HRIDAI: A Tale of Two Categories of ECGs

Priya Ranjan[1(✉)], Kumar Dron Shrivastav[4], Satya Vadlamani[2], and Rajiv Janardhanan[3]

[1] SRM University, Neerukonda, Mangalagiri Mandal, Guntur District, Mangalagiri 522502, Andhra Pradesh, India
ranjan.p@srmap.edu.in

[2] Laboratory of Disease Dynamics and Molecular Epidemiology, Amity Institute of Public Health, Amity University Uttar Pradesh, Sector 125, Noida, India
vadlamani.satya93@gmail.com

[3] Laboratory of Disease Dynamics and Molecular Epidemiology, Health Data Analytics and Visualization Environment, Amity Institute of Public Health, Amity University Uttar Pradesh, Sector 125, Noida, India
rjanardhanan@amity.edu

[4] Health Data Analytics and Visualization Environment, Amity Institute of Public Health, Amity University Uttar Pradesh, Sector 125, Noida, India
kdshrivastav@amity.edu

Abstract. This work presents a geometric study of computational disease tagging of ECGs problems. Using ideas like Earthmover's distance (EMD) and Euclidean distance, it clusters category 1 and category −1 ECGs in two clusters, computes their average and then predicts the category of 100 test ECGs, if they belong to category 1 or category −1. We report 80% success rate using Euclidean distance at the cost of intense computation investment and 69% success using EMD. We suggest further ways to augment and enhance this automated classification scheme using bio-markers like Troponin isoforms, CKMB, BNP. Future directions include study of larger sets of ECGs from diverse populations and collected from a heterogeneous mix of patients with different CVD conditions. Further we advocate the robustness of this programmatic approach as compared to deep learning kind of schemes which are amenable to dynamic instabilities. This work is a part of our ongoing framework Heart Regulated Intelligent Decision Assisted Information (HRIDAI) system.

Keywords: ECG · Computational disease tagging · Biomarkers · ECG-visualization

1 Introduction

Cardiovascular diseases (CVD) account for 17.9 million (31%) deaths each year worldwide. More than 75% of CVD deaths occur in low- to middle-income coun-

S. M. Thampi et al. (Eds.): SIRS 2020, CCIS 1365, pp. 1–21, 2021.
https://doi.org/10.1007/978-981-16-0425-6_18

tries (LMICs). India has seen a significant rise in CVD-associated mortality rates with an epidemiological transition to Noncommunicable Diseases. This pattern is uniform throughout the country despite wide variation in risk factors and socioeconomic status. India faces a great challenge in providing quality healthcare especially in rural domains due to lack of resources and trained healthcare providers. The lack of resources for triaging or stratification of patients based on the severity of their condition leads to prolonged waiting period for treatment further worsening the prognosis of the patients. Moreover, the scarcity of specialized cardiologists significantly impacts the clinical prognosis of the patients intensifying cardiovascular disease burden. The electrocardiogram (ECG) is a fundamental tool in the everyday practice of clinical medicine, with more than 300 million ECGs obtained annually worldwide. The ECG is pivotal for diagnosing a wide spectrum of cardiovascular abnormalities ranging from Arrhythmias to myocardial infarction (MI). The hospital based registries have neither been able to provide accurate estimates of the cardiovascular disease (CVD) burden nor identify the disease drivers for the CVD epidemic despite it being the largest cause of mortality. This necessitates the need to develop novel bottoms-up strategies to map out the burden of CVDs at community level as well. Such a strategy would not only help us to elucidate the niche specific disease drivers but also augment the hospital based registries in visualizing the realistic burden of CVDs across the Indian sub-continent. Although a large number of drugs have been designed for treating patients afflicted with CVDs across the Indian subcontinent, absence of a systematic database taking into account the vast genetic base of the Indian populace has been perhaps the Achilles heel in developing policies or programs aimed for better management of CVDs. Development of novel Artificial Intelligence (AI) enabled Electrocardiogram (ECG) interpretation has become increasingly important in the clinical ECG workflow since its inception over 50 years ago, serving as a crucial adjunct to physician interpretation not only in resource limited clinical settings prevalent across the Indian sub-continent but also elsewhere across the world. The availability of affordable, accessible, and scalable computational platforms with capabilities to process large-scale raw data will not only improve expert human ECG interpretation by accurately triaging or prioritizing the most urgent conditions but also importantly reduce the rates of misdiagnosed ECG interpretations. To this end, we are proposing a novel R-based open source software with inherent capability to classify different kinds of automated geometric visualizations of ECGs along with its categorization based upon similarity indices as measured by Earth Movers Distance (EMD). We anticipate that integration of this robust automated classifier along with minimally invasive detection of molecular biomarkers [14] such as N-terminal prohormone of brain natriuretic peptide (NTpro-BNP), Creatinine Phospho Kinase Muscle-Brain (CPKM/B), and Troponin isoforms will form the rationale for development of effective and precision oriented triage system for achieving not only high screening rates for Myocardial Infarction (MI) but also accurately triaging or prioritizing the most urgent conditions. Biomarkers are emerging as a new technique to find precursors of many diseases including

Epilepsy [18]. The rest of the paper is organised as follows. Section 2 contains the global and national perspective on the CVDs and related literature review. Section 3 contains the motivation for this work. Section 4 outlines the methodology which is Earthmover's distance (EMD) computation in our case. Section 5 contains the results from numerical study of ECG training and test dataset. Section 6 collects the conclusions, outlines future directions and discusses them in ongoing COVID-19 context. Finally in Sect. 7, practical aspects of this proposed algorithm and its possible implementation on an ECG device with applications to clinical practice is described.

2 Global and National Perspective

ECG captures the cardiac electrical activity from the body surface and is a cheap, non-invasive modality used as a basic cardiac diagnostic tool. ECG analysis is therefore a crucial first step in diagnosing, understanding and predicting cardiovascular aberrations representing 31% deaths globally in 2016 [4]. The premise of ECG remains the same, while the testing and analytical systems are evolving with the technological advancements incorporating increased mobility, ease of use, streamlined workflow and interoperability so as to interface data easily with electronic health/medical records (EHR/EMRs). Digitization of medical data and powerful computing platforms have laid the foundation for use of AI in revolutionizing healthcare innovation to optimize the processes and decisions.

Globally, ECG interpretation algorithms have slowly become the standard on many systems for automation of diagnosis and anatomical detection. These algorithms use gender and age-specific criteria to provide a virtual analysis for resting ECG interpretation, including detection of various cardiovascular diseases (NSTEMI, STEMI, Ventricular dysfunction, etc.) and diagnostic aids to provide interpretation of rhythm and morphology for a variety of patient populations and an effective system for triaging patients to prioritize the most severe conditions.

A major drawback of the existing ECG interpretation algorithms based on Artificial Neural Networks (ANN) is its black box nature. When you feed an image into a neural network and it matches it to another image, it is difficult to understand why and how it came up with a particular match [9–11]. Further, the nonlinear dynamical behavior of deep neural networks are prone to chaotic nature and fundamental underlying unpredictability [12]. Adding to this issue is extreme demand for computation in deep learning which is forcing researchers to explore other techniques [8,17]. On the other hand, static and predictable algorithms like, Earth Movers Distance (EMDs) perform image match by computing perceptual similarity and provide more meaningful and interpretable solutions to matching problems. To re-emphasize, ANNs have a long and very well researched history of inherent instability and its automated decisions can not be entrusted to make decisions critical to the survival of a patient afflicted with a severe case of a cardiac episode. India faces many challenges in delivering healthcare, especially in the rural domain due to shortage of resources and skilled personnel. The fact that CVD associated deaths account for 26% of mortalities in India clearly undermines the magnitude of the Public Health Challenge it imposes. This is further

compounded by the lack of suitable algorithms to map out the clinical resource allocation especially in rural underserved population where doctor-patient ratio is low and access to affordable and quality healthcare options is significantly marginalized. The 12 lead ECG is still most relied upon modality for diagnosing cardiac conditions despite the limitations of extant algorithms in vogue to accurately triage the patients in the order of urgency. A more prudent step therefore, will be to develop tools and applications that can be easily integrated into the current healthcare system. Additionally, the lack of well-structured databases for referencing and analysis hinders the progression of research from aiding and optimizing processes and clinical decision making with the help of AI. India endowed with diverse genetic base and socio-cultural norms, presents a unique landscape of disease burden necessitating the need for niche specific databases for enhancing the accuracy of AI tools. Apart from the above-mentioned complexities, the development of these technologies come with substantial technical, ethical, confidentiality and clinical challenges. In spite of the afore-mentioned hindrances, the socioeconomic impact and benefits of AI based automation of ECG analysis for LMICs like India will be significant. The potential healthcare application of AI-based platforms are vast, encompassing screening, disease detection, patient risk stratification along with niche specific optimal intervention strategies. Since the penetration of the mobile platforms and Internet is extensive across the Indian sub-continent, development of AI enabled computational applications on these platforms would provide a valuable, precision public health tool for better management of CVD epidemic alleviating a significant burden on the national exchequer.

3 Motivation for Building This Tool

This work is motivated by following two objectives.

1. Development and validation of an Intelligent Decision Support System for segregating ECG traces to detect CVD anomalies such as Myocardial Infarction (MI) in both tertiary care settings and extended community along with tracking of patients through low end mobile health applications.
2. Integration and validation of multi-modal tool in clinical practice involving automated processing of anonymized ECG-traces along with conventional molecular biomarkers [14] of cardiovascular diseases such as MI forming the rationale of effective triage methods for prioritizing the most urgent conditions to wait listed ones.

The race and sex-specific variations in the levels of conventional biomarkers such as Troponin indeed necessitate the validation and confirmation by a modality, which can crunch a large amount of data in an affordable and accessible manner. Our fruganomic AI-enabled tool will not only facilitate the same by incorporating the clinical-epidemiological features of the subjects evaluated at both tertiary care centers but also in the extended community. This computational modality, when integrated with digital signals from surrogate molecular markers, will form

the rationale of a multi-modal multi fusion sensor technology [19–21] which will aim at not only resolving the dogma of missed and misdiagnosis of CVD such as Myocardial Infarction at tertiary care centers and extended community but also individualize the risk assessment of patients with suspected myocardial infarction or to categorize patients into low- or high-risk groups.

We are further looking at directions to integrate signals from cardiac auscultation [13] and computational fluid dynamics [15,16] that will hopefully resolve the engineering dogma of delineating the ECG traces falling under the category of "mildly normal" and "mildly abnormal". This when integrated with clinical-epidemiological features of the subjects would significantly help in identification and delineation of modifiable vascular risk factors (particularly hypertension and smoking) along with social and environmental factors, including reduced exposure to air pollution through provisioning of Health Advisories and Literacy modules thereby empowering the vulnerable population to risks of even COVID-19 pandemic, whose clinical outcomes are extremely poor in patients with CVDs or risk factors of CVDs.

Taken together, evidently there is a vast potential of automated and computation based techniques in the detection of CVDs. Using ideas like the Earths Movers Distance (EMD) [22], visibility graphs [7] as shown in Fig. 13, we hope to further the state of art technologies in the detection of CVDs at the community level to understand the patterns and processes of CVDs in resource-limited healthcare systems in Low and Middle-Income Countries such as India. This will necessarily help in a realistic evaluation of the cardiovascular health of the Indian populace endowed with wide genetic and divergent socio-cultural norms as well as lifestyle choices.

4 Earthmover's Distance (EMD)

Earthmover's Distance (EMD) is a method to calculate the disparity between two multi-dimensional distribution in some space where a distance magnitude between single ones (ground distance) is given. Suppose the two distributions are there, one can be considered as the area with the mass of earth, and the other as a collection of holes in that same area. Then, the EMD is the measure of the least amount of work required to fill the holes with earth. Here the unit of work is the force needed in transporting unit earth by a unit of ground distance. So it can also be defined as the minimum cost that must be provided to convert one histogram into other. Measuring of EMD is based on a solution of *transportation problem* [1]. For finding mathematical representation, firstly we formalised it as the following linear programming problem:

Let X be the first signature with n clusters, x_i is the cluster representative, and w_{xi} is the weight of cluster.

Let Y be the second signature with m clusters, y_i is the cluster representative, and w_{yi} is the weight of cluster.

Let D be the ground distance matrix, d_{ij} is the ground distance between clusters x_i and y_j.

Let F be the flow matrix and f_{ij} is the between x_i and y_j.
Then,

$$X = \{(x_1, w_{x1}), (x_2, w_{x2}), (x_3, w_{x3}), ...(x_n, w_{xn})\} \tag{1}$$

$$Y = \{(y_1, w_{y1}), (y_2, w_{y2}), (y_3, w_{y3}), ...(y_m, w_{ym})\} \tag{2}$$

$$D = [d_{ij}] \tag{3}$$

$$F = [f_{ij}] \tag{4}$$

Now, the WORK (X, Y, F) = $\sum_{i=1}^{n} \sum_{i=1}^{m} f_{ij} d_{ij}$
Subject to constraints:

1. (i) $f_{ij} \geq 0$, where $0 \leq i \leq n$, $0 \leq j \leq m$
2. (ii) $\sum_{j=1}^{m} f_{ij} \leq w_{xi}$, where $0 \leq i \leq n$
3. (iii) $\sum_{j=1}^{m} f_{ij} \leq w_{yj}$, where $0 \leq j \leq m$
4. (iv) $\sum_{i=1}^{n} \sum_{i=1}^{m} f_{ij} = \min \sum_{i=1}^{n} w_{xi} \cdot \sum_{j=1}^{m} w_{yi}$

The constraint (i) enables mass moving from X to Y. (ii) and (iii) restricts the amount of mass that can be sent by the clusters in X to their weights and the clusters in Y to receive no more mass than their weights. (iv) one forces to move the maximum amount of mass possible. It is also known as the total flow. Once we solve the transportation problem, we will get the optimal flow F. Now the Earth Mover's Distance is defined as the work normalised by the total flow:

$$EMD(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} d_{ij} \div \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \tag{5}$$

5 Numerical Study of ECG Training and Test Dataset

5.1 Dataset

This dataset was formatted by R. Olszewski as part of his PhD thesis (see [1]). Each series traces the electrical activity recorded during one heartbeat. The two classes are a normal heartbeat verus a myocardial infarction event (heart attack due to prolonged cardiac ischemia). Train size: 100 Test size: 100 Missing value: No Number of classes: 2 Time series length: 96 This data has been donated by Robert Olszewski (see [2,3]). Just to first visualize this data, we plot two representative ECG time-series in Fig. 1, a two dimensional view in Fig. 2, three dimensional view in Fig. 6 and averages of ECGs in both categories in Fig. 3.

5.2 EMD Based Analysis of Training Data

We start implementation of EMD on training data first. Original training file has 100 ECGs in two categories, one is healthy and another is unhealthy with MI. We separate these two categories in two different spread sheets. Now major questions is that there are wide variations among healthy ECGs and unhealthy ECGs and a research question is which is ECG in the worlds of normal ECGs is representative

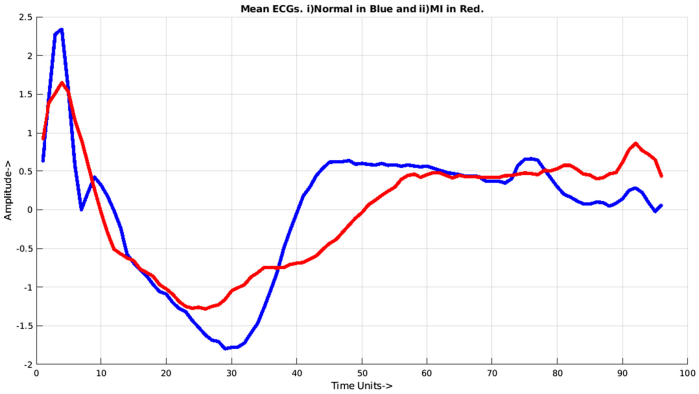


Fig. 1. Two representative ECGs from normal and diseased category

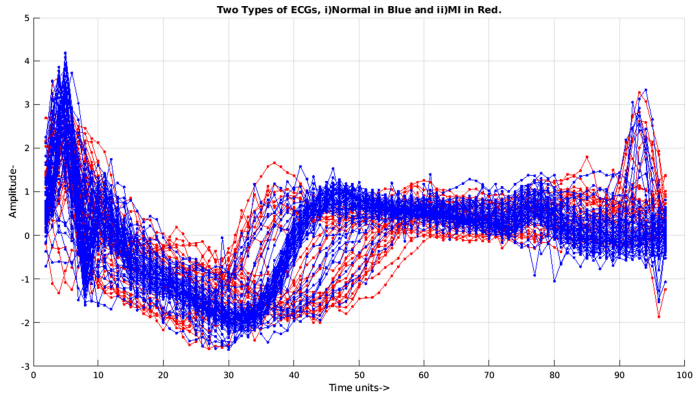


Fig. 2. Two dimensional view of different ECGs (healthy-blue, diseased-red) as temporal waveforms. (Color figure online)

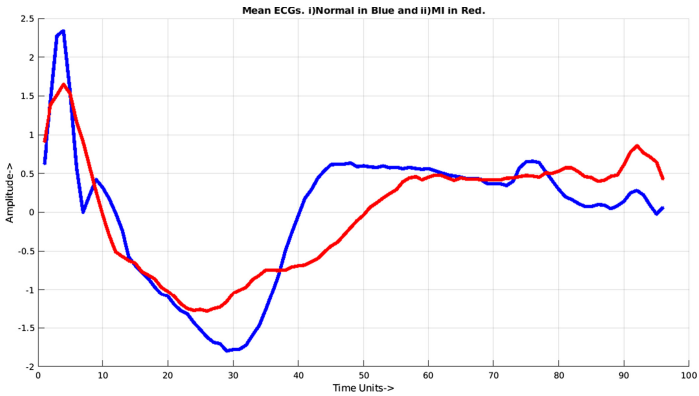
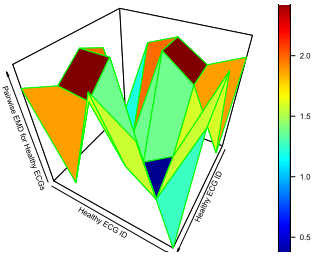
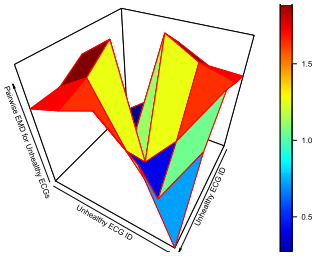


Fig. 3. Averaged out healthy and diseased ECG characteristics. This is used to compare a test ECG to decide its category as healthy or diseased.



(a) Pairwise EMD for first five healthy ECGs. This method is used to compute representative of healthy ECGs



(b) Pairwise EMD for first five unhealthy ECGs. This method is used to compute representative of unhealthy ECGs

Fig. 4. Pairwise EMD visualization for healthy and unhealthy ECGs

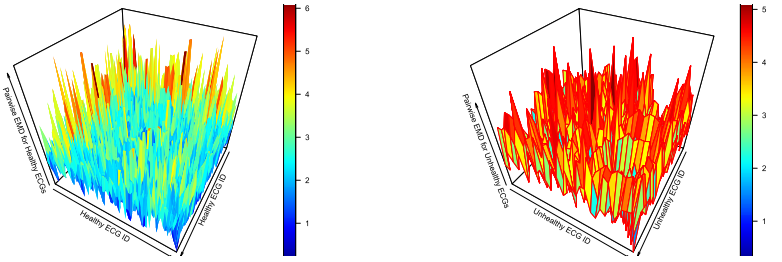
Table 1. Pairwise EMD for first five unhealthy ECGs

	Unhealthy ECG 1	Unhealthy ECG 2	Unhealthy ECG 3	Unhealthy ECG 4	Unhealthy ECG 5
Unhealthy ECG 1	0	0.4691725	1.887169	1.608353	1.586724
Unhealthy ECG 2	0.4691725	0	2.258859	1.943039	1.781622
Unhealthy ECG 3	1.887169	2.258859	0	0.835795	2.008787
Unhealthy ECG 4	1.608353	1.943039	0.835795	0	1.438485
Unhealthy ECG 5	1.586724	1.781622	2.008787	1.438485	0
Sum of EMDs	**5.5514185**	6.4526925	6.99061	5.825672	6.815618

of normal ECGs which can be used to compare the test ECGs with. Similar question arises to find a representative unhealthy ECG which can be used to compare the test ECGs with. It turns out that EMD answers these questions affirmatively and we will show how representative healthy and unhealthy ECGs are picked by calculating EMD amond pairwise healthy and pairwise unhealthy ECGs as shown in Table below. Number has been kept as five only otherwise matrix will become very large and can not be printed here. Full EMD matrix for both unhealthy case [5] with dimension 31×31 and healthy case [6] with dimension 69×69 is available with authors. EMD matrix among unhealthy ECGs is shown in Table 1 where last row shows the sum of EMDs of first ECG in first column with respect to other ECGs, second column for second ECG and so on. A simple inspection shows that in the row of Sum of EMDs, first column is minimum (as highlighted in bold and bigger size) which is interpreted that first ECG among these five unhealthy ECGs is most similar to all of them hence it can be taken as representative unhealthy ECG. Similar exercise was done for all 31 training ECGs and ECG number 25 was selected as representative among all unhealthy ECGs and will be used to compare the test ECGs with. A visual representation of this Table 1 is illustrated in Fig. 4b and for full 31 training unhealthy ECGs, pairwise EMD is show in Fig. 5b.

Table 2. Pairwise EMD for first five healthy ECGs

	Healthy ECG 1	**Healthy ECG 2**	**Healthy ECG 3**	Healthy ECG 4	Healthy ECG 5
Healthy ECG 1	0	2.427596	2.746186	2.250335	2.64763
Healthy ECG 2	2.427596	0	2.740578	2.143251	0.4748635
Healthy ECG 3	2.746186	2.740578	0	0.6655363	3.120786
Healthy ECG 4	2.250335	2.143251	0.6655367	0	2.522722
Healthy ECG 5	2.64763	0.4748636	3.120786	2.522722	0
Sum of EMDs	10.071747	7.7862886	9.2730867	**7.5818443**	8.7660015



(a) Pairwise EMD for all 69 training healthy ECGs. This is used to compute representative of training healthy ECGs

(b) Pairwise EMD for all 31 training unhealthy ECGs. This is needed to compute representative of training unhealthy ECGs

Fig. 5. Pairwise EMD 3D visualization for healthy and unhealthy ECGs at larger scale

Similarly, EMD matrix among healthy ECGs is shown in Table 2 where last row shows the sum of EMDs of first ECG in first column with respect to other ECGs, second column for second ECG and so on. A simple inspection shows that in the row of Sum of EMDs, fourth column is minimum (as highlighted in bold and bigger size) which is interpreted that fourth ECG among these five healthy ECGs is most similar to all of them hence it can be taken as representative healthy ECG. Similar exercise was done for all 69 training ECGs and ECG number 39 was selected as representative healthy ECG among all healthy ECGs and will be used to compare the test ECGs with. A visual representation of this Table 2 is illustrated in Fig. 4a and for full 69 healthy training ECGs is shown in Fig. 5a.

5.3 Disease Tagging with EMD

Now that we have representative healthy and unhealthy ECGs, we are going to compare the given 100 test ECGs with them. If test ECG is closer to healthy ECG i.e. its EMD is smaller to healthy one, then it is tagged as a healthy ECG (With tag 1) and if its resembles unhealthy ECG i.e. its EMD is small with respect to representative unhealthy ECG, then it is tagged as an unhealthy ECG (With tag -1) as shown in Table 3 next. The full process has been shown as a flowchart in Fig. 9. We compute that our success rate is 69 out of hundred or

69% which calls for multimodality and that is where biomarkers [14] walk in as a natural basis of CVD classification to further enhance the automated tagging of CVD with enhanced confidence.

5.4 Euclidean Distance Based Analysis of Training Data

Fundamental insight in this work which serves as a basis of Euclidean distance based analysis is that we can view each given ECG as a normal vector in 96 dimensional space and it gives us some fascinating results. We compute Euclidean distance as compared to EMD which we did in last section, for all 100 training ECGs and plot them in Fig. 7. Based on this Euclidean distance we cluster healthy and unhealthy ECGs. Figure 8 shows the two dimensional view of this clustering i.e. using first two coordinate of ECGs and Fig. 10 shows the three dimensional view of this clustering. The problem is as numbers increase, these clusters are intermixed and higher dimensional discrimination is needed for hi-fidelity classification. Finally, Fig. 11 shows the plot of Euclidean distance based automated detection success. Zero shows success and nonzero shows failure to correctly categorize. Software is able to predict correct category for 80% of the times based on Euclidean distance which is more than EMD which gives 69% of success rate. Although results look better for Euclidean based scheme, it has its cost in intense computational investment and likely will not scale if length of ECGs increase. In this work, since ECGs were only 96 dimensional, computational load was manageable. In future, lets say if ECG is 2000-dimensional, then it can lead to formidable computational problems and whole point of portability of such decision module is compromised.

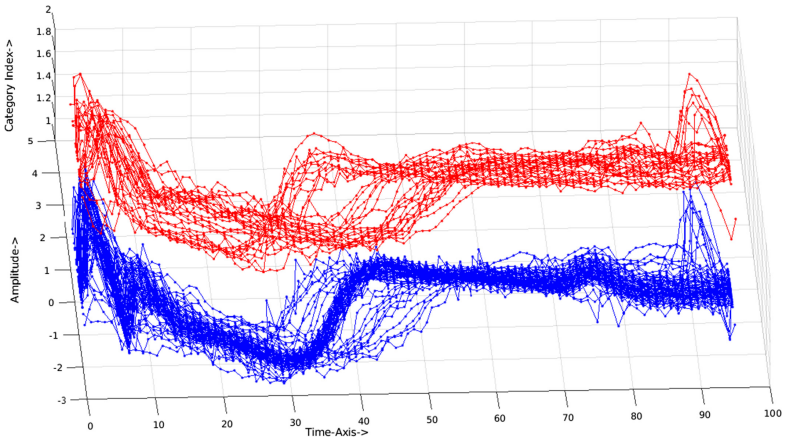


Fig. 6. Three dimensional view of different ECGs (healthy-blue, diseased-red) as temporal waveforms. (Color figure online)

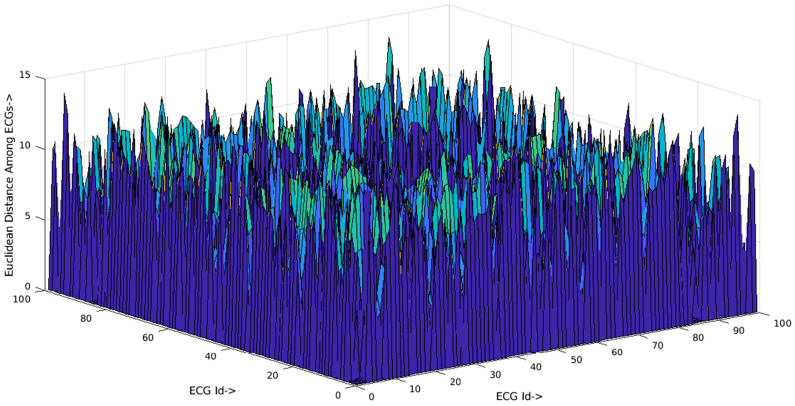


Fig. 7. Euclidean distance among training ECGs

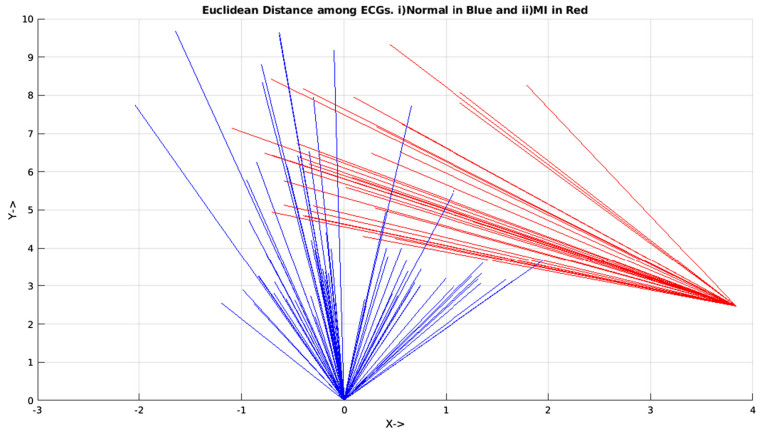


Fig. 8. Euclidean distance based two dimensional clustering for test ECGs

5.5 Error Analysis

Generally speaking, we have two kind of errors.

- Healthy ECG is tagged as an unhealthy one which shows up as 2 in error column.
- Unhealthy ECG is tagged as a healthy one which shows up as −2 in error column.

What is needed is other sensing modalities to cut into these false positive spaces of EMD or Euclidean distance based techniques. This is where a natural need for technique based on Biomarkers arises which can further enhance our confidence in automated disease tagging schemes and help us build next generation of biomedical machines which can save and prolong human lives.

5.6 Computational Infrastructure Deployed

Matlab has been used for performing geometrical part of the work. EMD aspect of this work has been performed in R software (Rstudio Version 1.3.1093 ©2009–2020 RStudio, PBC "Apricot Nasturtium" (aee44535, 2020-09-17) for Ubuntu Bionic Mozilla/5.0 (X11; Linux x86i_64) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.8 Chrome/69.0.3497.128 Safari/537.36) on a HP Probook laptop.

Laptop's operating system and other basic information from command uname -a is given below:

Linux Krishna 5.4.0-48-generic #52-Ubuntu SMP Thu Sep 10 10:58:49 UTC 2020 x86_64 x86i_64 x86_64 GNU/Linux

Output of hardware attributes of the laptop is as follows:

– memory
  description: System memory
  physical id: 0
  size: 8320 MiB

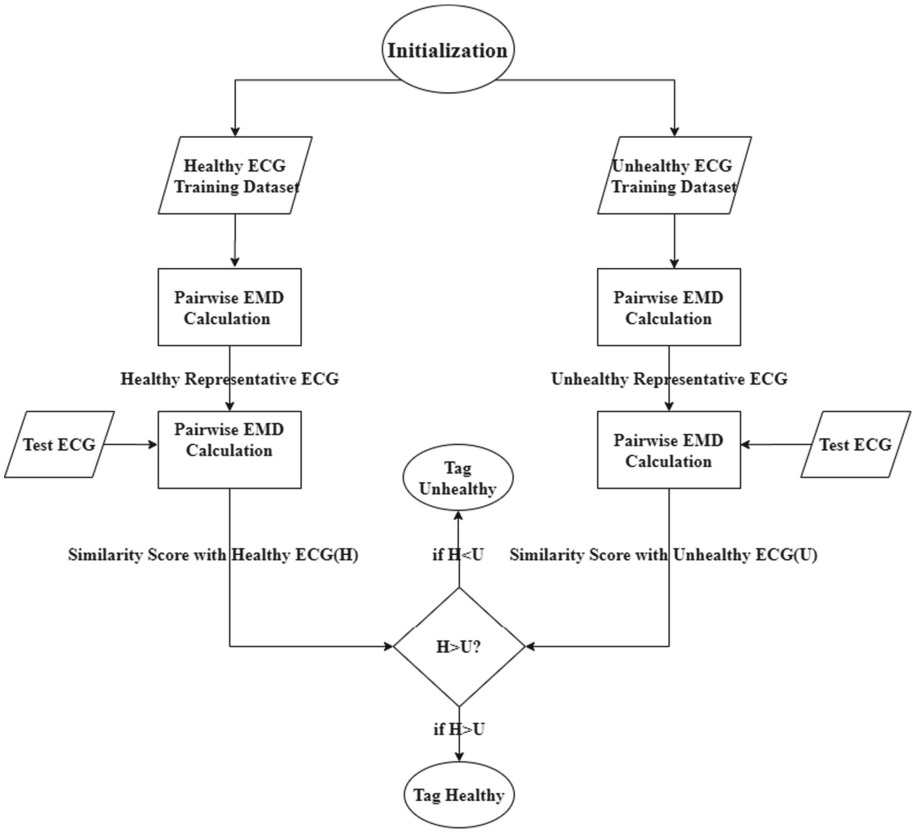


Fig. 9. Flow chart for computational disease tagging algorithm

– cpu
  product: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz
  vendor: Intel Corp.
  physical id: 1
  bus info: cpu@0
  size: 3304 MHz
  capacity: 3400 MHz

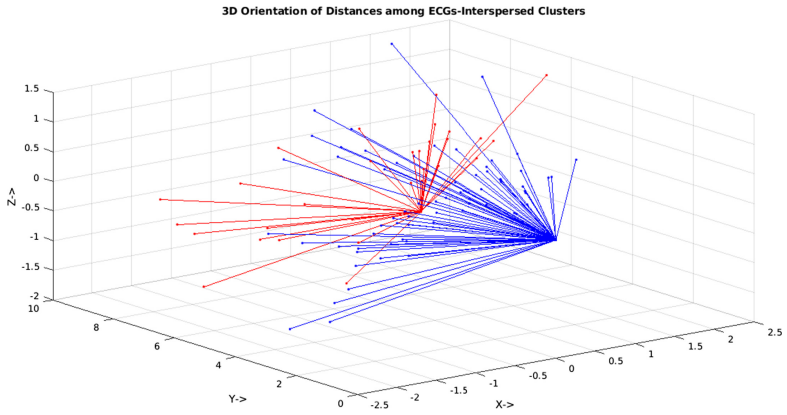


Fig. 10. Euclidean distance based three dimensional clustering for test ECGs

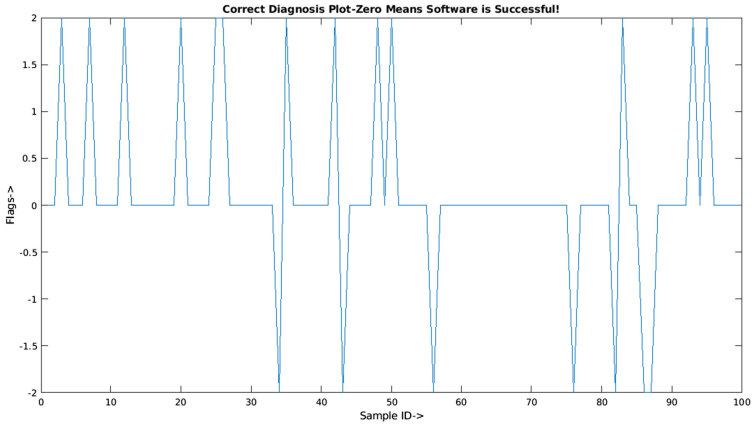


Fig. 11. Plot of Euclidean distance based automated detection success. Zero shows success and nonzero shows failure to correctly categorize. Software is able to predict correct category for 80% of the times based on Euclidean distance which is more than EMD which gives 69% of success rate.

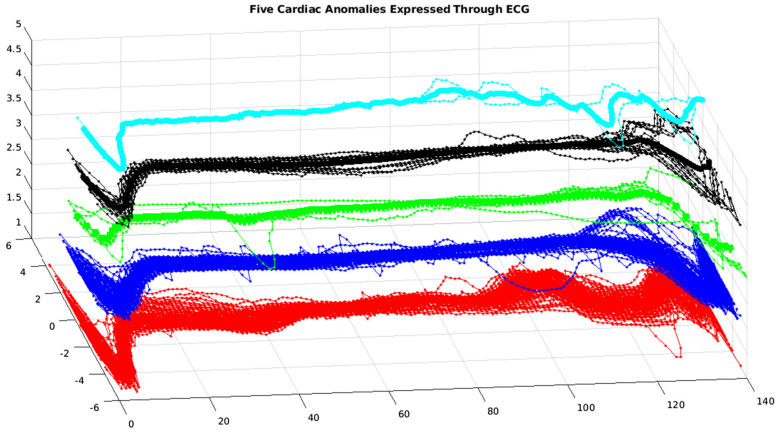


Fig. 12. Plot of five categories of ECGs in a 5000 ECGs training collection. Next generation of software is expected to classify given ECG in these five categories with some confidence using similarity based algorithms

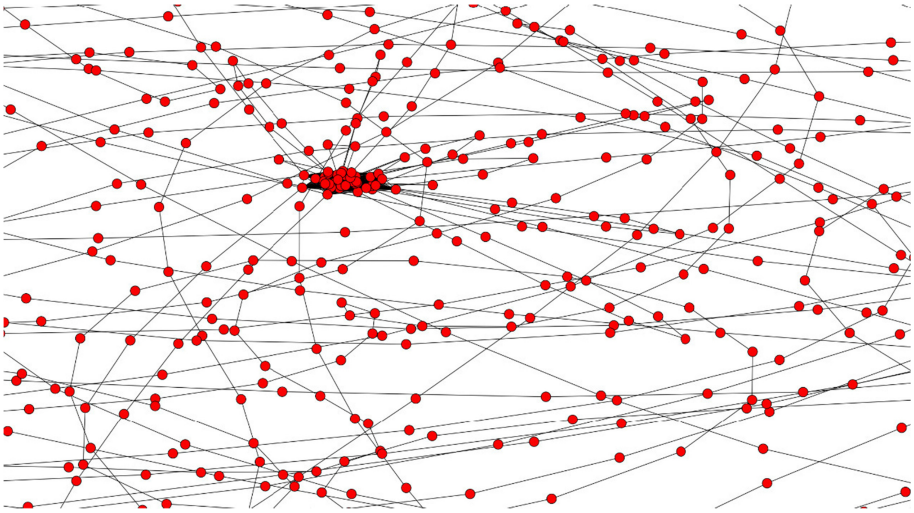


Fig. 13. Visibility graph of an ECG

Table 3. Computational disease tagging based on Earthmover's distance among ECGs

Test ECG ID	EMD with respect to healthy ECG	EMD with respect to unhealthy ECG	Our disease tag	Ground truth	Error
1	0.4508325	1.402279	1	1	0
2	0.3107516	1.449949	1	1	0
3	1.587482	0.9081023	−1	1	**2**
4	0.1772564	1.131663	1	1	0
5	1.595484	0.5994735	−1	−1	0
6	0.2734459	1.412014	1	1	0
7	1.394385	1.046782	−1	1	**2**
8	0.5062051	1.314533	1	1	0
9	0.6683038	1.3623	1	1	0
10	1.382945	0.7414857	−1	−1	0
11	1.097528	1.948595	1	−1	**−2**
12	2.33229	2.060142	−1	1	**2**
13	**2.801901**	**2.801836**	−1	−1	0
14	1.441449	0.4516492	−1	−1	0
15	2.60526	2.319956	−1	−1	0
16	1.128573	0.3905744	−1	1	**2**
17	0.5604649	1.60503	1	1	0
18	1.310016	1.296043	−1	1	**2**
19	1.284669	1.890638	1	−1	**−2**
20	0.8839369	1.700325	1	1	0
21	1.308087	2.08699	1	−1	**−2**
22	0.4983962	1.559452	1	1	0
23	1.017531	1.466081	1	1	0
24	0.2503529	1.404161	1	1	0
25	1.090369	1.815023	1	1	0
26	1.512469	2.003669	1	1	0
27	1.175686	0.4669774	−1	−1	0
28	1.223159	1.018802	−1	1	**2**
29	1.225785	1.987391	1	−1	**−2**
30	0.4989944	1.4843	1	1	0
31	0.6800542	1.324818	1	1	0
32	0.4493727	1.599889	1	1	0
33	1.495537	0.665113	−1	−1	0

(*continued*)

Table 3. (*continued*)

Test ECG ID	EMD with respect to healthy ECG	EMD with respect to unhealthy ECG	Our disease tag	Ground truth	Error
34	0.8376239	1.624659	1	−1	**−2**
35	2.35989	2.082147	−1	1	**2**
36	0.9122198	0.3954042	−1	1	**2**
37	1.009897	1.553917	1	1	0
38	0.6517924	1.202491	1	1	0
39	1.246308	2.086131	1	−1	**−2**
40	1.256244	0.4942391	−1	−1	0
41	0.3680083	1.306147	1	1	0
42	2.935643	2.657054	−1	1	**2**
43	0.6171389	1.675256	1	−1	**−2**
44	0.2309919	1.34059	1	1	0
45	0.4137751	1.554165	1	1	0
46	0.3178781	1.348103	1	1	0
47	0.5051144	1.602979	1	1	0
48	1.360489	1.012411	−1	1	**2**
49	0.1103903	1.23562	1	1	0
50	1.202057	1.964359	1	1	0
51	0.9608793	1.374716	1	1	0
52	1.413128	1.96717	1	−1	**−2**
53	1.062305	1.02717	−1	−1	0
54	1.412209	0.4888431	−1	−1	0
55	2.049299	2.128684	1	−1	**−2**
56	0.513019	1.294244	1	−1	**−2**
57	0.3262479	1.315604	1	1	0
58	1.543188	0.5049611	−1	−1	0
59	2.610557	2.388494	−1	−1	0
60	1.500736	0.9894223	−1	−1	0
61	1.797978	1.692139	−1	−1	0
62	0.5567521	1.708674	1	1	0
63	0.9454798	2.007701	1	1	0
64	1.123959	0.4063717	−1	−1	0
65	1.882564	1.081164	−1	−1	0
66	1.490418	0.4881482	−1	−1	0
67	0.5267136	1.68411	1	1	0
68	1.330896	1.28545	−1	1	**2**

(*continued*)

Table 3. (*continued*)

Test ECG ID	EMD with respect to healthy ECG	EMD with respect to unhealthy ECG	Our disease tag	Ground truth	Error
69	1.425453	1.92742	1	1	0
70	0.3167376	1.415025	1	1	0
71	1.238598	1.576925	1	1	0
72	0.3772002	1.463139	1	1	0
73	0.3584798	1.415153	1	1	0
74	1.489268	1.774857	1	1	0
75	0.780029	1.413753	1	1	0
76	0.5997958	0.9639898	1	−1	**−2**
77	0.1913777	1.277019	1	1	0
78	0.9003933	0.7171211	−1	1	**2**
79	0.2772272	1.406237	1	1	0
80	0.5605359	1.52097	1	1	0
81	0.842902	1.330435	1	1	0
82	0.8392607	0.671752	−1	−1	0
83	2.35084	2.075138	−1	1	**2**
84	0.2239769	1.382525	1	1	0
85	1.778807	1.406144	−1	−1	0
86	0.8491539	0.9131678	1	−1	**−2**
87	0.5174375	1.187815	1	−1	**−2**
88	1.451983	0.4261783	−1	−1	0
89	0.5274906	0.8449174	1	1	0
90	0.5280408	1.424953	1	1	0
91	0.3098607	1.319613	1	1	0
92	1.581227	1.747838	1	−1	**−2**
93	1.55048	1.024037	−1	1	**2**
94	0.9683289	1.768002	1	1	0
95	1.095088	0.2158081	−1	1	**2**
96	0.8317844	1.407165	1	1	0
97	1.264388	1.690354	1	1	0
98	1.776811	1.644949	−1	−1	**−2**
99	0.9506655	1.976846	1	−1	**−2**
100	0.6648064	0.5372591	−1	1	**2**

6 Conclusions, Future Directions and COVID-19 Context

This work shows that using basic geometric ideas like Euclidean distance or similarity detection metrics like Earthmovers distance used frequently in face-detection problems can provide upto eighty percent accuracy in healthy ECG tagging. This can help the PHCs to make sense of ECGs using this work and further provide referral advice to patients. Integration of visual imaging based tools with biomarkers [14] like Troponin isoforms, CKMB, BNP may enhance the capability to automatically categorize ECGs correctly but also may add to the instrumentation and costs. Future software will be expected to classify given ECG in these five categories with some confidence using similarity based algorithms as shown in Fig. 12. EMD has many different implementations based on what kind of distancing we pick, i.e. "manhattan" distancing is likely to give different result from $emdL1$ scheme. Even we can deploy drastically different methods like Visibility graph (VG) [7] as shown in Fig. 13 where different aspect of time series data can be used to automatically label healthy and diseased ECGs. Both EMD and VG can give rise to even a hybrid scheme if performance is boosted! A major future direction is to study the interspersed nature of healthy and unhealthy ECG vectors in clustering space to provide guidelines for disease tagging when it is too close to call or it is obfuscating as shown in Fig. 8 and in Fig. 10. Further indecision points arise in Table 3, row number 13 where EMD from both healthy and unhealthy ECGs is equal upto second place of decimal. This ECG is equidistant from healthy and unhealthy world in equal magnitude. Such problems arise because of limited size and diversity of datasets and there is a natural need for larger and more heterogeneous datasets of ECGs apart from other modalities like Cardiac Auscultation [13] and bio-markers [14].

Although India has witnessed an epidemiological transition from communicable diseases to NCDs in the last two decades with Cardiovascular Diseases and Cancer accounting for a significant proportion of morbidities and mortalities, the unfinished agenda of communicable diseases has led to emergence of recent pandemics such as COVID-19. With the emergence of COVID19, there is deep, broad recognition of our deeply fractured system of healthcare and public health and the long road ahead to improve on it. The question is no longer whether we should strengthen our healthcare infrastructure but how. The COVID-19 pandemic has overshadowed developmental activities across the world. The global political, financial and technical resources have been mobilized to contain COVID-19 pandemic. Impact of this pandemic shall be long-lasting, influencing all spheres of human lives and slowing all developmental activities including ambitious and aspirational Sustainable Development Goals (SDGs). It is an irony, that it has taken a highly communicable condition like COVID-19 to highlight the dangers inherent in non-communicable diseases (NCDs), such as CVDs along with risk factors such as diabetes, hypertension, and obesity to health. These are usually bracketed away as previously underlying or pre-existing medical conditions when, in fact, they should be in the foreground. It should come as no surprise that population with underlying co-morbid risk factors of NCDs are at the highest risk of suffering complications and reduced life

expectancy due to COVID-19 pandemic, are not only the elderly or geriatric population but also unfortunately comprise the population contributing to the nation building. Yet this aspect is never quite stressed. Worldwide data show that annually 15 million people die from an NCD between the ages of 30 years and 69 years. Once these figures are seen in the context of COVID-19 fatalities the strike rate of this dreaded pandemic is not nearly as random as it may first appear. For years now, the recently much maligned World Health Organization (WHO) has been warning and reminding countries about the dangers inherent in the spread of NCDs. The epidemic of NCDs poses devastating health consequences for individuals, families and communities, and threatens to overwhelm health systems. The socio-economic costs associated with NCDs make the prevention and control of these diseases a major development imperative for the 21st century. Healthcare technologies enable us to understand the patterns and processes associated with the landscape of disease burden in the Low and middle income countries (LMICs) burdened with fractious and fractionated healthcare ecosystems. India endowed with heterogeneous genetic base, divergent socio-cultural norms, along with varied geological relief structures provides a veritable landscape of disease burden resulting in the prevalence of healthcare disparities. Use of AI-enabled technologies combined with big data enabled platforms provides rationale for strengthening resource-deprived healthcare systems. Our AI-enabled solution is aimed at not only large-scale screening of CVDs in resource limited rural milieus of the Indian sub-continent for disease labeling of subjects afflicted with CVDs such as MI but also triaging subjects in the order of disease severity. This will significantly alleviate the clinical prognosis of the patients through timely clinical interventions.

Finally, one of the major question with which Indian clinical community has been struggling is which ECG should be tagged "Healthy" when there is so much variation in what clinicians mean by a "Healthy" ECG. This work successfully picks a representative "Healthy" ECG from an ensemble of "Healthy" ECGs using Earthmover's distance (EMD) based on the premise that this representative "Healthy" ECG should demonstrate maximum similarity to all other "Healthy" ECGs in given ensemble. It is interesting to note that a serious issue of medical community finally gets resolved in the world of computational similarity.

7 Device Utility

1. Has potential application as a Adjunct Clinical Aid for the Cardiologists/Medical Professionals
2. Automatic Classification of ECG images facilitating large scale screening of subjects in remote health camps.
3. Easy, fast and robust technology with capabilities to be implemented in web-based, desktop-based and smartphone-based applications when coupled with an ECG measurement device on the internet.
4. It has potential of turning CVD disease management as a self care exercise. Control moves from the hands of expensive hospital to cheap and affordable selfcare.

References

1. Rubner, Y., Tomasi, C.: The earth movers distance perceptual metrics for image database navigation (2001)
2. Olszewski, R.T.: Generalized feature extraction for structural pattern recognition in time-series data. No. CMU-CS-01-108. Carnegie Mellow University, School of Computer Science (2001). https://www.cs.cmu.edu/bobski/pubs/tr01108-twosided.pdf
3. http://www.timeseriesclassification.com/description.php?Dataset=ECG200
4. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
5. https://www.mediafire.com/file/rdper56710dvekx/un-emd.pdf/file
6. https://www.mediafire.com/file/2xzsmp3cufj9a0a/healthy-emd.pdf/file
7. Ji, H., Xu, T., Wu, W., Wang, J.: Visibility graph analysis on EEG signal. In: 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, pp. 1557–1561 (2016). https://doi.org/10.1109/CISP-BMEI.2016.7852963
8. Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F.: The computational limits of deep learning. arXiv arXiv:2007.05558 (2020)
9. Yampolskiy, R.V.: Unpredictability of AI. arXiv arxiv:1905.13053 (2019)
10. Yampolskiy, R.V.: Unexplainability and incomprehensibility of artificial intelligence. arXiv arXiv:1907.03869 (2019)
11. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. arXiv arXiv:1503.03585 (2018)
12. Li, H.: Analysis on the nonlinear dynamics of deep neural networks: topological entropy and chaos. arXiv arXiv:1804.03987 (2018)
13. Wang, F., Syeda-Mahmood, T., Beymer, D.: Finding disease similarity by combining ECG with heart auscultation sound. In: Computers in Cardiology, Durham, NC, pp. 261–264 (2007). https://doi.org/10.1109/CIC.2007.4745471
14. Than, M.P., et al.: Machine learning to predict the likelihood of acute myocardial infarction. Circulation **140**(11), 899–909 (2019)
15. Morris, P.D., et al.: Computational fluid dynamics modelling in cardiovascular medicine. Heart **102**(1), 18–28 (2016)
16. Gharleghi, R., Samarasinghe, G., Sowmya, A., Beier, S.: Deep learning for time averaged wall shear stress prediction in left main coronary bifurcations. In: IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, pp. 1–4 (2020). https://doi.org/10.1109/ISBI45749.2020.9098715
17. https://venturebeat.com/2020/07/15/mit-researchers-warn-that-deep-learning-is-approaching-computational-limits/. Accessed 15 July 2020
18. Scassellati, C., Bonvicini, C., Benussi, L., Ghidoni, R., Squitti, R.: Neurodevelopmental disorders: metallomics studies for the identification of potential biomarkers associated to diagnosis and treatment. J. Trace Elem. Med. Biol. **60** (2020). ISSN 0946–672X. https://doi.org/10.1016/j.jtemb.2020.126499
19. Chen, R.J., et al.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. https://arxiv.org/abs/1912.08937 (2019)
20. Carmichael, I., et al.: Joint and individual analysis of breast cancer histologic images and genomic covariates. https://arxiv.org/abs/1912.00434 (2019)

21. Guo, A., Chen, Z., Li, F., Li, W., Luo, Q.: Towards more reliable unsupervised tissue segmentation via integrating mass spectrometry imaging and hematoxylin-erosin stained histopathological image (2020). https://www.biorxiv.org/content/10.1101/2020.07.17.208025v1
22. Shrivastav, K.D., et al.: Earth mover's distance-based automated geometric visualization/classification of electrocardiogram signals. In: Sarma, H., Bhuyan, B., Borah, S., Dutta, N. (eds.) Trends in Communication, Cloud, and Big Data. LNCS, vol. 99, pp. 75–85. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1624-5_8