order to eliminate the random noise while keeping the sperm shapes, we used wavelet-based Modified Overlapping Groups Shrinkage (MOGS) technique. The observed signal x is accepted as group-sparse signal. By group-sparse, it is meant that large magnitude components of x tend not to be isolated. Unlike, large magnitude components tend to form groups. In this respect, large magnitude components of the signal are kept for further process [23]. In such cases, as a common practice, convex and non-convex optimisation methods can be employed to estimate sparse vectors from the noisy observation data [9]. Therefore, in our image enhancement problem, a solution $x^* \in \mathbb{R}^N$ can be formulated as

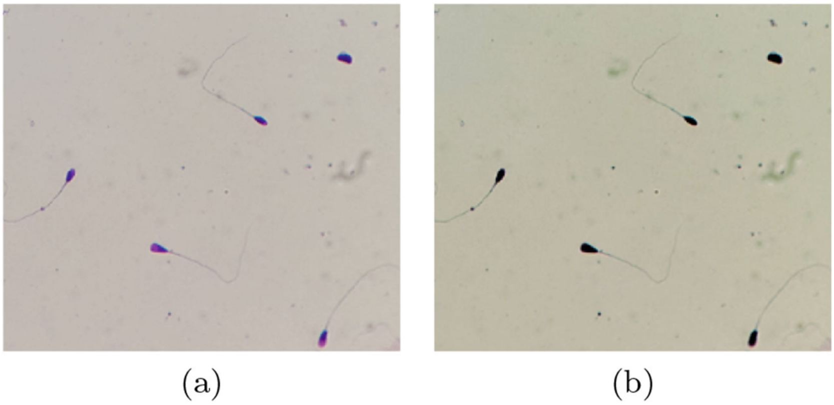
$$x^* = argmin_x \left\{ F(x) = \frac{1}{2} \|y - x\|_2^2 + \lambda R(x) \right\} \tag{2}$$

where $R(x)$ and λ denote the penalty function and regularisation parameter, respectively. $\lambda > 0$ is the satisfaction criteria. y is the measured signal (the sperm image in our case), x^* is the optimal solution of the MOGS regularisation optimisation and x is the group-sparse signal. In MOGS, the denoising problem is a convex function referring to a non-convex regularisation term. More detail about the MOGS approach can be found in [23, 27, 28, 35]. The MOGS effect is given in Fig. 4 over an example image. As expected, the magnitudes of the pixels forming a cluster (pixels belonging to the sperm body and tail) are enhanced with respect to background activity.

2.3 Segmentation

Medical image segmentation can be defined as the partitioning of a medical image (ocular images in our case) into non-overlapped, consistent regions which are homogeneous with respect to some characteristics such as the gray value or texture [19]. In this study, as an initial process, we employed the segmentation over full

Fig. 4 **a** Original ocular image **b** MOGS denoised ocular image



slice ocular images to extract the sperm and non-sperm parts as cropped ROIs. These parts have smooth variations in the color space which refers to hard detection in corners and transitions. Therefore, rather than defining a thresholding point, we focused on the color histogram based segmentation. In this respect, we tested two of the most popular clustering approaches using the distance metrics in histogram space as Fuzzy C-Means and regular K-Means similar to many studies in literature such as in [32, 55]. Additionally, a hybrid method that combines the MOGS and one of these two techniques was also performed. This idea was firstly proposed in [40] for a small size data set. In the current proposed study, the performance of the same strategy is measured for a larger data set consisting of 200 ocular images (the number of ocular images was 13 in [40]) and the results were presented. In order to visualise the clustering effect, an example of the segmented ROIs is given in Fig. 5. The results show that the MOGS increases the performance of segmentation and decreases the miss rates for both clustering techniques. The best segmentation

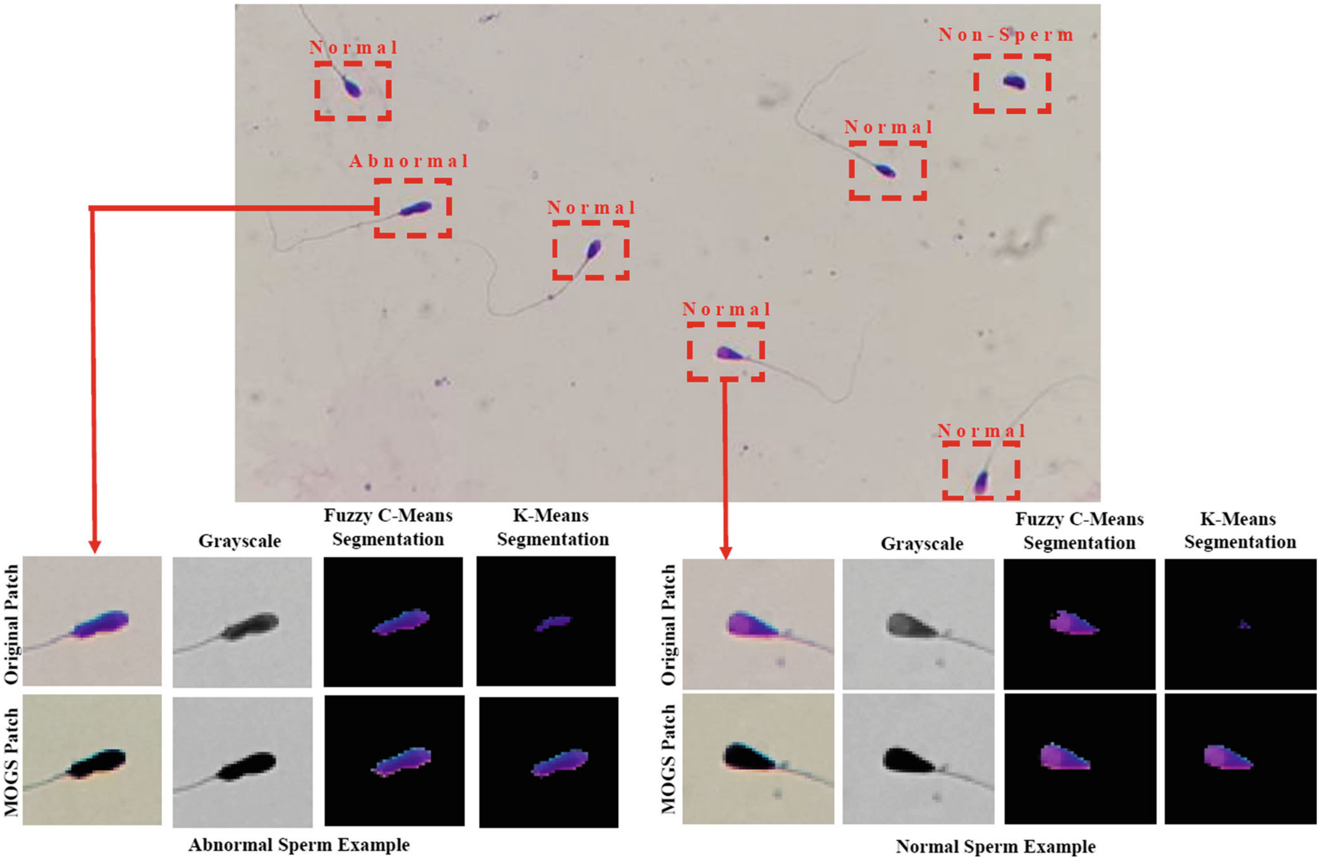


Fig. 5 Segmentation results of fuzzy C-means and K-means with/without MOGS

performance was obtained with Fuzzy C-Means clustering technique when the MOGS is applied to images as a pre-processing step. In contrast to the K-Means clustering, which restrict each pixel value to exactly one cluster, the Fuzzy C-Means clustering uses the idea of fuzziness for the belongingness of each image pixels [76] and this increases the segmentation performance.

2.4 Creation of the sperm morphology image data set

The proposed hybrid method consists of the segmentation and classification steps. Subsequent to automatically ROI segmentation and cropping steps for sperm and non-sperm parts, the sperm parts must be classified as normal or abnormal sperm. This part have been arranged manually by experts in order to create the Sperm Morphology Image Data Set (SMIDS). Therefore, all process is a semi-automated approach including an automated blob analysis step that is employed to assign class labels using spatial-based features for roughly identification for sperm/non-sperm parts and a manual validation step that is applied by human experts for defining normal/abnormal classes in sperm parts.

In the automated class labelling for sperm/non-sperm parts, spatial-based features including sperm eccentricity and sperm area, whose reference values were calculated from the mean values of randomly selected 100 sperm patches, were employed. The dimensions of normal sperm shape are defined as 2.5–3.5-μm width and 5–6-μm length. Normal sperm is characterised by an oval head in a single long tail. In the first phase of creation SMIDS, sperm (normal+abnormal) and non-sperm (occurred by the staining process) class labelling was performed by using an automated pre-classification process that uses the sperm eccentricity and sperm area values defined as in Eq. 3. The sperm eccentricity is defined as ratio of the major axis (sperm length) and the minor axis (sperm width). In addition, the area (S) of the blobs were measured by counting the pixel values located in segmented patches. The extracted blob features are illustrated in Fig. 6 over an example sperm. On the other hand, abnormal sperm shapes have minimal differences in spatial-based metrics, which makes hard to differentiate them from normal sperm shapes. Therefore, subsequent to automated sperm/non-sperm identification, the sperm shapes were manually labelled by experts as normal or abnormal, in order to finalise the SMIDS that includes the ground truth classes. Regarding to segmentation output, the missed sperms were manually corrected by the experts and added to SMIDS while the false-positives (stained blobs) were excluded. After all, SMIDS consists of 3000 patches (normal, abnormal, non-sperm) that were extracted from the microscopic 200 ocular images of 17 subjects. A total of 2027 of these patches were manually labelled as normal and abnormal sperm, whereas the 973 samples were assigned to non-sperm by the spatial-based automated features. This unique data set is freely available for benchmark studies. The detailed information of created data set is given in Table 2 and example images are presented in Fig. 7.

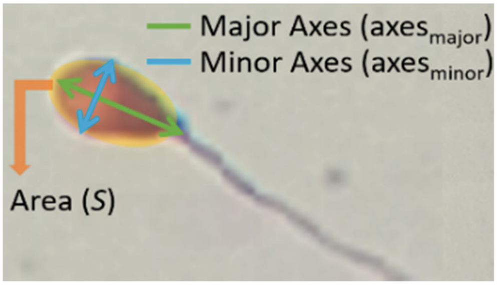


Fig. 6 Extracted spatial features by the blob analysis

$$1.4 < \frac{\text{axis}_{\text{major}}}{\text{axis}_{\text{minor}}} < 3 \qquad 30 < S < 60 \tag{3}$$

2.5 Feature extraction

In the classification phase of the proposed study, firstly, various feature extraction approaches including the DWT, DTWT and DTCWT as texture features; and SURF, MSER and KAZE as the descriptor-based features were applied. Later, their outputs were fed into the ensemble and conventional individual machine learning techniques. Secondly, as an end-to-end learning process, three CNNs architectures, named as the InceptionV3, VGG19 and MobileNets, were tested with the aim of sperm morphology classification.

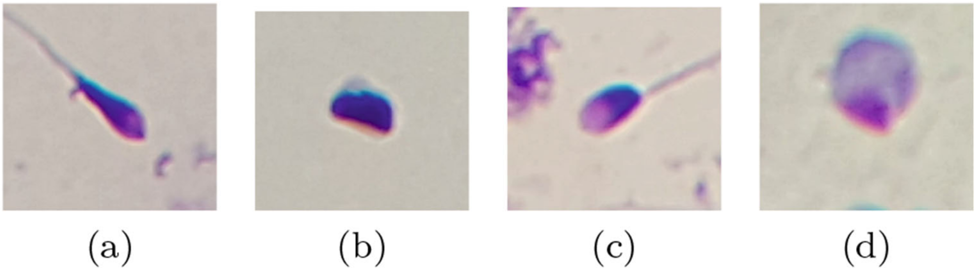
2.5.1 Wavelet-based feature extraction

Wavelet transform (WT) is a non-stationary signal analysing tool due to its flexible time-frequency representation ability. It provides a time-scale representation by scaling and shifting a mother wavelet ($\Psi(t)$). In this representation, signal can be examined in a good frequency resolution

Table 2 Data Set (SMIDS) class information

Normal sperm	1021
Abnormal sperm	1005
Non-sperm	974
Total	3000

Fig. 7 Example images from SMIDS. **a** Abnormal sperm. **b** Clusters occurred by the staining (non-sperm). **c** Normal sperm. **d** Optical microscope effects (non-sperm)



at low frequencies and good time resolution at high frequencies. The continuous WT can be formulated as.

$$CWT(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\Psi\left(\frac{t-b}{a}\right) dt \qquad a, b \in \mathbb{R}, a \neq 0 \tag{4}$$

where a and b indicate the scale and translation parameters of the input signal $(f(t))$ respectively. In theory, although the continuous WT provides powerful insight into a signal's time and frequency characteristics at the same time, it suffers from huge computation complexity and the redundant data that occur due to the continuous change of parameters. Therefore, in practice, it cannot be implemented in real time. A computerised version of the ordinary continuous WT is obtained by using discrete scale and translation parameters for the real-time applications. This computerised transform, which is called as discrete WT (DWT), results in countable set of coefficients with reduced memory requirements and higher computation speed for both one-dimensional signals and images. The DWT of a signal *f(t)* can be found by the following equation.

$$W_s(m, n) = \frac{1}{\sqrt{a_0^m}} \int_{-\infty}^{\infty} f(t)\Psi\left(\frac{t - nb_0a_0^m}{a_0^m}\right) dt \qquad a, b \in \mathbb{R}, a \neq 0 \tag{5}$$

where m and n are discrete scale and translation steps, while a_0 and b_0 are the discrete scale and translation step sizes [63]. In order to obtain even better performance, the DWT can be implemented by using multi-resolution analysis (MRA) approach [48, 68]. In MRA, a discrete lattice is employed in order to obtain even more computational efficiency as given in Eq. 6

$$\mathrm{DWT}_{f(n)} = \begin{cases} A_{j,k}(n) = \sum\limits_{n} f(n)G_j^*(n - 2^j k) \\ D_{j,k}(n) = \sum\limits_{n} f(n)H_j^*(n - 2^j k) \end{cases} \tag{6}$$

where $A_{j,k}(n)$ and $D_{j,k}(n)$ are the approximation and detail coefficients, while $G(n)$ and $H(n)$ refer to the low-pass and high-pass filters respectively. The two parameters j and k indicate the dyadic wavelet scale and translation factors [78]. A DWT structure is implemented by employing these digital filter pairs in an iterative pattern and additional down-sampler (which halves the number of samples in each iteration) operators are used to prevent redundancy. This iterative procedure can be applied with a binary tree in which the approximation coefficients of each level are used as the new input to the following level of analysis. This process can be repeated up to a certain number of levels for further analysis until a single sample is left at the end. More details about the mathematical structure and implementation procedure of the DWT for one-dimensional signals can be found in [48, 68].

In the decomposition of images (two-dimensional signals) case, a one-dimensional DWT tree is applied to both x and y dimensions separately in order to obtain the spatial-frequency information of both the rows and columns of the image. As a result of one-level decomposition of an image by using a two-dimensional DWT (2D DWT), four sub-band coefficient sets that are named as LL: low-low, LH: low-high, HH: high-high and HL: high-low are obtained. These LH, HL and HH sub-bands mostly give information about the high-frequency components of the analysed image along horizontal, vertical and diagonal directions, respectively. On the other hand, the LL sub-band mostly represents low-frequency components of the image. In order to interpret the image in a hierarchical framework, the LL component, also named as the approximation component, is employed as the input of next analysis level resulting in the coarser components during the full decomposition [53, 78].

Despite the advantages of DWT applications in both one-dimensional signals and images, the DWT suffers from being shift variance causing dramatic changes in the energy distribution of decomposed sub-bands as a reaction to small shifts in the original image. Obviously, this will result in degradation of the classification performance of the learning model when DWT coefficients are utilised as features. Dual tree wavelet transform (DTWT) and dual tree complex wavelet transform (DTCWT), which have enhanced shift invariance property, were proposed as two modified versions of classical DWT with limited redundancy [61, 63]. The DTWT utilises six oriented wavelets. It consists of two real separable 2-D wavelet transforms in parallel resulting in a two times expansive transform. As a further improvement, in the DTCWT, both oriented and complex wavelets can be obtained by using DWTs having 90° phase difference. In this implementation, the low- and high-pass filter pairs are complex resulting in four times computational cost comparing with DWT. Applying the DTWT to images results in six directional selective sub-bands having ± 15°, ± 45° and ± 75° for each scale. On the other hand, when the DTCWT is employed, an additional complexity property, which induces near shift-invariance, is also achieved for the same six orientation. In this study, mean, standard deviation and entropy of the raw sub-band coefficients obtained by 2D DWT, DTWT and DTCWT were extracted as statistical features for each image in the SMIDS. Then, these features were combined and fed to the individual and ensemble classifiers as a single row vector.

2.5.2 Descriptor-based feature extraction

Descriptors are generally called as the interest point extraction techniques. Interest points are defined as the key points of the images which represent the objects in a frame with different aspects related with the utilised descriptor algorithm. An object can be represented by one or more interest points according to the applied descriptor technique. Some points such as noises or halos caused by light reflection in microscopic images are undesirable. Therefore, a majority voting procedure, which determines the results by observing the dominant features, should be applied. In this study, the performance of three descriptors, Speed Up Robust Features (SURF), Maximally Stable Extremal Regions (MSER) and KAZE, were tested by employing a majority voting idea that was used to obtain the feature vectors of created SMIDS.

The SURF-algorithm [12] is based on a very similar principle with the Scale-Invariant Feature Transform (SIFT), which was proposed by Lowe [47]. However, the SURF uses a different scheme and it is faster than SIFT with better performance. SURF mainly uses the idea of integrating the image, which is used as a quick way of calculating the sum of pixel values in a given image—or a rectangular grid subset of an image. In this respect, Eq. 7 is utilised for each rectangular subsets having *x* width and *y* height filter sizes.

$$I_{\sum}(x, y) = \sum_{i=1}^{i \le x} \sum_{j=1}^{j \le y} I(i, j) \tag{7}$$

where $I(x, y)$ represents the sum of intensity values. Summing process is performed over *(i, j)* pairs starting at (0, 0) up to the indicated rectangular subset sizes (*(x, y)*). This process is performed for each pixels by changing the size of rectangular subset sizes (*x* and *y*) in the whole image. As a result, each point contains its own integration information. Figure 8 indicates the integration process for 4 points in the image. Integration information of each point was calculated by a rectangular filter starting from *(0, 0)* to $I_n(x, y)$ coordinates.

The density of the rectangle taken from any part of the images is calculated in four steps. The sum of the values of two corner points is subtracted from the corner with the greatest value. Afterwards, the corner with the smallest value is added to the result. This allows field intensity calculations to be performed independently of the size.

$$I_{\text{sum}} = I_{\sum}(x_a, y_a) + I_{\sum}(x_d, y_d) - I_{\sum}(x_b, y_b) - I_{\sum}(x_c, y_c) \tag{8}$$

Then, the Hessian matrix is created by employing the 2nd-order Gaussian kernels in 3 direction as *x*, *y* and *xy* and then determinant is calculated to detect the variations over integral rectangle filters. The determinant responses are normalised to scale. At the last step, the Haar wavelets of size 4σ in *x* and *y* directions are computed for the orientation assignment.

$$H(x, y) = det \begin{bmatrix} A & C \\ C & B \end{bmatrix} = det \begin{bmatrix} \partial^2 f/\partial x^2 & \partial^2 f/\partial x \partial y \\ \partial^2 f/\partial x \partial y & \partial^2 f/\partial y^2 \end{bmatrix} \tag{9}$$

Another desriptor, Maximally Stable Extremal Regions (MSER), uses the idea of detection for the maximum stable blobs between the multiple versions of the image. Then, the features of the stable regions are extracted for

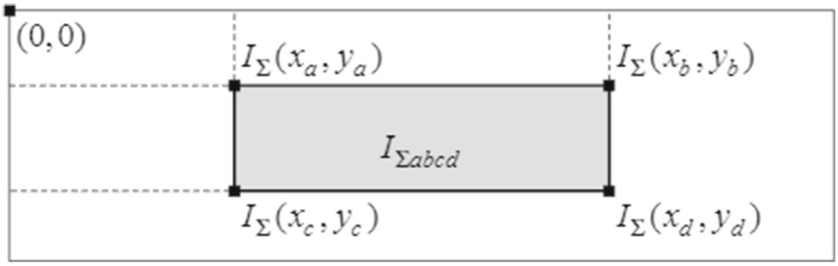


Fig. 8 The calculation of SURF integrated density information for 4 points; *a, b, c, d*

classification process. This technique was proposed by Matas et al. [49]. Generally, the stability detection process is performed by using a thresholding process. All the pixels below a given threshold are assigned to ‘black’, while all those above or equal this threshold are assigned to ‘white’. In this technique, a sequence of thresholded images, in which each image corresponds to an increasing threshold value, is obtained based on a source image. The first thresholded image will be a white image, then ‘black’ spots corresponding to local intensity minimas will appear and then grow larger. These ‘black’ spots will eventually merge into a whole black image. The set of all connected components in the sequence is the set of all extremal regions to be used for feature extraction process. MSER analysis takes more time when compared with SURF due to the checking and validating processes of stability regions between all images. MSER is also robust descriptor to affine transformations.

KAZE features, which are employed as the last set of descriptors, are obtained by a multiscale 2D feature detection and description algorithm that is applied in non-linear scale spaces. Previous approaches detect and describe features at different scale levels by building or approximating the Gaussian scale space of an image. However, Gaussian blurring does not take into account the natural boundaries of objects. Therefore, it smoothes details and noise components in the same degree resulting in a reduced localisation accuracy and distinctiveness. In contrast, KAZE features are 2D features calculated in a non-linear scale space by means of non-linear diffusion filtering. In this way, blurring can be implemented locally and adaptive to the image data resulting in reduced noise while retaining object boundaries. This improvement in noise reduction results in a superior localisation accuracy and distinctiveness when compared with the previous approaches. Similar to MSER, KAZE features have higher computational complexity than SURF as a result of the non-linear scale space construction process.

An illustration of interest points and extracted feature matrices by the implemented descriptors is given in Fig. 9. Each descriptor is represented by a different color including red, blue and green colors for the SURF, MSER and KAZE respectively. s_p, k_p and m_p represent the SURF, KAZE and MSER feature matrices extracted from an image patch p, respectively. Each descriptor may define different number of interest points in the image, but a 1×128 feature vector is extracted from each interest point at the end. It is observed that the maximum number of interest points was extracted by MSER descriptors (m_p) for the SMIDS due to the multiple sequential thresholding processes. MSER mostly identified the interest points at the same location of the image where the black pixels are clustered, such as in acrosome part of the sperm. KAZE (k_p) has the second highest number of interest points because of the blurring process. KAZE reduced the unwanted effect of the small particles by blurring them. SURF described the minimum number of interest points in the image due to the Hessian matrix. However, when the number of the defined interest points with various descriptors is investigated, it is seen that there is no direct relation between the interest points numbers and classification performance. The higher classification performance is more related to the power of informative points rather than defining high number interest points, which mostly located in a limited region of the image. This behaviour can be seen in Fig. 9 for the MSER-based key point description outputs, in which most of the MSER features are located in the acrosome of analysed sperm. On the other hand, the SURF has also limited performance due to the insufficient number of represented objects even if they are located in different regions. Finally, when the KAZE descriptor is investigated, it is observed that sufficient number of multiple interest points located at

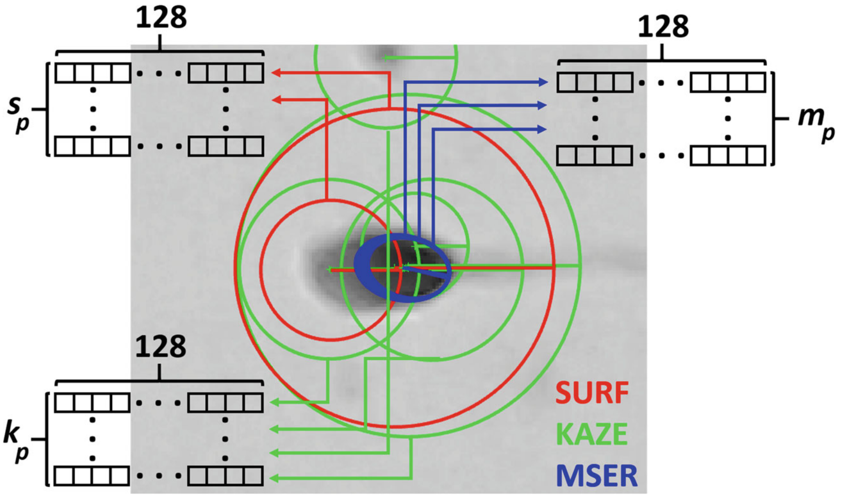


Fig. 9 Interest point demonstration of SURF (red), MSER (blue) and KAZE (green) descriptor

different regions having higher object representation ability can be obtained.

2.6 Classification

In the classification phase of the proposed study, three main classifier approaches to find optimum algorithm have been tested. Firstly, the obtained wavelet- and descriptor-based features were individually fed into conventional individual machine learning techniques consisting of SVM, KNN and DT models. Secondly, the performance of the ensemble versions of DTs in terms of Adaboost, Rusboost and Bagging was investigated under the ensemble learning idea. Thirdly, the performance of three different CNN configurations, as the deep learning-based state-of-art techniques, was obtained.

2.6.1 Individual machine learning techniques

In the individual classification part, the SVM, DT and *k*-NN classification techniques are selected as the conventional individual machine learning models. SVM is known to be a prominent classification algorithm that can be used in large-scale data sets and is more efficient than statistical and basic neural classifiers [58, 72]. Kernel-based learning, which aims to separate data in high-dimensional feature space by mapping data points with a kernel function, is the core idea in the SVM [5]. It gives higher classification accuracies with even small size training sets [59]. In this study, SVM was tested with three different kernels (Linear, RBF, Polynominal) and the penalty parameters of each classifier were set to 1. As an another well-known technique, DT, which is based on a tree terminology, was chosen as the second classifier [18]. In DT terminology, the root represents the most informative features and the leaves are created by some mathematical formulas that results in information gain for other features such as Entropy, Twoing, KNN and Gini. Leaves positions are arranged by the order of calculated information gains. Generally, DT is an easy to implement method, in which the interpretation of the classification is much easier than the other methods such as *k*-NN. However, DT classifiers have an important drawback as attaining low discrimination performance on large-scale data sets (like the SMIDS) due to their model simplicity. Furthermore, the pruning process, which needs to be controlled to avoid over fitting, is an important issue in DT. In this study, a DT model having the pruning functionality and Gini's Diversity Index was used to calculate information gain. As the last classifier, *k*-NN method, which uses the cumulative distance between the testing sample and *k* nearest members of the training set, to define the final class label was employed as the distance-based classification model for performance comparison of extracted features.

2.6.2 Ensemble learning techniques

As the second approach of the classification phase, the ensemble methods, which employ the combinations of several individual learners, were tested. In the ensemble methods, some base individual learners are applied to the data and the decisions of these individual models are combined by using voting approaches. In our study, DT models were chosen as the base learners and the majority voting, in which the final decision is obtained by averaging the outputs of individual classifiers, was utilised. As the first ensemble classifier, Bootstrap aggregating (Bagging) [17], which improves the quality of estimates by the help of well-formed train samples (also named as resampling), was tested. In the Bagging, new data sets having smaller number of samples when compared with the original SMIDS were produced by random sampling with replacement from the original set. Subsequent to resampling, each new data set, which were statistically independent, was given into a weak learner and a final label was obtained by using the majority voted output of weak learners' decisions. When the Bagging is considered, any element has the same probability to appear in a new data set. However, in Boosting, which is an alternative to Bagging, a biased resampling method, in which the observations are weighted and therefore some of them take part in the new sets more often, can be used. In Boosting algorithms, each weak classifier was trained on a small data set, taking into account the previous weak classifiers' performance. After each training step, the weights were redistributed. If there were some misclassified samples, the weights of these samples were increased to emphasise the most difficult cases. By doing this, the subsequent learners of the algorithm was forced to focus on difficult cases during the training and this increases the discriminating capability of general ensemble model. In the proposed study, Adaboost [30] and Rusboost [60] algorithms were employed as the ensemble Boosting methods. Adaboost is an abbreviation of adaptive boosting which is mainly outperforms regular boosting technique in the manner of solving over fitting problem, while Rusboost is another popular boosting technique mainly dealing with data imbalance problem.

2.6.3 Deep learning techniques

As seen in the previous sections, extensive domain-specific feature extraction methods must be applied before the traditional individual and ensemble classification approaches. By contrast, DNNs require no conventional feature extraction steps and they can be trained end-to-end from image patches resulting in a more straightforward approach. DNNs are currently employed in a myriad of visual applications and they have been shown to exceed the human performance [51, 57, 66]. In case of biomedical applications, the DNNs have played an important role in various applications such as genomics to gain insight to the genetic diseases such as autism, cancers and spinal muscular atrophy [4, 75, 77, 79]. Additionally, DNNs were successfully employed in various medical imaging problems related to the skin cancer [29], brain cancer [42] and breast cancer [74].

As a specific type of DNNs, the CNN architecture allows to greatly decrease the number of network parameters compared with a traditional fully connected network by employing convolutional operations to only small regions of the input space and by sharing parameters between regions [7]. In CNNs, the entire computation process is performed as a sequence of operations on the outputs of a previous layer. In the final stage of the operations, generally, fully connected layers are used to generate the output of the network, for example in our case, a probability that an image patch contains one of the three classes: non-sperm, normal sperm or abnormal sperm. In CNNs, the network has no memory and the output for an input is alfways the same irrespective of the sequence of inputs previously given to the network [69].

In our study, as the third approach for the classification phase, three CNNs named as the InceptionV3 [71], VGG19 [67] and MobileNets [36] were tested as DNN classifiers. Normally, the performance of the utilised DNNs was shown by using the ImageNet dataset (ILSVRC2014). This data set contains 1000 categories and 1.2 million images, which indicates 1200 images per class. Our data set includes 3000 images for 3 categories. For all three CNNs, the networks were fully trained by using the Adamax algorithm [45] with 900 images for each class similarly as in ImageNet data set. The batch sizes were set to 5 and the learning rates were set to 0.0001. The training algorithms were adjusted to minimise the categorical cross entropy function. The networks were trained by using Python Keras [24] library, which allows the users easy and fast prototyping of the neural networks, running on top of Tensorflow [1]. All work was performed on a computer with Nvidia Quadro P5000 GPU with 16 GB dedicated memory, Intel Core i7-6700K processor and 48 GB RAM memory.

As the first approach for DNN-based learning, InceptionV3 CNN architecture [71] was employed. The first version of the Inception architecture (InceptionV1) [70] was proposed in ImageNet ILSVRC-2014 [56] and won the first place in the classification competition. The InceptionV1 introduced an inception module, which was composed of parallel connections, whereas previously there had been only a single serial connection. The InceptionV1 consists of 22 layers, which includes various sized filters (for example, 1 × 1, 3 × 3, 5 × 5) that are used for each parallel connection. As an improvement, smaller 1D filters to decompose the convolutions with reduced number of multiply-and-accumulate operations were employed in InceptionV3. It was reported in [71] that InceptionV3 achieves over 3% lower top-5 error than InceptionV1 with 2.5× increase in computation cost.

The VGG19 [67] net was tested as the second DNN-based learning approach. The first version of VGG nets (VGG16 [67]) was also tested in ImageNet ILSVRC-2014 and it won the first place in the localisation competition. VGG16 consists of 16 layers including 13 convolution and 3 fully connected layers. It uses large convolution filters, e.g. (5 × 5) that are built from multiple smaller filters (e.g., 3 × 3, having fewer weights) to control the computational cost of going deeper. VGG19 was proposed as an improvement to VGG16 with increased number of layers (19 layers are utilised) and it performed 0.1% lower top-5 error rate than VGG16 at the cost of 1.27× more multiply-and-accumulate operations.

Thirdly, as a DNN method that can deployable on mobile devices, the performance of MobileNets [36] was also tested on SMIDS. In the implementation of MobileNet, the depthwise separable convolutions (DSC), which is a form of factorised convolutions, are employed in order to reduce the computational cost and model size. By employing the factorised convolutions, the standard convolution operation is factorised into a depthwise convolution (DC) and an 1 × 1 convolution called a pointwise convolution (PC). The DCs are used to apply a single filter per each input channel (input depth), while PC (a simple 1 × 1 convolution) is used to create a linear combination of the output of the depthwise layer. DC is extremely efficient relative to standard convolution to reduce the number of multiply-and-accumulate operations. However, by using the DC, only the input channels are filtered and the necessary combination operation to create new features is missing. Therefore, a PC layer that computes a linear combination of the output of DC via 1 × 1 convolution is applied in order to generate needed new features. As a result of employing 3 × 3 DSC, a significant computational improvement reaching up to 8 to 9 times less computation cost than the standard convolutions with only a small reduction in accuracy can be obtained [36].

In medical image analysis, data size may be small and the trained models may have difficulty to generalise data from the validation and test set. Therefore, these DNN models may suffer from the overfitting problem. In such cases, the training set can be augmented by scaling, rotating or cropping the existing images to reduce the overfitting and enhance robustness [46]. In this study, in order to reduce the overfitting and increase the classification performance by using the relevant library of Keras tool, the existing SMIDS size was augmented using various augmentation sizes (1 to 5)[24].

2.7 Performance metrics

True-positive (TP), false-positive (FP) and missed sperm numbers (Misses) were used as the performance criteria for measuring the performance of segmentation and pre-classification phases by comparing the results of automated system with ground truths labelled by human experts. TP and FP show the correctly pre-classified sperm numbers, and the objects that are wrongly pre-classified as sperms, respectively. The numbers indicated under the 'misses' refer to the system segmentation failure in sperm detection step.

After the segmentation of image patches from the ocular full slice images, final classification of the SMIDS was evaluated in terms of 'Accuracy', 'Precision', 'Recall' and '*F*-Measure' metrics. Accuracy is the key benchmark metric for any classification study. It indicates the number of correctly classified images within all data set. However, higher rates in accuracy do not reflect the model success. The distribution of the confusion matrix is more important than observing the only one case. Therefore, Recall and Precision which focus on the incorrect classified samples were also derived as other performance metrics in order to evaluate the model in all aspects. Additionally, *F*-Measure score was used in the performance analysis of the classification techniques. It reaches to the best score at 1 and worst score at 0.

3 Experimental results

In the proposed sperm morphology segmentation and classification framework, the sperm detection performance was investigated by evaluating the pre-classification step and then the three-class classification performance of proposed framework was tested on the created data set (SMIDS) by using the confusion based metrics.

3.1 Segmentation/pre-classification results

In the ocular images, three different class of objects may occur; (i) non-sperm noise components that may be formed due to the excessive use of the substance in the staining process, (ii) abnormal sperm that may be the indicators of various problems related with infertility and (iii) normal sperm. The non-sperm noise components may affect the abnormal sperm counting due to the high similarities to the abnormal spermatozoa, and therefore, the correct detection of possible sperm components has great importance to prevent the misdiagnosis of sperm morphology-based diseases. In this respect, an automated sperm segmentation phase was proposed as the first step of this study. In order to detect the sperm components, two clustering techniques, Fuzzy *C*-Means and *K*-Means, were applied to 200 full slice ocular images obtained from 17 subjects containing 2026 sperm (normal + abnormal) and 974 non-sperm segments. Additionally, in order to enhance the segmentation performance of these two methods, a wavelet-based Modified Overlapping Groups Shrinkage (MOGS) technique, which reduces the noise level while keeping possible sperm shapes, was employed as a pre-processing step of clustering techniques. Subsequent to possible sperm detection phase, a sperm morphology data set (SMIDS) was created by using the detected objects with the aim of employing them to test the classification performance of proposed framework. To do so, a pre-classification step based on predefined spatial features, such as the eccentricity and area, was applied and the segmented sperm objects were obtained. Then, the results of the pre-classification step were controlled by the human experts and the final labels of the SMIDS were determined. It is observed form Table 3 that the segmentation performance of the Fuzzy *C*-Means detector revealed superior results over the well-known *K*-Means detector in the both MOGS applied/non-applied cases. The highest average recall and precision rates were achieved as 94.7% and 90.9% respectively, when the Fuzzy *C*-Means clustering accompanied by the MOGS denoising was applied. In all the 17 subjects, the usage of MOGS enhanced the precision and recall rates. Additionally, significant increase in percentages, as 16.6% and 42.5% respectively, for the average precision and recall rates were achieved with MOGS when the Fuzzy *C*-Means detector was used. As highlighted in Fig. 5, the sperm regions could not be fully represented in the lack of MOGS usage due to the weak gray-scaled histogram information of the blobs. This dramatic loss of detected sperm pixels distorts the spatial discriminative capability of the eccentricity and area features. When the results of each subject were investigated, it was seen that the most incorrect detections were observed for subjects 3, 5 and 11, whose ocular images were mostly contaminated with the overdose usage of staining process. Proposed system could only achieve 65.3%, 71.1% and 61.4% precision rates because of the detection of non-sperm structures as sperms. In another example, system failed in detection of sperms for subject 2, whose sperm forms gathered in a single region of microscope ocular images. Recall ratios with 76.9% emphasise that 9 out of 39 sperm cannot be identified as sperm due to the occlusion problem. Except these, in all the other subject cases, the proposed system resulted in performance metrics over 85% and this shows the robustness of segmentation process.

When the contribution of MOGS is examined in more detail, it is observed that MOGS denoising increased all the performance metrics for both clustering techniques. It enhanced the precision rates of Fuzzy *C*-Means and *K*-Means from 74.3% and 46.4% to 90.9% and 89.0%,

Table 3 Pre-classification results of segmentation techniques for sperm patches with/without MOGS

			Patient ID																	
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
		GT	341	39	17	94	28	242	172	144	106	81	28	114	161	135	112	58	155	2027
MOGS	Fuzzy C-Means	TP	314	30	17	85	27	224	163	140	104	79	27	111	158	130	108	57	149	1920
		FP	9	0	9	1	11	7	17	15	8	12	17	18	20	18	13	3	15	193
		Misses	27	9	0	9	1	18	9	4	2	2	1	3	3	5	4	1	9	107
		Precision	97.2	100	65.3	98.8	71.1	96.9	90.6	90.3	92.9	86.8	61.4	86.1	88.7	87.8	89.3	95	90.8	90.9
		Recall	92.1	76.9	100	90.4	96.4	92.6	94.8	97.2	98.1	97.5	96.4	97.3	98.1	96.3	96.4	98.3	94.3	94.7
	K-Means	TP	308	35	17	83	28	213	160	139	103	79	27	111	154	130	108	58	145	1898
		FP	17	0	12	0	14	11	21	17	9	13	19	21	19	25	14	4	18	234
		Misses	33	4	0	11	0	29	12	5	3	2	1	3	7	5	4	0	10	129
		Precision	94.8	100	58.6	100	66.7	95.1	88.4	89.1	92.0	85.9	58.7	84.1	89.0	83.9	88.5	93.5	89.0	89.0
		Recall	90.3	89.7	100	88.3	100	88.0	93.0	96.5	97.2	97.5	96.4	97.4	95.7	96.3	96.4	100	93.5	93.6
No MOGS	Fuzzy C-Means	TP	133	10	10	46	17	105	103	87	68	40	10	66	102	68	78	29	87	1059
		FP	14	18	27	8	7	25	25	19	23	19	21	26	35	21	37	19	22	366
		Misses	208	29	7	48	11	137	69	57	38	41	18	48	59	67	34	29	68	968
		Precision	90.5	35.7	27.0	85.2	70.8	80.8	80.5	82.1	74.7	67.8	32.3	71.7	74.5	76.4	67.8	60.4	79.8	74.3
		Recall	39.0	25.6	58.8	48.9	60.7	43.4	59.9	60.4	64.2	49.4	35.7	57.9	63.4	50.4	69.6	50.0	56.1	52.2
	K-Means	TP	93	5	8	17	9	31	19	12	37	19	7	43	53	18	50	16	56	493
		FP	21	13	42	17	11	63	28	37	45	39	31	51	39	28	45	28	31	569
		Misses	248	34	9	77	19	211	153	132	69	62	21	71	108	117	62	42	99	1534
		Precision	81.6	27.8	16.0	50.0	45.0	33.0	40.4	24.5	45.1	32.8	18.4	45.7	57.6	39.1	52.6	36.4	64.4	46.4
		Recall	27.3	12.8	47.1	18.1	32.1	12.8	11.0	8.3	34.9	23.5	25.0	37.7	32.9	13.3	44.6	27.6	36.1	24.3

respectively. The most significant increments were observed in sensitivity rates which directly refers to correct sperm detection performance. It was increased from 52.2% and 24.3% to 94.7% and 93.6% for Fuzzy *C*-Means and *K*-Means respectively. In both scenarios as with and without MOGS applied images, Fuzzy *C*-Means mostly surpassed *K*-Means. The maximum difference between these two clustering methods was observed at segmentation of images without MOGS denoising process. Overall scores were increased by Fuzzy *C*-Means more than 25% in both performance metrics. In the segmentation of MOGS denoised images, Fuzzy *C*-Means performed 1% higher performances than *K*-Means in both metrics. In other words, MOGS denoising process increased the *K*-Means performance more than Fuzzy *C*-Means clustering when the original image segmentation results are taken as reference (without MOGS) and this proves that MOGS denoising has a big impact on robustness of both clustering algorithms. Regarding to the obtained pre-classification results, Fuzzy *C*-Means over the MOGS denoised ocular images are dedicated as the sperm and non-sperm detection approach of the proposed framework. Then, each detected region is extracted from original image to create a data set.

3.2 Classification results

After the segmentation and pre-classification of ROIs in ocular images, the labels of pre-classified images were checked by human experts and the sperm shapes were sub-categorised as normal/abnormal sperms. Then the sperm morphological data set, nominated as SMIDS, was created with the ground truth label information. In the final classification step, we aimed to find the most successful machine learning model that can be used in the decision step of our framework by investigating several models. In this respect, several conventional individual and ensemble machine learning methods employing various feature extraction techniques, and novel DNN models that are trained/tested with end-to-end approach were employed. As the validation method, *K*-Fold Cross Validation, with $K=5$, was selected in all the testing/training phases to avoid bias problem. The models have been trained with 80% of data and the performances measured over rest of 20% data set in each fold. Lastly, the average performance of folds is accepted as the general performance of the models. This technique is accepted as an alternative and more objective way to standard validation (hold-out) technique in literature [8]. DWT, DTWT and DTCWT wavelet-based statistical features, and KAZE, SURF and MSER descriptor-based features were extracted and used in the conventional and ensemble machine learning techniques. In deep learning models, InceptionV3, VGG19 and MobileNet networks were utilised and their results were reported in this study.

In the evaluation of wavelet-based features, for all the traditional classification approaches, various decomposition levels were tested. Highest accuracy value was obtained with the SVM classifier when the RBF kernel was employed similar to our preliminary study performed on small size of created data set [39]. The accuracy values changing with decomposition level for SVM classifier with RBF kernel are presented in Fig. 10. DTWT-based statistical features gave the highest accuracy rate (80.47%) when the decomposition level was chosen as seven. This implies that seven level analysis results in optimum frequency resolution. Higher and lower number of decomposition levels than seven give redundant and deficient information for sperm morphology classification problem, especially for the presented data set (SMIDS).

The combined representation of all the accuracy values related with the conventional individual and ensemble learning models for the wavelet transform-based features (seven level decomposition) and descriptor-based spatial features are given in Table 4. The highest accuracy rate was obtained as 83.82% with the RBF kernel SVM model trained and tested by KAZE descriptor-based features. The worst classifier was observed as *k*-NN over DWT-based

Fig. 10 The Accuracy ratios over wavelet decomposition levels - SVM RBF classifier

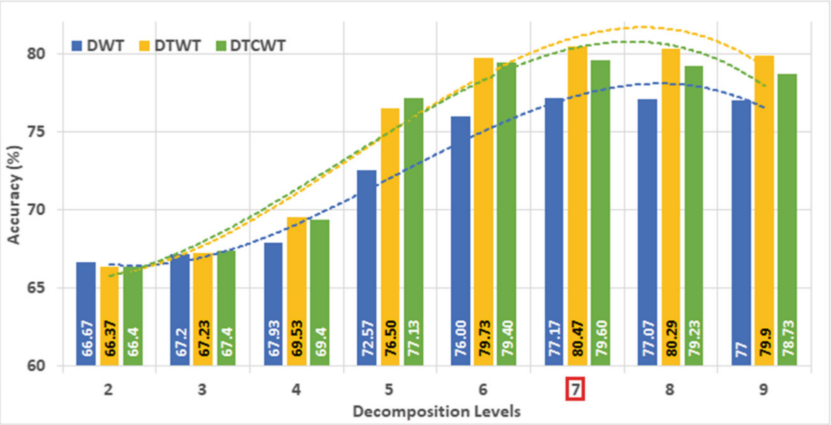


Table 4 Classification results of wavelet and descriptor based features with conventional and ensemble techniques

		Ensemble classification techniques			Conventional classification techniques					
		Adaboost	Bagging	RusBoost	SVM (Kernels)				DT	KNN
					Polynomial (d)		RBF	Linear		
					4	3				
Wavelet	DTCWT	78.20	78.88	66.02	77.70	80.07	79.60	75.93	66.93	50.60
	DTWT	76.33	78.42	66.21	78.60	80.07	80.47	75.83	66.30	48.80
	DWT	74.37	76.32	65.37	77.33	76.43	77.17	72.30	66.53	47.23
Descriptor	KAZE	77.15	82.73	74.28	81.85	83.16	83.82	73.18	72.11	74.62
	SURF	64.03	71.30	67.40	77.60	72.93	75.00	55.13	60.30	71.87
	MSER	74.87	77.37	70.57	80.67	78.87	80.17	73.83	72.27	77.70

features (47.23%). According to the results, KAZE features are the most informative features for Bagging, RusBoost, Polynomial and RBF kernel SVM classifiers in the sperm morphology representation problem. For the DT and KNN models, highest accuracy rates were achieved with MSER descriptor features. In particular to Adaboost and Linear SVM classifiers, wavelet-based features performed better than the descriptor-based features.

Addition to conventional individual and ensemble classifiers, three CNNs were also tested in sperm morphology classification problem as deep learning-based classification approaches. We employed different size of data augmentation in terms of scaling, resizing, shifting, cropping etc. due to the necessity of large amount of data in deep learning-based approaches. We carefully chose the ranges of the applied augmentation techniques in order to avoid disrupting the original normal/abnormal sperm shapes. Otherwise, normal sperm shapes might be detected as abnormal with incorrectly configured augmentation techniques. These implemented data augmentation effects are given in Fig. 11 over an example abnormal sperm image.

More data augmentation causes the extreme processing times and system resource consuming. In this respect, we evaluated the networks due to their performance metrics and training processing times. The results are graphically illustrated in Fig. 12, in which the *x*-axis refers the augmentation size where 1 indicates the only original data and others are the augmented versions.

The classification accuracy values that are given in Fig. 12 are increasing in an aggressive manner between augmentation values 1 to 4 for all networks. But, results are decreased in the level 5 for VGG19 and Inception networks

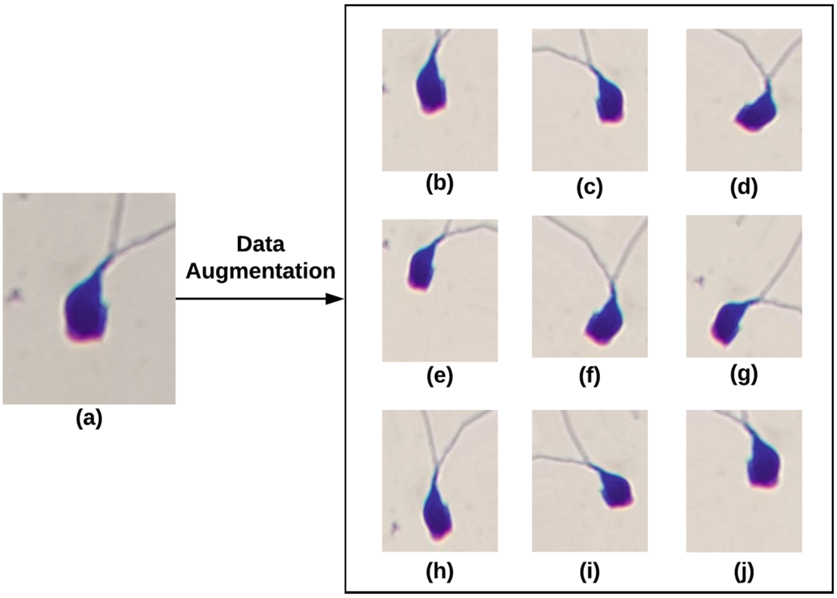


Fig. 11 The effects of the implemented data augmentation technique. **a** Original abnormal sperm image, **b** height shifted and sheared, **c** horizontal flipped, **d** horizontal flipped and zoomed, **e** width and height shifted, **f** horizontal flipped and height shifted, **g** rotated, **h** height shifted, **i** horizontal flipped and width shifted, **j** horizontal flipped and height shifted

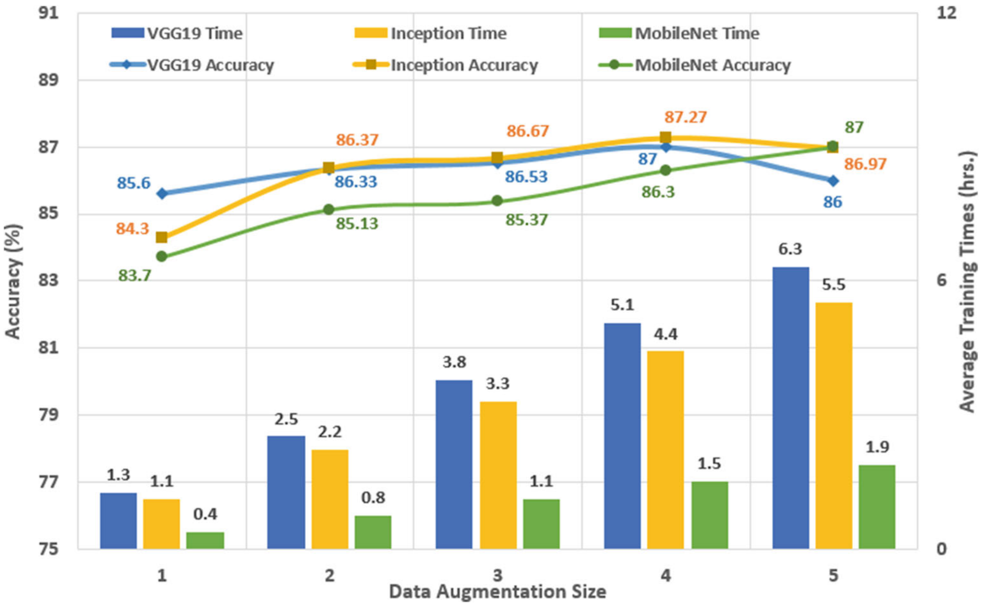


Fig. 12 Accuracy results and training times of deep learning networks per each fold over different data augmentation size (1 = original data)

where MobileNet still has an increment. Addition to the classification performance increase, the processing times of the trained networks are also rising in parallel. The required training time values for each network are shown as bar charts in Fig. 12. The maximum data augmentation level was limited as 5 in training sessions, because the required processing time values are dramatically rising while the classification accuracy values remain stable or decrease for the levels higher than 5.

In the deep learning tests, the highest accuracy rate (87.3%) was achieved by InceptionV3 network with 4 times data augmentation size. However, the average processing time for training is excessive (4.4 h). Another network, VGG19, resulted in 87% accuracy rate. But, it is more time and source consuming network than InceptionV3. VGG19 completed its training process in 25.5 h totally where the average is 5.1 h for each fold. Additionally, both networks are not suitable to use in mobile devices due to their excessive network sizes and they need to be executed in the server side. On the other hand, MobileNet which is a small network and deployable in phone type applications also resulted in 87% accuracy rate with less amount of training time and source consuming for each fold (1.9 h). More detailed results such as individual precision and recall metrics are given in Table 5 for the best obtained accuracy rates.

Figure 13 indicates the best obtained accuracy rates of conventional individual/ensemble models and DNN techniques. The highest accuracy for the conventional techniques were measured with RBF kernel SVM classifer trained with KAZE features. However, all CNNs surpassed the conventional techniques due to their ability to represent high-level features with advanced learning concepts. The accuracy variations between the folds, which indicates the model consistency, are also presented in Fig. 13. In this respect, VGG19 is the most consistent classification model and resulted in ±1.12% accuracy variation between folds. However, when the models are compared in terms

Table 5 Detailed best classification results of deep learning networks

	VGG19			InceptionV3			MobileNet		
	Normal	Abnormal	Non-sperm	Normal	Abnormal	Non-sperm	Normal	Abnormal	Non-sperm
Precision	0.83	0.94	0.85	0.84	0.9	0.88	0.85	0.93	0.84
Recall	0.82	0.91	0.87	0.82	0.94	0.86	0.80	0.92	0.88
F-Measure	0.83	0.93	0.86	0.83	0.92	0.87	0.83	0.93	0.86
Accuracy	88.47	95.23	90.3	88.7	94.77	91.07	88.57	95.37	90.07
Overall Acc.	87			87.27			87		
Training time (hrs)	5.1			4.4			1.9		

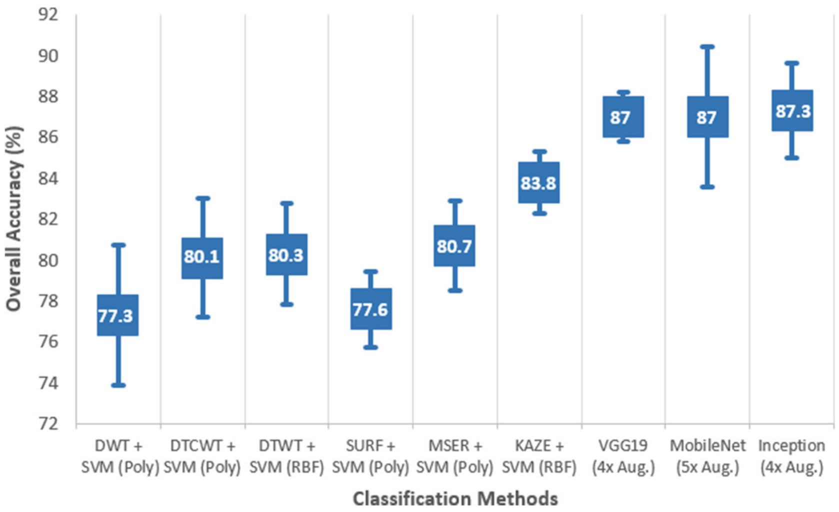


Fig. 13 The best obtained accuracies and the accuracy variations between each folds

of the maximum fold accuracy, MobileNet gave the highest classification accuracy for fold 4 with 90.4%. MobileNets is more sensitive to image noises due to being a small size of network when compared with VGG19 and InceptionV3. Even though the results obtained with MobileNet have more variation when compared with other deep learning models, MobileNet still outperforms other tested conventional machine learning methods, which makes it an optimum solution for sperm morphology analysis. The variation in MobileNet is measured as ±3.4%. In the conventional techniques, the wavelet-based features have more variations in results compared with descriptor-based features. KAZE features with SVM classifier is observed as the most consistent model within all conventional techniques (±1.43%).

4 Discussion

In infertility diagnosis, one of the important steps is the sperm morphology assessment. Traditionally, this assessment is performed visually and this results in a highly observer expertness dependent conditions. Therefore, in order to eliminate the human factor, computerised techniques should be utilised in the laboratories. In this study, we proposed fully automated sperm morphology analysis approach over the images obtained by a smartphone data acquisition technique.

In our approach, we initially segmented the ROI zones of the stained ocular images by employing two clustering methods. But these methods can be affected by the variation of the gray scale histogram. The light changes cause the incorrect segmentation. For this reason, we utilised MOGS technique. This technique isolates the large magnitude signal parts. In our application, sperm can be identified as large magnitude objects in the patches. In the MOGS denoised images, the visual quality of sperm morphology was improved and a cleaner representation of possible blobs (sperm or non-sperm) was obtained. Clustering methods resulted more detailed segmentation of sperm/non-sperm regions when they were applied as a subsequent step of MOGS. After automated pre-classification, each segmented sperm region was manually labelled as normal/abnormal for forming the ground truths of SMIDS. Then, we evaluated the classification techniques over the MOGS based automatically segmented and manually labelled patches as normal, abnormal and non-sperm. We tested conventional and ensemble machine learning techniques with wavelet- and descriptor-based feature extraction methods. Additionally, three convolutional neural networks in terms of deep learning idea were implemented over the SMIDS.

In the wavelet-based feature extraction analysis, which offers an alternative representation of an image to reveal additional information difficult to obtain in the original domain, DWT, DTWT and DTCWT forms were performed in different decomposition levels. It was observed that the highest classification accuracy values were obtained by using DTCWT due to their nearly shift invarional selectivity properties. Regarding to the number of decomposition levels, employing seven levels was found to be more informative feature extraction strategy for the sperm morphology in all wavelet forms.

Additionally, descriptor-based spatial feature extraction techniques such as SURF, MSER and KAZE were also tested on SMIDS with the same classifiers. In descriptor-based analysis, KAZE method, which builds a non-linear scale space in an iterative way to reduce noise while retaining the object boundary structure in order to obtain accurate positions of image keypoints, gave the most accurate results (83.82%). However, the computational cost is more than SURF and MSER due to being an iterative

Table 6 Result comparison with our preliminary studies

Data set size (No. images)	Ilhan et al. [39]		Ilhan et al. [41]	Proposed study
	DTCWT	DWT	SURF	MobileNET
Small (536)	82.33	78.15	83.39	86.58
Full (3000)	80.07	77.33	77.6	87

analysing approach. The minimum computational cost was observed in SURF which is a fast version of regular SIFT technique. The highest accuracy was obtained as 77.60% for SURF-based feature classification.

In the deep learning-based classification approaches, all the three employed CNNs obtained higher classification accuracy values when compared with the conventional feature extraction techniques. But, their extreme computational cost and training times are important criteria to be considered in deep learning-based systems. In this respect, MobileNet, which holds potential for substantial clinical impact as an application in mobile devices, is dedicated as the most useful network because of its minimum processing time with almost the same accuracy rates with other tested deep learning networks (87%). The main reason of being faster is the parameters reduced by organizing the standard convolution processes as spatial and channel based in the architecture. In order to provide the functionality of standard convolution process, depthwise (channel) and pointwise (spatial) convolutions are applied sequentially. In this way, the parameter size and calculation cost are minimised. [36].

As a fair comparison of the results with our preliminary studies [39, 41] which have been performed on a small size of SMIDS, all the relevant results are gathered in Table 6. The distribution of small data set and the implementation of DTCWT, DWT and SURF methods are explained in the corresponding papers in detail. In proposed study, the same configurations are tested over full size of SMIDS in order to test the generalisation capability of the relevant methods and it was seen that increasing the sample size has not a direct positive effect on the performance of conventional techniques due to the noise components appeared in stained images. However, the MobileNet approach benefits from this sample size increase with an even better performance when data augmentation is applied.

As the minor limitation of this study, it would be good to point out that the performance of the hybrid system is not directly measured on the raw patches obtained from the image acquisition system. The segmentation performance, which was obtained from the ocular images, and classification performance, which was obtained from SMIDS, are given separately for the following reasons: (i) to show the effect of the MOGS and the performance of Fuzzy *C*-means in the segmentation, (ii) to create a complete database which can be used publicly in sperm morphology studies, to do that the misses and the false positive samples obtained at the end of segmentation step were corrected and added to SMIDS, (iii) to find the best discriminating feature set for the sperm morphology problem, (iv) to decide the individual and ensemble learning methods that shows highest representative capability for normal/abnormal classification problem and (v) to find the optimum CNN model extracting the most informative high-level features.

5 Conclusion

In the light of the obtained results, we propose a hybrid sperm morphology analysis framework, which consists of a MOGS plus Fuzzy *C*-Means clustering approach in the segmentation part, and an end-to-end MobileNet-based learning approach in the classification part, as given in Fig. 2. The superiority of the proposed framework was validated by comparing its performance with the conventional features plus traditional individual/ensemble machine learning approaches and other two well-known CNNs named as InceptionV3 and VGG19 in terms of accuracy and computational complexity.

In further studies, we are planning to test our data with already trained networks in terms of ‘Transfer Learning’ concept [65] to accelerate the training procedure and to get benefits of reusing the pieces or modules of already developed robust models. In this scenario, if the transfer learning results in better performance, we would eliminate exhaustive and time-consuming training steps of deep learning. Additionally, we intend to test the proposed deep learning-based system on other previously published sperm morphology data sets such as SCIAN-Morpho and HuSHeM [22, 64].

Regarding to head-tail orientation information, which might be crucial in the performance of classification, a directional masking technique has been tested in a small data set (HuSHeM) and promising results have been obtained [37]. In a future study, we intend to apply this technique to SMIDS with the aim of obtaining better classification performance.

Eventually, our final objective is finding a common hybrid deep neural network-based system that can be employed in mobile devices with high discrimination performance that is proven over all publicly available sperm morphology data sets including the SMIDS. We also plan to implement a real-time working version of the proposed framework in an Android-based embedded system and test this system in clinical environment with the aim of fast and robust diagnosis.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed were in accordance and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards and ethical approval was obtained from Istanbul University, Faculty of Medicine.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M et al (2016) Tensorflow: large-scale machine learning on heterogeneous systems. arXiv preprint arXiv:1603.04467
2. Alegre E, Biehl M, Petkov N, Sanchez L (2013) Assessment of acrosome state in boar spermatozoa heads using n-contours descriptor and rlvq. Comput Methods Program Biomed 111(3):525–536
3. Alegre E, GonzáLez-Castro V, Alaiz-rodríguez R, GarcíA-OrdáS MT (2012) Texture and moments-based classification of the acrosome integrity of boar spermatozoa images. Comput Methods Program Biomed 108(2):873–881
4. Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. Nat Biotechnol 33:831–838
5. Alpaydin E (2014) Introduction to machine learning the. MIT Press, Cambridge
6. Amann RP, Waberski D (2014) Computer-assisted sperm analysis (casa): capabilities and potential developments. Theriogenology 81(1):5–17
7. Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7):878
8. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79
9. Bach F, Jenatton R, Mairal J, Obozinski G et al (2012) Optimization with sparsity-inducing penalties. Found Trends® Mach Learn 4(1):1–106
10. Bao P, Zhang L (2003) Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. IEEE Trans Med Imaging 22(9):1089–1099
11. Barroso G, Mercan R, Ozgur K, Morshedi M, Kolm P, Coetzee K, Kruger T, Oehninger S (1999) Intra-and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation, staining techniques and manual versus computerized analysis. Hum Reprod 14(8):2036–2040
12. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision. Springer, pp 404–417
13. Belsey M, Moghissi K, Eliasson R, Paulsen C, Gallegos A, Prasad M (1980) Laboratory manual for the examination of human semen and semen-cervical mucus interaction
14. Bijar A, Benavent AP, Mikaeili M et al (2012) Fully automatic identification and discrimination of sperms parts in microscopic images of stained human semen smear. J Biomed Sci Eng 5(07):384
15. Bijar A, Mikaeili M, Benavent AP, Khayati R (2012) Segmentation of sperm's acrosome, nucleus and mid-piece in microscopic images of stained human semen smear. In: 2012 8th international symposium on Communication systems, networks & digital signal processing (CSNDSP). IEEE, pp 1–6
16. Björndahl L, Barratt CL, Mortimer D, Jouannet P (2015) How to count sperm properly: checklist for acceptability of studies based on human semen analysis. Hum Reprod 31(2):227–232
17. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
18. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and brooks, Monterey
19. Cai W, Chen S, Zhang D (2007) Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recogn 40(3):825–838
20. Centola GM (2014) Semen assessment. Urol Clin 41(1):163–167
21. Chang V, Garcia A, Hitschfeld N, Härtel S (2017) Gold-standard for computer-assisted morphological sperm analysis. Comput Biol Med 83:143–150
22. Chang V, Saavedra JM, Castañeda V, Sarabia L, Hitschfeld N, Härtel S (2014) Gold-standard and improved framework for sperm head segmentation. Comput Methods Program Biomed 117(2):225–237
23. Chen PY, Selesnick IW (2013) Group-sparse signal denoising: non-convex regularization, convex optimization. arXiv:1308.5038
24. Chollet F et al (2015) Keras
25. Cui W (2010) Mother or nothing: the agony of infertility
26. DeLamater J, Plante RF (2015) Handbook of the sociology of sexualities. Springer, Berlin
27. Deng SW, Han JQ (2018) Adaptive overlapping-group sparse denoising for heart sound signals. Biomed Signal Process Control 40:49–57
28. Ding Y, He W, Chen B, Zi Y, Selesnick IW (2016) Detection of faults in rotating machinery using periodic time-frequency sparsity. J Sound Vib 382:357–378
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118
30. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
31. García-Olalla O, Alegre E, Fernández-Robles L, Malm P, Bengtsson E (2015) Acrosome integrity assessment of boar spermatozoa images using an early fusion of texture and contour descriptors. Comput Methods Programs Biomed 120(1):49–64
32. Ghosh S, Dubey SK (2013) Comparative analysis of k-means and fuzzy c-means algorithms. Int J Adv Comput Scie Appl 4(4):35–39
33. Gonzalez-Castro VGC, Alegre E, Morala-Arguello P, Suarez S (2009) A combined and intelligent new segmentation method for boar semen based on thresholding and watershed transform. Int J Imaging Robot 2(S09):70–80
34. Gupta S, Chauhan RC, Sexana SC (2004) Wavelet-based statistical approach for speckle reduction in medical ultrasound images. Med Biol Eng Comput 42(2):189–192

35. He W, Ding Y, Zi Y, Selesnick IW (2016) Sparsity-based algorithm for detecting faults in rotating machines. Mech Syst Signal Process 72:46–64
36. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
37. Ilhan H, Serbes G, Aydin N (2019) Automatic directional masking technique for better sperm morphology segmentation and classification analysis. Electron Lett 55(5):256–258
38. Ilhan HO, Aydin N (2018) A novel data acquisition and analyzing approach to spermiogram tests. Biomed Signal Process Control 41:129–139
39. Ilhan HO, Serbes G, Aydin N (2018) Dual tree complex wavelet transform based sperm abnormality classification. In: 2018 41St international conference on telecommunications and signal processing (TSP). IEEE, pp 1–5
40. Ilhan HO, Serbes G, Aydin N (2018) The effects of the modified overlapping group shrinkage technique on the sperm segmentation in the stained images. In: 2018 41St international conference on telecommunications and signal processing (TSP). IEEE, pp 1–4
41. Ilhan HO, Sigirci IO, Serbes G, Aydin N (2018) The effect of nonlinear wavelet transform based de-noising in sperm abnormality classification. In: 2018 3Rd international conference on computer science and engineering (UBMK). IEEE, pp 658–661
42. Jermyn M, Desroches J, Mercier J, Tremblay MA, St-Arnaud K, Guiot M, Petrecca K, Leblond F (2016) Neural networks improve brain cancer detection with raman spectroscopy in the presence of operating room light artifacts. J Biomed Opt 94002:21–9
43. Kabir MA, Shahnaz C (2012) Denoising of ecg signals based on noise reduction algorithms in emd and wavelet domains. Biomed Signal Process Control 7(5):481–489
44. Khachane MY, Manza R, Ramteke R (2015) Fuzzy rule based classification of human spermatozoa. In: 2015 international conference on Electrical, electronics, signals, communication and optimization (EESCO). IEEE, pp 1–5
45. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
46. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
47. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
48. Mallat S (1999) A wavelet tour of signal processing. Elsevier, Amsterdam
49. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 22(10):761–767
50. Meurant G (2012) Wavelets: a tutorial in theory and applications, vol 2. Academic press, Cambridge
51. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller MA, Fidjeland A, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518:529–533
52. Nafisi VR, Moradi MH, Nasr-Esfahani MH (2005) Sperm identification using elliptic model and tail detection. World Acad Sci Eng Technol 6:205–208
53. Nayak DR, Dash R, Majhi B (2016) Brain mr image classification using two-dimensional discrete wavelet transform and adaboost with random forests. Neurocomputing 177:188–197
54. Organization WH et al (2010) Who laboratory manual for the examination and processing of human semen
55. Panda S, Sahu S, Jena P, Chattopadhyay S (2012) Comparing fuzzy-c means and k-means clustering techniques: a comprehensive study. In: Advances in computer science, engineering and applications. Springer, pp 451–460
56. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV) 115(3):211–252
57. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
58. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H (2019) A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. Appl Soft Comput 74:255–263
59. Schölkopf B, Smola AJ, Bach F et al (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
60. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2010) Rusboost: a hybrid approach to alleviating class imbalance. Trans Sys Man Cyber Part A 40(1):185–197
61. Selesnick IW, Baraniuk RG, Kingsbury NC (2005) The dual-tree complex wavelet transform. IEEE Signal Process Mag 22(6):123–151
62. Serbes G, Aydin N (2014) Denoising performance of modified dual-tree complex wavelet transform for processing quadrature embolic doppler signals. Med Biol Eng Comput 52(1):29–43
63. Serbes G, Sakar BE, Gulcur HO, Aydin N (2015) An emboli detection system based on dual tree complex wavelet transform and ensemble learning. Appl Soft Comput 37:87–94
64. Shaker F, Monadjemi SA, Alirezaie J, Naghsh-Nilchi AR (2017) A dictionary learning approach for human sperm heads classification. Comput Biol Med 91:181–190
65. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298
66. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484
67. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
68. Strang G, Nguyen T (1996) Wavelets and filter banks. SIAM, Thailand
69. Sze V, Chen YH, Yang TJ, Emer JS (2017) Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE 105(12):2295–2329
70. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
71. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
72. Ulukaya S, Serbes G, Kahya YP (2019) Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. Comput Biol Med 104:175–182
73. Wang C, Leung A, Tsoi WL, Leung J, Ng V, Lee KF, Chan SY (1991) Computer-assisted assessment of human sperm

morphology: comparison with visual assessment. Fertility Sterility 55(5):983–988

74. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH (2016) Deep learning for identifying metastatic breast cancer. CoRR arXiv:1606.05718
75. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR et al (2015) The human splicing code reveals new insights into the genetic determinants of disease. Science 347(6218):1254806
76. Yang MS, Hu YJ, Lin KCR, Lin CCL (2002) Segmentation techniques for tissue differentiation in mri of ophthalmology using fuzzy clustering algorithms. Magn Reson Imaging 20(2):173–179
77. Zeng H, Edwards MD, Liu G, Gifford DK (2016) Convolutional neural network architectures for predicting dna–protein binding. Bioinformatics 32(12):i121–i127
78. Zhang Y, Dong Z, Wu L, Wang S (2011) A hybrid method for mri brain image classification. Expert Syst Appl 38(8):10049–10053
79. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods 12:931–934

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Hamza O. Ilhan is research assistant in YTU. His research interests are in the areas of image and signal processing, machine learning and pattern recognition with applications to biomedical engineering.

I. Onur Sigirci is a PhD student at the Computer Engineering at the Yildiz Technical University (YTU). His main research interests are hyperspectral and biomedical image processing, machine learning and algorithms.

Gorkem Serbes has experience in biomedical engineering, digital signal processing and pattern recognition for more than 10 years, and he has authored over 40 peer-reviewed manuscripts in these fields.

Nizamettin Aydin is now with Computer Engineering Department and serves as the head of the department at YTU. He has experience in biomedical engineering, digital signal processing and pattern recognition.

Nizamettin Aydin received the B.Sc. (1984) and M.Sc. (1987) degrees in Electronics and Communication Engineering from Yildiz Technical University, Turkey, and Ph.D. in Medical Physics (1994) from the University of Leicester, UK. He worked in Electronics Engineering Department, Gebze Institute of Technology, Turkey, from 1995 to 1999. He also worked in the Department of Clinical Neurosciences at Kings College London and the Division of Clinical Neuroscience at St George's Hospital Medical School as a Research Fellow between 1998 and 2001. He was a Senior Research Fellow in the Institute for Integrated Micro and Nano Systems, School of Engineering and Electronics at the University of Edinburgh from 2001 to 2004. He is now a Professor of Biomedical Engineering and Computing in the Computer Engineering Department, Yildiz Technical University, Turkey. He was awarded the IEE the Institute Premium Award for 2000/2001. His research interest s include time-frequency and time-scale analysis, physiological measurements, biomedical signal processing and computing, AI, data science and analytics, VLSI, bioinformatics, digital communications, and software engineering.