DIAGNOSING OSTEOPOROSIS FROM TAPPING RESPONSES USING MACHINE LEARNING

Francis F. Li, Olga Umnova, Jamie Scanlan, Gyorgy Rakoczy

Acoustics Research Centre, University of Salford, UK
emails: f.f.li@salford.ac.uk, o.umnova@salford.ac.uk, j.scanlan1@edu.salford.ac.uk,
george.rakoczy04@gmail.com

Nóra Lövey

Warrington Hospital, Warrington, Cheshire, UK
e-mail: n.love@nhs.net

Osteoporosis is a prevalent but asymptomatic condition that affects a large number of people in elderly population, causing bone fragility and resulting further in a significantly increased risk of fracture. Several methods have been developed and are currently available in specialist departments of general hospitals to assess the bone quality and diagnose osteoporosis; however these are accessed via referrals of general practitioners, not broadly available for screening testing in primary healthcare settings such as general practitioner or family doctors' clinics for a number of logistic and cost constraints. This paper describes a new method that uses a medical grade reflex hammer to exert testing stimuli, i.e. tapping, an electronic stethoscope to pick up impulse responses of tibia, and intelligent signal processing based on machine learning to determine the likelihood of osteoporosis.

Keywords: resonant frequencies, osteoporosis, machine learning, classification, stethoscope.

1. Introduction

Bone continuously undergoes a dynamic regeneration process, in which new collagen and mineral are added and old ones removed for remodeling. As the skeleton develops more bone is added than is being taken away. The process eventually stabilizes and the bone density remains constant. If however, more bone is removed than added, a condition called osteoporosis (OP) occurs. Osteoporosis literally means ‘porous bone’ and describes a period of largely asymptomatic bone loss leading to skeletal fragility and increased risk of fracture. One in three women and one in five men over the age of 50 will break a bone attributed to osteoporosis according to some well-known public surveys by the International Osteoporosis Foundation [1]. Osteoporosis affects more than 200 million people worldwide, being the cause of more than 8.9 million fractures every year [2,3]. It is well established that early diagnosis and treatments are the key to prevent further complications and fractures. The lack of a simple and practical diagnostic method for screening has been identified as a cause of delayed diagnosis, poor prognosis, increased risk of fracture and subsequent complications. The commonly used diagnostic method for OP is dual energy X-ray absorptiometry (DEXA/DXA), which measures the bone mineral
density (BMD). The World Health Organization (WHO) recommended the use of a t-score of the BMD as a diagnostic criterion \[4, 5\]. The t-score is a comparison of one’s BMD with that of a young and healthy population and is reported in standard deviations (SD). A t-score below -2.5 SD is deemed as being osteoporotic. However the BMD is not a reliable measure of the bone’s strength and can only imply bone quality \[4, 6, 7\]. Even so, DEXA scans are not readily accessible in general practitioners clinics for screening testing but are only available in general hospitals via primary care professionals’ referrals and often have long waiting list times.

In everyday life, one often taps on a structure or material to determine its solidity. Similarly, percussion or tapping techniques are used by doctors in clinical examinations to determine density or cavity and assess certain conditions of the thorax and abdomen \[8\]. They can also be used for assessing conditions in other parts of the body. Percussion sound transmitted through bones listened through the chest using a stethoscope was reported to be able to detect osteoporosis \[9\]. A close correlation between bone resonant frequencies and the BMD was confirmed recently \[10\]. The lowest resonant frequency of tibia and other physiological information were mapped onto the ‘FRAX’ algorithm to give a diagnosis of osteoporosis \[11\]. More recently the authors proposed a machine learning method to differentiate vibro-acoustic signals, and detect osteoporosis \[12, 13\].

This paper summarizes our works and suggests some further developments. The proposed method and setup are illustrated in Figure 1: A clinician taps on a patient’s proximal tibia bone with a medical grade reflex hammer and an electronic stethoscope picks up the sound at the midpoint and/or the distal end of the tibia; the signal is transmitted via a Bluetooth data-link to a computer for signal processing and pattern recognition, leading to a diagnostic decision. The hypothesis is that the bone’s bending stiffness, mass, and densities can be interpreted from its resonant frequencies with some necessary assumptions. The decision to suggest that a patient might have osteoporosis is based on machine learning. The algorithm looks for common features in the time-frequency domain from a good number of acoustical examples of normal and osteoporotic subjects, then generalizes the knowledge to cases not presented in the training for decision making.

![Figure 1: Illustration of the proposed method](image)

2. **Rationale and mechanisms**

Much research done in physiology indicated that the lowest resonant frequency is closely related to the bending stiffness of a long bone, hence the quality of the bone \[14,15\]. This is in line with a simple rod model:

\[
fo = \sqrt{\frac{k}{m}}
\]

where \(k\) (N/m) is the stiffness of the rod, \(m\) (kg) is the mass and \(fo\) (Hz) is the lowest resonant frequency. Therefore, an isotropic rod might be used as a coarse but intuitive model for a long bone. Reduced stiffness lowers the resonant frequency, but lessened mass results in an increased resonant frequency. Osteoporosis reduces bone stiffness; it also removes mass from the bone \[16\]. Such a decrease in mass
will counter-act on the lowering of the resonant frequency. However, the pores formed in long bones due to bone losses are filled with fat and other body substances, the mass reduction is not as significant as forming cavities. Much of the research indicated that lowered resonant frequencies and shifted modal frequency distributions are associated with osteoporosis, though the strict proportional relation between the resonant frequency and the square root of stiffness does not generally hold [16-21]. Clear correlation relationships between resonant frequencies of long bones and whole body BMDs are also evident [10]. There are other efforts and paradigms of using vibro-acoustic techniques in vivo to detect osteoporosis and fracture, e.g. [9, 10, 22].

Bone has a complex anisotropic structure, which is made of two main layers: a surface layer of calcium and an internal network formed by trabeculae: a porous mass of micro-rods and branches. The trabeculae are arranged in cross-connected plates along the bone to give strength in typical load directions [23]. The complexity and diversity in the bony structures of individuals make accurate mathematical modeling and analytical solutions extremely difficult. Machine learning seems to be a sensible solution for this type of complex problems, if a large number of examples and reliable “teacher” values are available.

Impulse responses to reflect hammer percussion acquired in vivo have artifacts: the reflex hammer has a semi-rigid rubber head and the soft-tissue layers introduce nonlinear damping. These artifacts vary depending on many factors such as the force and velocity of tapping and the thickness of soft tissue layer. Feature extraction methods that can de-convolve complex signals might be beneficial. Machine learning is expected to learn from a large number of examples to disregard these varying artifacts found in signals. As a block-box approach to complicated mapping problems, where analytical methods are difficult to apply but examples are available, machine learning with neural networks found itself useful here; and in applying such a method, generalization capability of the model is validated via testing with examples not included in training.

3. Methods

3.1 Source and transducer

A typical method to study structure borne sound is to use an impact hammer to generate an impulsive stimulus and accelerometers at various receiving positions to acquire the impulse responses. Previous research into the long bones resonances used exclusively similar methods. Accelerometers were used to pick up signals. In in vitro studies, impact hammers were used; in in vivo cases a pendulum with a 20 g glass ball was used to exert impact forces [10]. To facilitate data acquisition in a clinical setting, the use of reflex hammer as an electronic-stethoscope was first proposed by the authors [12, 13].

An electronic stethoscope with Bluetooth data-link (3M Littmann model 3200) was used in this study. A side by side comparison of the electronic stethoscope and an accelerometer reveals that the stethoscope has nontrivial resonances in the region of 10 to 35 Hz, a generally flat frequency response from about 50 Hz to 600Hz and then a circa -18 dB/Oct rolling off in higher frequency extension to 1.2 kHz. The lowest resonant frequency of tibia is around 100 Hz and major resonant frequencies of interest are within several hundred Hz in the current study, as resonances above 1 kHz are barely detectable in vivo due to soft-tissue damping. The electronic stethoscope seems adequate for the purpose and if preferred, the frequency response can be equalized.

3.2 Dataset

Recordings of 110 patients were made by one of the authors (Rakoczy) in 2016 & 2018 following the method described in Figure 1. Ten tapping responses were recorded at mid-point and distal ends of tibia in each patients and controls. A summary of the patient statistics is in given in Table 1, in which OP means osteoporosis and OK means normal. These diagnostic decisions were made by experienced
human doctors based on DEXA scores and other relevant information of the patients (e.g. gender, age, body mass and other health conditions) and further checked and confirmed by a consultant rheumatologist (one of the authors, Lövey). The impulse responses acquired at the mid-point of tibia are used in the current study.

<table>
<thead>
<tr>
<th>Table 1. Population Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>M: 19; W: 90;</td>
</tr>
<tr>
<td><strong>DXA</strong></td>
</tr>
<tr>
<td><strong>No. OP Cases</strong></td>
</tr>
<tr>
<td><strong>No. OK Cases</strong></td>
</tr>
</tbody>
</table>

### 3.3 Signal pre-processing

Pre-processing is performed to isolate and align the individual impulse responses from a series of 10 in each recording. This is done according to the gradients of RMS envelopes. An “error checking” stage is used to discard “bad” ones mostly due to inconsistency in percussion. A variety of feature spaces have been experimented with. It is found that data representation in cepstrum domains generally outperforms others in terms of detection sensitivity and specificity. It also facilitates training, evidenced by faster convergence. This finding not surprising, as cepstrum domain representation makes unwanted and uncertain components in signals through a convolution process easier to be removed or averaged out in the subsequent machine learning stage. A non-uniformed frequency sampling with more emphasis on lower end than higher end of the spectrum is found computationally efficient. Taking these into account and as a convenient choice (not necessarily the optimal one), a time-frequency domain representations, Mel Frequency Cepstru Coefficients (MFCCs), are used to capture the features and resonance patterns. Each set of the MFCCs has 21 coefficients over 18 windows. This gives 378 coefficients arranged as a column vector to feed into the artificial neural network. Table 2 gives the detail of window length and overlap when calculating the MFCCs.

<table>
<thead>
<tr>
<th>Table 2: Dataset parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>MFCC Window Length</td>
</tr>
<tr>
<td>MFCC Window Overlap</td>
</tr>
<tr>
<td>MFCC Total</td>
</tr>
</tbody>
</table>

### 3.3 Machine Learning Methods

Artificial Neural Networks (ANNs) are made from a large number of interconnected neuron models, which empowers a learning capability [24]. The purpose of apply ANNs here is to train the ANN on percussion sounds and the human doctor’s diagnosis so that the trained system can generalize to cases not included in the training. The ANN used in this study are built around neuron models with a linear basis function and a continuous sigmoid function is used in this study.

\[
a_i = \frac{1}{1 + e^{-u_i}}
\]  

(2)
where \( a_i \) is the output of the \( i \)th neuron, and \( u_i \) is the weighted sum of the inputs into that neuron.

A multi-layered feed-forward ANN was found adequate. The input layer takes feature vector and distributes data to the hidden layer(s). The hidden layers reduce the data and find solution for the desirable mapping relationships. The output layer then takes the outputs of the last hidden layer and sums them for a final result. The network was built with 378 neurons for the input layer, 120 for one hidden layer and 1 output neuron for the early proof-of-concept study using only part of the data (with only 12 cases). For larger datasets the ANN was modified to include a second hidden layer of 40 neurons.

The ANN starts with random weights and the MFCCs are passed through the ANN. The output of the network is compared with the doctor’s diagnosis (“teacher value”) and the error between them is calculated. The teacher values are: 0 for osteoporotic (OP) patients; 1 for healthy (OK) patients. The teacher values are informed by senior doctors’ diagnoses of the patient in question, which are generally in line with the aforementioned WHO guidelines i.e. t-scores below -2.5 as osteoporotic. But some other conditions and aspects are included in the diagnosis such as lifestyle and family history, which becomes part of the teacher value. The aim of training is to minimize the total square error \( E \) as defined in Equation 7 over all training examples:

\[
E = \frac{1}{2} \sum_{m=1}^{M} [o^{(m)} - t^{(m)}]^2 \tag{3}
\]

where \( o \) is the ANN output \( t \) is teacher value \( m \) is example number.

This error is then used in the back-propagation algorithm which updates the weights for each neuron [24]. This process means that the most important or distinctive data points in each layer are weighted greater in the sums and decisions than other points, reducing the data required to represent the information.

3.4. Additional Classification & Voting Stage

The ANN was designed to have a continuous output for reasons. While the general aim is to classify OP and OK through a binary classifier. The WHO classification of osteopenia means there is borderline conditions. Therefore the output of the ANN is passed through a post-hoc stage which uses a threshold to make decisions. This is shown in Figure 2. The case is deemed as being OK or OP if its error from the teacher value is within 0.4. If the error is substantial then it is classified as a false positive or negative. The advantage of having this threshold stage is that the thresholds can be adjusted the trade-off between detection sensitivity and specificity.

![Figure 2: Error threshold diagram. Green indicates the catchment areas for classifying OK or OP. Yellow indicates excluded ambiguous results.](image)

4. Results and Discussion

Training and validation tests were experimented with using various pre-processing algorithms, with a variety of ANN sizes and training strategies. First effort was made with a small dataset of only 12 patients, and later the dataset has gradually enlarged to include more recordings. With the latest
dataset, as discussed earlier, including 110 patients, validation testing shows a maximum error rate of circa 30%. This slightly varies with each new start of the ANN and the step size used in training.

To evaluate the clinical usefulness of the method, it is important to find what specificity and sensitivity the method can offer. Specificity describes how accurate the algorithm is in detecting OK patients while reducing false negatives. Sensitivity describes the opposite: how accurate the algorithm is in detecting the OP subjects while reducing false positives. To understand these, training and validation with a small dataset (12 patients) were further looked into to explore the decision making behaviour, with a range of learning rates and 3 practical stop criteria.

1) Training terminates when the total squared error falls below 0.05. Figure 3 a) and b) give some of the results for illustration. It is observed in Figure 3 b) that there is a period of rapid oscillation in overall error. This indicates the likelihood of missing some valuable local minima, leading to poor results even at a late stage due to a large learning rate.

(2) Training terminates when the algorithm reaches 80% accuracy for both the OK and OP cases. This is presented in Figure 4 a) and b).

For a diagnostic system the possibility of false negatives is a concern. It is observed in Figure 4 a) that for small learning rates, the false negative counts are low in the initial stages (0 - 2 seconds run time) but once the algorithm starts to balance to reduce the false positives, the false negative rate rises. At high learning rates such as 0.5, the network rebalances at the expense of increasing false negatives. However this does not affect the patient results, which still shows the osteoporotic patients being diagnosed as such.
(3) When the algorithm was able to correctly identify all the patients that were OK and OP. Figure 5 a) and b) give examples of this. Overall, as the training error decreases, the number of patients correctly identified increases, while the number of IR correctly identified converges to stable values.

![Figure 5. a) Patient sensitivity b) sensitivity to all 35 impulse responses.](image)

These finding suggested an empirical or systematic method is needed to thoroughly fine turn the machine learning algorithm to achieve its optimal performance.

5. Concluding remarks

A new method suitable for screening testing of osteoporosis has been proposed and validated with a reasonable sized dataset. Using common clinical apparatus, a reflex hammer and an electronic-stethoscope, a piece of computer software analyzes the percussion sounds and makes diagnostic decisions. Further refinement to the algorithm and data representation can potentially improve the specificity and sensitivity of diagnosis. Full scale clinical validations side by side with other diagnostic methods are suggested to fully establish the method.

REFERENCES


