Vilnius university
Institute of Data Science and Digital Technologies
L I T H U A N I A

INFORMATICS ENGINEERING (T007)

DATA FUSION FOR BETTER DECISION MAKING IN THE MEDICAL FIELD

Jaroslava Arsenjeva

October 2020

Technical Report DMSTI-DS-T007-20-12

VU Institute of Data Science and Digital Technologies, Akademijos str. 4, Vilnius
LT-08412, Lithuania
www.mii.lt

Abstract

This paper overviews most common algorithms and state-of-art methods for predicting predicting a medical condition called ARDS – acute respiratory distress syndrome: a lethal condition that can be stopped if predicted early enough. A novelty algorithm that uses data fusion has been proposed in 2018. The proposed method used the MIMIC II database to extract data streams measuring heart rate, respiratory rate, oxygen saturation and the mean abdominal pressure. After combining four data streams together a patient was assigned a weighted decision index *D* form -1 to 1 where -1 meant that the patient was at risk of acquiring ARDS and +1 patients were considered stable. Authors of the proposed method managed to get an 85% accuracy rate with a possible prediction of 39 hours before the onset of the syndrome. Author of this paper has set the task to replicate the methodology that already exists and find a way to improve the accuracy of ARDS prediction.

Keywords: data fusion, ARDS prediction, data analysis.

Contents

Introduction

Data fusion has existed for decades however recently with the evolution if smart homes and devices and industry 4.0 it had gained popularity. Data fusion itself is the process of combining information fro, multiple sensors. It is obvious that the a human with all five senses is better at making decisions (let’s say when to cross the street) that a human, who is blind or deaf (has only less than 5 senses). Similarly decision making for technical algorithms is easier when more information about the object is available.

The term “data fusion” was first mentioned by JDL in 1985. The initial model had 4 fusion levels and since then it has been modified multiple times. [1]

Being mainly a militaristic methodology in the beginning, data fusion is now being applied in many fields such as automotive, agricultural, medical and others. The main tasks data fusion tackles are target acquisition [3], health diagnosis estimation [4], image combining [5], risk analysis [6], fingerprint and/or object recognition [7] [8] [9], sensor information processing in wireless networks [10] [11] [12] and many others.

Data fusion can obviously be applied in many fields for various tasks however in this paper the main focus will be on the medical field. Even in medicine the number of problems data fusion can aid in solving varies however from predicting future syndromes and illnesses form the data available now humankind will benefit the most. One of the syndromes – ARDS, or acute respiratory distress syndrome – is quite crucial nowadays because it may be caused by infection of SARS-CoV-2.

ARDS occurs when fluid builds up in the alveoli in lungs. The fluid keeps the lungs from filling with enough air, which deprives organs of the oxygen they need to function. Although some people who develop ARDS don't survive there is possibility to recover completely if the person was taken care of timely. Since symptoms usually develop within a few hours to a few days after the previous injury or infection, it is possible to monitor ill people to predict the onset of the syndrome. [14]

2 Research methodology for data fusion

2.2 Data streams

Sometimes the data from the sensor is presented in a way of a continuous data flow. The data usually is recorded into smaller bits which are stored in separate files and is need in pre-processing to be analyzed. These data streams usually come in a format of a measurement on y axis and the time on x axis so it is easy to compare multiple heterogeneous data streams with the same data stamp. With the overwhelming number of smart watches and bands a lot of information like this can be collected even by a general user.

A great example of such data streams is the MIMIC database [15]. MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising anonymous health data associated with ~60,000 intensive care unit admissions. The example from physionet on how the data might look like may be found below (Figure 2.1.)

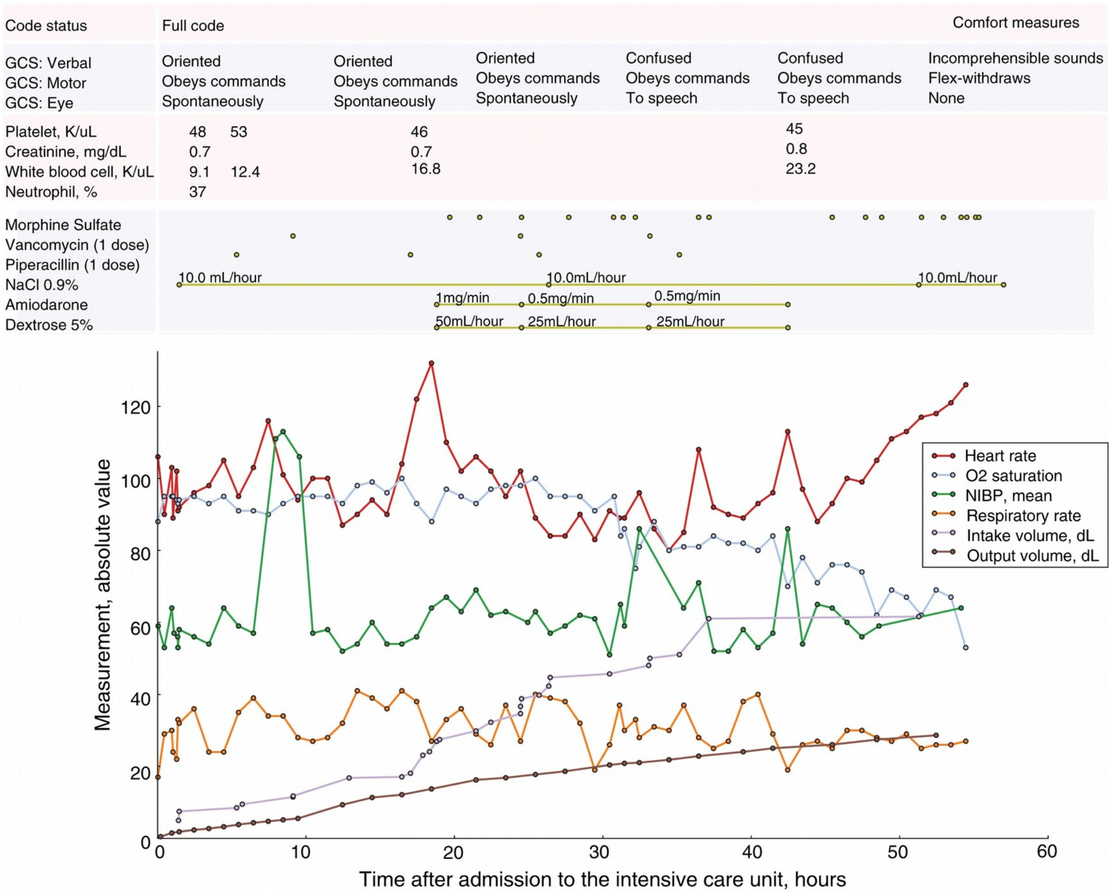

Figure 2.1. Example of medical data streams [16]

2.3 Prediction technique

In [17] and [18] authors have proposed a novelty algorithm that helps predicting ARDS with the use of data fusion. The authors propose a framework that is analysing four different data streams that measure heart rate, respiratory rate, oxygen saturation and the mean airway blood pressure. The frequency of taken measurements is 0.016 Hz. For each data stream *1:n* (usually at least 24h long) a starting ‘normal’ period *1:k* is taken. After the standard deviation is calculated for the whole data stream. An outlier definition is proposed as any point existing in the stream outside the three standard deviation yields.

When data is normally distributed, in case of a number of outliers in the k:n exceeds 1 percent the patient is considered to have a risk of developing ARDS. In the more likely case of abnormal distribution, the distribution of outliers is calculated for *1:k* and *k:n* periods. Although the outlier and k can be selected by the researcher, it is proposed in the paper to optimize these values using real data. In case the number of outliers in the *k:n* period is bigger than in *1:k* then the data stream is assigned with a weight +1 (likely to develop ARDS), otherwise weight is -1.

Since four data streams were analysed, now four inferences (*D)* with values -1 and +1 are obtained. The fusion model proposed takes into account these inferences to develop a final prediction (Figure 2.2.).

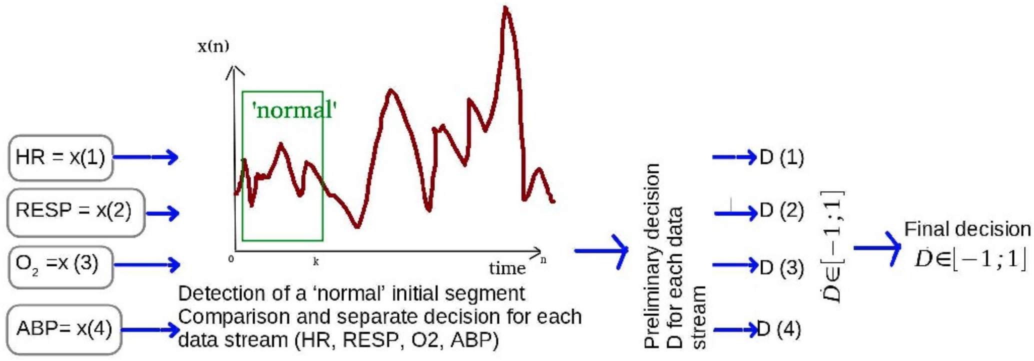

Figure 2.2. Example of ARDS predicting model.

In case the data would be normally distributed, authors consider an outlier any value outside of the three standard deviations window however for the real conditions a more general approach is proposed. An outlier value T is selected, then again if the original 'normal' segment has less points which value exceeds T, than the segment k:n, then the patient is considered unstable and an inference D=1 is assigned.

Of course, the value of T can be selected by the researcher however for better results is was optimized together with the length of 'normal' segment K. for this sensitivity - the proportion of positives that are correctly identified and specificity - the proportion of negatives that are correctly identified were estimated. Performance indexes sensitivity Se and the specificity Sp of the whole data set were computed for each couple of values of K and T within a set range. Then the Youden index [19] was maximized ($Se + Sp - 100$), leaving authors with most accurate T and K values.

2.4. Data fusion model for decision making

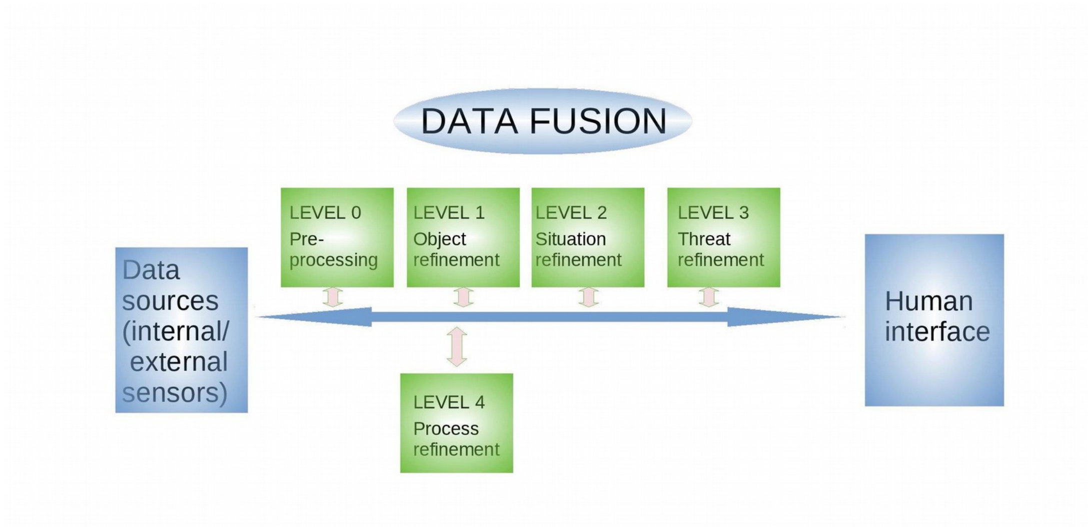

Figure 2.3. Original JDL model

As mentioned before, the original JDL model (Figure 2.3.) has four levels of fusion. The newer model have deteriorated to just three levels: raw data fusion, feature fusion and inference fusion. Raw data fusion can only be applied in case of homogeneous data. The easiest intuitive way to combine data from homogeneous sensors is to simply take the weighted average however a better method for raw data processing is Kalman Filter.

If the obtained data from sensors is not of the same kind (like in the case with ARDS prediction: HR and RESP are measured in beats per minute, ABP is measured in mmHg and oxygen saturation in percentage) , the fusion must happen at feature extraction or inference levels. Authors [17,18] have used the inference fusion level, this means that data was combined to make a final decision after making inferences for each data stream separately.

2.5. Inference fusion situation prediction

To combine all the individual decisions, a multivariate data analysis is performed using the Chair and Varshney fusion technique [18,20]. The final decision is calculated as follows:

Final decision = $D_0 + \sum \bar{D}_i W_i$

$$D_0 = \log\left(\left(Amount\ of\ stable\right) / \left(Amount\ of\ unstable\right)\right)$$

W_i for (stable, D=-1) is calculated as follows:
log((sensitivity)/(1-specificity))

W_i for (unstable, D=1) is calculated as follows:
log((specificity)/(1-sensitivity))

After if the final decision is >0 then the patient is considered unstable and at risk at developing ARDS. This fusion method was tested with the ARDS onset in 6, 12, 18 and 24 hours. The proposed data fusion algorithm succeeded in predicting ARDS 24 hours before of its occurrence with a sensitivity higher than 78% (compared to 64 – 77 if using singular data streams for prediction models), and to identify 50% of stable subjects [18].

3 Future work

Since the proposed above method managed to improve the sensitivity but lacked in the way of improving specificity (e.g. identifying not at risk patients that could be left unsupervised) I have the intention to use a different approach to try to improve the algorithm. Since the MIMIC database is available to general public, the goal for the future semester would be to duplicate the carried out research in the paper mentioned above and try to get similar results.

After the goal would be to create 6, 12, 18, 24 hour graphs with all four data stream lines on them and analyze them as images using CNN (convolution neural network) with the goal of creating same +1 or -1 outputs for unstable and stable patients accordingly.

The main problem when attempting tor each this goal would be the data per-processing because in the MIMIC database downloaded data is stored in separate files

(Figure 3.1.) and sometimes measurements must be combined into continuous streams by using their ID.

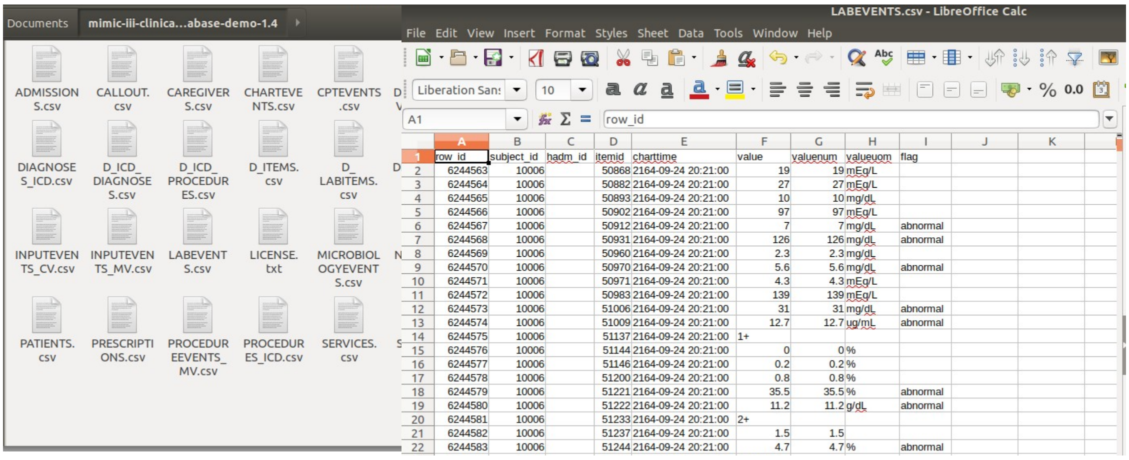

Figure 3.1. MIMIC III database

4 Data fusion application in the medical field

As mentioned above, the data fusion model has slightly changed over the years and now has 3 main levels: raw data fusion, feature (vector) fusion and inference fusion. Most common algorithms [21] and their applications can be found in Table 4. [13]. In medicine image fusion mainly feature and inference fusion techniques are used and currently more and more papers appear on this topic. Challenges arising are the size of data needed to be collected (also images must be of a good quality/ resolution); the computing time is quite long and the data cannot be processed in real time and for inference fusion the extracted vectors might have contradicting values so the fusion result might be inaccurate of inconclusive.

Table 4. Common algorithms for data fusion in medicine

Lvl	Algorithm	Example of application	Reference
Raw data fusion	AVG/ weighted AVG	The average of temperature measurements from different body parts / different thermometers (only homogeneous data)	
	Kalman filter	Acceleometer + gyroscope data (detecting if a person is standing by himself or leaning on a wall)	[22]
	Particle filter	Detecting the position location of a human; segmentation of noisy medical images (internal body parts)	[23]

Vector/ feature fusion	ANN (deep learning)	Activity recognition, natural language processing, cancer detection, identification and diagnosis of micro-calcification, remote medical diagnosis, image processing (convolution layers)	[36], [37], [38], [39], [40]
	Decision trees	Mostly activity recognition, work better with categorized data; condition surveillance; heath system improvement	[28], [29], [30], [31]
	kMeans	Activity rcognition, cancer diagnosis and treatment, healthy/ unhealthy cell classification	[41], [42], [43]
	kNN	Activity classification; healthy/unhealthy cell classification	[24], [25], [26] ,[27]
	SVM Support vector machines	Activity classification (linear and non-linear); cancer diagnosis and treatment, tumor classification, gene classification	[32], [33], [34], [35]
Inference fusion	Bayesian inference	Activity recognition, reverse engineering of transcription regulatory networks from genomics; cancer diagnosis; gene system modeling and analyzing	[45], [46], [47]
	Fuzzy logic	Image based decisions (cancer cells identification for example); diagnosis of brain conditions, cancer treatment; ovarian cancer detection and diagnosis, gene model reconstruction	[48], [49], [50], [51], [52]

Conclusion

With the increasing popularity of data fusion, medical task such as recognition of cancer cells, gene reverse engineering, detection of unhealthy body parts and diagnosis assessment are becoming less expensive and human-hour consuming. Doctors all over the world use machine learning to analyse images they have gathered from electronic microscopes and Xrays and MRI scans. In some places fully autonomic systems are entitled to question the patients and diagnose them.

Although data fusion techniques have come a long way since they have been first used in medicine, the methodology still has a long way to go: to the real-time big data processing frameworks that can detect or even predict an illness weeks before the first symptoms will arise. Hopefully scientists will continue coming up with novelty methods for medical data fusion and the technique will become even more advanced in the future.

References

1. Liggins, Martin E.; Hall, David L.; Llinas, James (20
2. 08). Multisensor Data Fusion, Second Edition: Theory and Practice (Multisensor Data Fusion).
3. Shen X., P.K. Varshney, Sensor selection based on generalized information gain for target tracking in large sensor networks, IEEE Trans. Signal Process., 62 (2), (2014), pp. 363 – 375
4. Dong M., He D., Hidden semi-markov model based methodology for multi-sensorequipment health diagnosis and prognosis, Eur. J. Oper. Res., 178 (3) (2007), pp. 858 - 878
5. Yang F., Wei H., Fusion of infrared polarization and intensity images using support value transform and fuzzy combination rules, Infrared Phys. Technol., 60 (2013), pp. 235 – 243
6. Zhang L., Wu X., Zhu H., S.M. AbouRizk, Perceiving safety risk of buildings adjacent to tunneling excavation: an information fusion approach, Autom. Constr., 73 (2017), pp. 88 -101.
7. D. Peralta, I. Triguero, S. García, Y. Saeys, J.M. Benitez, F. Herrera, Distributed incremental fingerprint identification with reduced database penetration rate using a hierarchical classification based on feature fusion and selection, Knowl. Based Syst., 126 (2017), pp. 91-103
8. D. Peralta, I. Triguero, S. García, F. Herrera, J.M. Benitez, DPD-DFF: a dual phase distributed scheme with double fingerprint fusion for fast and accurate identification in large databases, Inf. Fusion, 32 (2016), pp. 40-51
9. G. Fortino, S. Galzarano, R. Gravina, Li W., A framework for collaborative computing and multisensor data fusion in body sensor networks, Inf. Fusion, 22 (2015), pp. 50-70
10. Zhang Z., Zhang W., Chao H.C., Lai C.F., Toward belief function based cooperative sensing for interference resistant industrial wire less sensor networks, IEEE Trans. Ind. Inf., 12 (6) (2016), pp. 2115 - 2126
11. Zhang Z., Hao Z., S. Zeadally, Zhang J., Han B., Chao H. - C., Multiple attributes decision fusion for wireless sensor networks based on intuitionistic fuzzy set, IEEE Access, 5 (2017), pp. 12798 - 12809
12. Duan Y., Fu X., Li W., Zhang Y., G. Fortino, Evolution of scale -free wireless sensor networks with feature of small world networks, Complexity, 2017 (3) (2017).
13. James, Alex & Dasarathy, Belur V. (2015). A Review of Feature and Data Fusion with Medical Images.
14. Article: ARDS from MayoClinic website. https://www.mayoclinic.org/diseases-conditions/ards/symptoms-causes/syc-20355576.
15. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
16. https://www.nature.com/articles/sdata201635/figures/2
17. Taoum A, Mourad-Chehade F, Amoud H. Early-warning of ARDS using novelty detection and data fusion. Comput Biol Med. 2018 Nov 1;102:191-199.

18. T. Aline, M. Farah, A. Hassan, C. Aly, F. Ziad and D. Jacques, "Data fusion for predicting ARDS using the MIMIC II physiological database," 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), Munich, 2016, pp. 1-5,
19. https://www.sciencedirect.com/topics/medicine-and-dentistry/youden-index
20. A. I. Hernandez, G. Carrault, F. Mora, L. Thoraval, G. Passariello and J.-M. Schleich, "Multisensor fusion for atrial and ventricular activity detection in coronary care monitoring", IEEE Trans. Biomed. Eng., vol. 46, no. 10, pp. 1186-1190, Oct. 1999,
21. Rachel C. King, Emma Villeneuve, Ruth J. White, R. Simon Sherratt, William Holderbaum, William S. Harwin, Application of data fusion techniques and technologies for wearable health monitoring, Medical Engineering & Physics, Volume 42, 2017, pp. 1-12,
22. A. Al-Jawad, A. Barlit, M. Romanovas, M. Traechtler, Y. Manoli The use of an orientation Kalman filter for the static postural sway analysis APCBEE Procedia, 7 (2013), pp. 93-102,
23. Fischer, Lukas. (2009). Using Shape Particle Filters for Robust Medical Image Segmentation,
24. L. Atallah, Lo B., R. King, Yang G.-Z. Sensor positioning for activity recognition using wearable accelerometers IEEE Trans Biomed Circ Syst, 5 (4) (2011), pp. 320-329,
25. Liu S., Gao R.X., D. John, J.W. Staudenmayer, P.S. Freedson Multisensor data fusion for physical activity assessment IEEE Trans Biomed Eng, 59 (3) (2012), pp. 687-696,
26. N. Bicocchi, M. Mamei, F. Zambonelli Detecting activities from body-worn accelerometers via instance-based algorithms Pervasive Mob Comput, 6 (4) (2010), pp. 482-495,
27. Vitola J, Pozo F, Tibaduiza DA, Anaya M. A Sensor Data Fusion System Based on k-Nearest Neighbor Pattern Classification for Structural Health Monitoring Applications. Sensors (Basel). 2017 Feb 21;17(2):417,
28. Frantzidis C.A., Bratsas C., Klados M.A., Konstantinidis E., Lithari C.D., Vivas A.B., Papadelis C.L., Kaldoudi E., Pappas C., Bamidis P.D. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. Trans. Inf. Tech. Biomed. 2010;14:309–318,
29. L. Bao, S.S. Intille, A. Ferscha, F. Mattern Activity recognition from user-annotated acceleration data. pervasive computing, Lecture notes in computer science, vol. 3001, Springer, Berlin Heidelberg (2014),
30. O. Banos, M. Damas, H. Pomares, A. Prieto, I. Rojas, Daily living activity recognition based on statistical feature quality group selection, Expert Syst Appl, 39 (9) (2012), pp. 8013-8021,
31. L.C. Jatobá, U. Großmann, C. Kunze, J. Ottenbacher, W. Stork Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity. Proceedings off the IEEE conference on engineering in medicine and biology society (2008), pp. 5250-5253,
32. G. Pavesi, and G. Valentini, “Classi①cation of co-expressed genes from DNA regulatory regions,” Information Fusion, vol. 10, no. 3, pp. 233–241, 2009,
33. Vrigazova, B. (2020). Detection of Malignant and Benign Breast Cancer Using the ANOVA-BOOTSTRAP-SVM, Journal of Data and Information Science, 5(2), 62-75,

34. Wang H, Huang G. Application of support vector machine in cancer diagnosis. Med Oncol. 2011 Dec;28 Suppl 1:S613-8,
35. Sweilam, Nasser & Tharwat, Assem & Moniem, N.. (2010). Support vector machine for diagnosis cancer disease: A comparative study. Egyptian Informatics Journal. 11. 81-92. 10.1016/j.eij.2010.10.005.,
36. LeCun Y., Y. Bengio, G. Hinton, Deep learning, Nature, 521 (7553) (2015), pp. 436-444,
37. K. Altun, B. Barshan, O. Tunç el, Comparative study on classifying human activities with miniature inertial and magnetic sensors, Pattern Recogn, 43 (10) (2010), pp. 3605-3620,
38. Canbilen, Ayşe & Ceylan, Murat. (2015). Fusion and ANN Based Classification of Liver Focal Lesions Using Phases in Magnetic Resonance Imaging. 10.1109/ISBI.2015.7163900.,
39. D, Jude & Anitha, J. & Kwintiana, Bernadetta. (2017). Fusion of artificial neural networks for learning capability enhancement: Application to medical image classification. Expert Systems. e12225. 10.1111/exsy.12225.,
40. Ruholla Jafari-Marandi, Samaneh Davarzani, Maryam Soltanpour Gharibdousti, Brian K. Smith, An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals, Applied Soft Computing, Volume 72, 2018, Pages 108-120,
41. Gönen, Mehmet & Margolin, Adam. (2014). Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology. Advances in Neural Information Processing Systems. 2. 1305-1313.,
42. Zura Kakushadze, Willie Yu, K-means and cluster models for cancer signatures, Biomolecular Detection and Quantification, Volume 13, 2017, Pages 7-31,
43. Zhao S, Li W, Cao J. A User-Adaptive Algorithm for Activity Recognition Based on K-Means Clustering, Local Outlier Factor, and Multivariate Gaussian Distribution. Sensors (Basel, Switzerland). 2018 Jun;18(6),
44. L. Atallah, B. Lo, R. Ali, R. King, Yang G.-Z., Real-time activity classification using ambient and wearable sensors, IEEE Trans Inf Technol Biomed, 13 (6) (2009), pp. 1031-1038,
45. Xi Chen and Jianhua Xuan (August 28th 2019). Bayesian Inference of Gene Regulatory Network, Bayesian Inference on Complicated Data, Niansheng Tang, IntechOpen,
46. Sauta, E., Demartini, A., Vitali, F. et al. A Bayesian data fusion based approach for learning genome-wide transcriptional regulatory networks. BMC Bioinformatics 21, 219 (2020),
47. Yuan Y, Savage RS, Markowetz F (2011) Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. PLoS Comput Biol 7(10),
48. Qazlan, Tuqyah & Hamdi-Cherif, Aboubekeur & Kara-Mohamed, Chafia. (2015). State of the Art of Fuzzy Methods for Gene Regulatory Networks Inference. The Scientific World Journal. 2015. 1-11. 10.1155/2015/148010,
49. Raviajot Kaur, Abhishek Bhola, Shailendra Singh, A novel fuzzy logic based reverse engineering of gene regulatory network, Future Computing and Informatics Journal, Volume 2, Issue 2, 2017, Pages 79-86,
50. Ali, Shaker & Mutlag, Wamidh. (2018). Early detection for breast cancer by using fuzzy logic. Journal of Theoretical and Applied Information Technology. 96. 5717-5728,

51. Scrobotă, I., Băciuț, G., Filip, A. G., Todor, B., Blaga, F., & Băciuţ, M. F. (2017). Application of Fuzzy Logic in Oral Cancer Risk Assessment. Iranian journal of public health, 46(5), 612–619,
52. Sizilio, G. R., Leite, C. R., Guerreiro, A. M., & Neto, A. D. (2012). Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. Biomedical engineering online, 11, 83.