Investigation of the analysis of wearable data for cancer-specific mortality prediction in older adults

Salvatore Tedesco *, Martina Andrulli, Markus Åkerlund Larsson, Daniel Kelly, Suzanne Timmons, Antti Alamäki, John Barton, Joan Condell, Brendan O'Flynn, Anna Nordström

***Abstract*—Cancer is an aggressive disease which imparts a tremendous socio-economic burden on the international community. Early detection is an important aspect in improving survival rates for cancer sufferers; however, very few studies have investigated the possibility of predicting which people have the highest risk to develop this disease, even years before the traditional symptoms first occur. In this paper, a dataset from a longitudinal study which was collected among 2291 70-year olds in Sweden has been analyzed to investigate the possibility for predicting 2-7 year cancer-specific mortality. A tailored ensemble model has been developed to tackle this highly imbalanced dataset. The performance with different feature subsets has been investigated to evaluate the impact that heterogeneous data sources may have on the overall model. While a full-features model shows an Area Under the ROC Curve (AUC-ROC) of 0.882, a feature subset which only includes demographics, self-report health and lifestyle data, and wearable dataset collected in free-living environments presents similar performance (AUC-ROC: 0.857). This analysis confirms the importance of wearable technology for providing unbiased health markers and suggests its possible use in the accurate prediction of 2-7 year cancer-related mortality in older adults.**

Keywords—Cancer, Electronic Health Records, Mortality, Older Adults, Prediction, Wearables

I. INTRODUCTION

According to the WHO, cancer is associated with a large group of diseases that can start in almost any organ or tissue of the body and which occurs when abnormal cells grow uncontrollably [1]. It was estimated that 18.1 million new cancer cases and 9.6 million cancer deaths occurred in 2018 worldwide [2]. This number is expected to rise due to population ageing [3] worldwide. On average, currently there is a 20% risk of getting a cancer before age 75, and a 10% chance of dying from it [2]. The physical, emotional and financial strain exerted on individuals, families, communities, and health systems by cancer continues to grow globally [1], and large numbers of patients globally do not have access to a timely quality diagnosis and early treatment as a result. Early detection is one of the key factors in improving the survival rates of many types of cancers and, thus, cancer mortality prediction is an essential tool for both individualized disease management and effective health resource allocation [3]. A number of clinical indices or scores have been proposed in the literature to predict mortality for a wide range of cancers e.g. the UCLA Prostate Cancer Index [4], the Skin Cancer Index (SCI) [5], and many more. Also, established standard indices, such as the Carolina Frailty Score (CFI) were linked to cancer-related mortality in older adults [6]. Although these scores are well-established, and easy to use and understand; these models have been mostly built to predict survival and quality-of-life following a cancer-related treatment or surgery, are mostly based on patient-reported outcomes (which show several shortcomings [7]) and cannot be tailored to the individual patient [8].

Machine learning (ML) has the potential to transform several aspects of patient care, and its adoption has seen a rapid growth in health and medicine [9-12]. ML modelling has been generally applied to cancer-related datasets in a number of studies (such as [13-15]). However, despite the new heights in clinical cancer research reached through the use of artificial intelligence, those studies have only investigated aspects related to cancer prognosis (involving predictions of disease recurrence and patient survival following therapies or surgeries), or cancer diagnosis of solid and non-solid tumors [16]. The possibility to predict cancer-specific mortality in an older population, years before the cancer diagnosis even occurred, has not yet been deeply investigated.

Moreover, while standard scores generally rely on laboratory measurements to predict mortality, which can affect a timely prediction (especially in people living in rural areas) [17], very little attention has been paid to date to the possibility of only using non-invasive parameters (including those obtainable from wearable devices) for cancer-specific mortality prediction.

This work aims to develop a ML model able to predict cancer-specific mortality in a general population cohort of healthy older adults based on features including anthropometric variables, physical and lab examinations, questionnaires and lifestyles, as well as wearable data collected in free-living settings. Moreover, a targeted analysis on the impact of the wearable data on the overall model performance was also

This project is co-funded by the European Regional Development Fund (ERDF) under Ireland's European Structural and Investment Funds Programmes 2014-2020. Aspects of this work have been supported in part by INTERREG NPA funded project SenDOC. Aspects of this publication have emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289-P2 INSIGHT-2 which is co-funded under the ERDF. Aspects of this publication were supported by Enterprise Ireland and the Department of Business, Enterprise and Innovation under the DTIF project HOLISTICS.

S. Tedesco, M. Andrulli, J. Barton, B. O'Flynn are with the Tyndall National Institute, University College Cork, Lee Maltings Complex, Dyke Parade, T12R5CP Cork, Ireland.
D. Kelly, J. Condell are with School of Computing, Engineering and Intelligent Systems, Ulster University, Londonderry, UK.
S. Timmons is with the Centre for Gerontology and Rehabilitation, University College Cork, Cork, Ireland.
A. Alamaki is with the Department of Physiotherapy, Karelia University of Applied Sciences, Joensuu, Finland.
M. Akerlund Larsson, A. Nordstrom are with Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden.
A. Nordstrom is also with School of Sport Sciences, UiT The Arctic University of Norway, Tromsø, Norway.
* Corresponding author. e-mail: salvatore.tedesco@tyndall.ie

performed. This manuscript is organized as follows. Section II covers a description of the dataset adopted in this work, as well as the data processing steps and the developed model. The results of the analysis are shown and discussed in Section III, while conclusions are illustrated in Section IV.

II. Methods

A. Dataset

The dataset used in this investigation was provided by the "Healthy Ageing Initiative" (HAI) study [18], conducted in Umeå, Sweden. HAI is an ongoing study conducted at a single clinic in Umeå with the aim of identifying traditional and potentially new risk factors for cardiovascular disorders, falls, and fractures among 70-year-olds. The eligibility criteria for inclusion in the study are residency in the Umeå municipality and an age of exactly 70 at the time of the study. There are no exclusion criteria, and population registers are used for recruitment. The HAI study was approved by the Regional Ethical Review Board in Umeå, Sweden. For this work, the data collected in the period from January 2013 to December 2017 were taken into account. The data collection involved a 3-hour health examination for each participant who were then asked to wear an ActiGraph GT3X+ on the hip for 1 week. For logistic reasons, the ActiGraph data collection was limited to 1 week per subject. Subjects' conditions were longitudinally monitored via population registers to determine which subjects died in the time between the data collection and the end of study date (31st December 2019).

The overall dataset consisted of 156 parameters for 2291 recruited participants. Only 92 subjects (approx. 4%) died in the 2-7 years follow-up period, and of these, 50 died from cancer-related conditions (Table I) and have been included in this analysis. Only the records of the 42 subjects who died because of non-cancer related conditions were excluded from the dataset. All the considered predictive variables have been divided into five main subsets as below:

1) Demographics/Anthropometry: gender, height, weight, hip and waist circumference, Body Mass Index (BMI).

2) Self-report health/lifestyle: medications, past or current medical conditions (e.g., stroke), mental health, tobacco and alcohol consumption, physical activity (via IPAQ).

3) Wearable data: All metrics related to the accelerometer data collected via the ActiGraph over one week (e.g., steps taken, time in light, sedentary, moderate, vigorous activities, energy expenditure, etc.). For the data to be acceptable the minimum wear time per day was 600 min, for at least 4 days. Wearable data from subjects that did not have sufficient wear time were considered as missing data.

4) Laboratory tests: such as systolic-diastolic blood pressure, plasma glucose, heart rate, gait analysis data (i.e., step length, etc.), balance test (sway with full and no vision), hand grip strength non-dominant hand, Timed Up and Go (TUG), etc.

5) Others: All information related to body composition (e.g., bone mass, fat, and lean mass for each body segment, obtained via DXA), cholesterol, and feature engineered variables, i.e., Frailty Index [19] and Mortality Index [20].

It was decided to separate the laboratory data collected into two categories ('Laboratory tests' and 'Others') to separately examine variables that could be potentially obtained via wearable technology (i.e., gait analysis [21-22]) and variables not obtainable via wearables (i.e., DXA). Once these five categories above were identified, different subset combinations were evaluated and their results compared. The considered combinations were:

- Case 0: All features;
- Case 1: Demographics/Anthropometrics and self-report health and lifestyle data;
- Case 2: Demographic/Anthropometrics, Self-report health and lifestyle data, and Wearables data;
- Case 3: Demographics/Anthropometrics, Self-report health and lifestyle data, Wearables, and Lab tests data;
- Case 4: Demographic/Anthropometrics, Self-report health and lifestyle data, Wearables data, and Others.

TABLE I
Mortality Causes

Cause	Num. of people
Malignant neoplasm of pancreas	11
Malignant neoplasm of bronchus and lung	7
Malignant neoplasm of colon	6
Malignant neoplasm of prostate	3
Malignant neoplasm of liver, intrahepatic bile ducts, or unspecified parts of biliary tract	3
Malignant neoplasm of brain	2
Malignant neoplasm of bladder	2
Malignant neoplasm of ovary	2
Malignant neoplasm of rectum	2
Malignant neoplasm of breast	2
Malignant neoplasm of skin	2
Mesothelioma	2
Malignant neoplasm of stomach	1
Malignant neoplasm of renal pelvis	1
Multiple myeloma and malignant plasma cell neoplasms	1
Malignant neoplasm without specification	3

B. Ensemble Model

Ensembles have shown promising results when dealing with imbalanced problems and have been often applied on medical data. In this work, the dataset was split into four partitions: training, validation, test and hold-out sets. The hold-out was obtained from the 30% of the whole available data, while the remaining 70% was split again into 50%-25%-25% assigned to the training, validation, and test sets, respectively. The splitting was stratified in order to guarantee that the proportion between positive (subjects who died due to cancer-related conditions) and negative (subjects who did not die) cases was the same in every set. The different types of cancer were not treated differently when building the model. Every continuous feature was standardized by estimating the mean and the standard deviation of each feature and with the normalized variable obtained by subtracting the feature mean and dividing it by its standard deviation. The means and standard deviations were calculated for the training set and then used on the other sets to avoid any possible leakage. In case of missing entries in the dataset, imputation was carried out using the feature mean (estimated on the training set).

The training data was fed to the Forward Selection Component Analysis (FSCA) [23] for feature selection. FSCA can also be successfully adopted to build interpretable and robust systems for anomaly detection. This is possible because FSCA works differently from other feature selection techniques, since it focuses on selecting those features that are able to discriminate more easily between the two different

classes. Moreover, an Isolation Forest (contamination level set to 0.1, 50 decision trees) was applied on the training set to remove possible outliers.

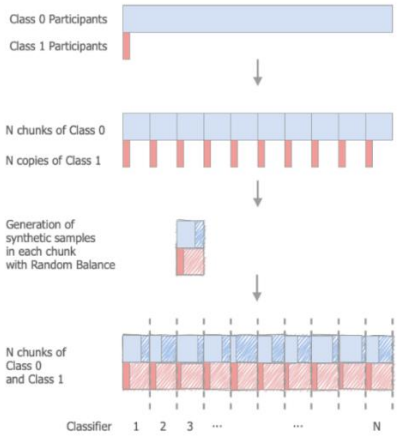

Fig 1. Random Balance graphical representation. Pseudocode in [24]

Following the data pre-processing, the algorithm separates the positive (Class 1) and negative (Class 0) samples in the training set. The samples in the Class 0 training set were split into N different chunks, while N copies of Class 1 samples were generated and each copy of the Class 1 samples was assigned to a different chunk of Class 0 samples, thus generating N different subsets with each subset composed by the same Class 1 samples but different Class 0 samples.

Then, the Random Balance algorithm [24] was applied on each subset in order to generate a set randomly balanced between Class 1 and Class 0 samples. Indeed, each subset differs from the others in terms of a ratio between the number of original and synthetically generated samples (Figure 1), thus increasing the diversity for the learning model. As in [24], SMOTE was adopted for the generation of synthetic samples. Once each subset was properly balanced, its data was used to train a different AdaBoost classifier via a stratified 5-fold cross-validation. As a result, in this ensemble model, N different AdaBoost classifiers have been used, each one properly trained on a different training subset. Each classifier's hyper-parameters were tuned by means of the validation set, to prevent over-fitting.

The performance of each of the N classifiers was evaluated on the test set. The accuracy reported on the test set by each classifier was used as a weight for that classifier at prediction time. After the training of the N AdaBoost classifiers and their individual evaluation on the test set, the whole model performance was evaluated on the hold-out set, with the predictions of each single classifier weighted based on the accuracy computed in the test set.

III. Results and Discussion

The results achieved using the presented ensemble model are reported in this section. The scoring metric utilized to optimize the overall model performance is the AUC-ROC; however, given the highly imbalanced dataset available, other useful metrics (AUC-PR, Brier score, F1 score, accuracy, precision, and recall) are also provided for evaluation. The number of features selected by FSCA was changed properly, together with the models' hyper-parameters, in order to prevent over-fitting, while the number of subsets N was set to 10. Furthermore, the analysis has been repeated six times with different data split across data sets to show the repeatability of the model performance. The results in this section are reported as the mean performance for every metric considered as well as its 95% confidence interval (C.I.). P-values were also calculated for pairwise comparisons of cases based on the estimated difference between group means, as indicated in [25]. The achieved results on the hold-out set are reported in Table II. As expected, the highest AUC-ROC (0.882) is obtained in Case 0, where all features are taken into account, while the lowest performance (AUC-ROC: 0.533) is shown in Case 1, as it is the combination with the least number of features. The difference between the two cases is statistically significant (p-value <0.001). However, when also considering the features from the wearable data (Case 2), the model performance is increased up to 0.857 which is comparable to the original case, e.g. Case 0 (p-value: 0.207). Interestingly, though, when lab tests data was included in the model, AUC-ROC significantly decreased to 0.67 (p-value <0.001 when compared to both Case 0 and 2), showing the detrimental effect of those features on the model. This can be explained by the fact that features selected via FSCA were not sufficiently representative of the characteristics able to differentiate between the two classes despite maximizing the amount of information available. If, instead of the lab tests data, the variables indicated in the 'Others' category were included, the performance increased to 0.875, therefore in between Case 0 and Case 2 (p-value: 0.362 when compared to Case 2, p-value: 0.437 when compared to Case 0). Moreover, AUC-PR was also comparable between Case 0, Case 2, and Case 4 (0.516-0.54); however, AUC-PR was much lower than AUC-ROC for every studied case (highlighting a possible high number of false positives) which is a behavior already seen in literature on similar datasets (i.e., MIMIC-III [26]).

TABLE II
Model Performance on Hold-Out Set

Model	AUC-ROC	AUC-PR	Brier Score	F1 Score	Accuracy	Recall	Precision
Case 0	0.882 (0.870 - 0.896)	0.540 (0.523 - 0.550)	0.218 (0.199 - 0.234)	0.169 (0.159 - 0.186)	0.783 (0.767 - 0.803)	0.987 (0.960 - 1.000)	0.093 (0.085 - 0.102)
Case1	0.533 (0.505 - 0.567)	0.259 (0.192 - 0.312)	0.415 (0.296 - 0.510)	0.048 (0.043 - 0.058)	0.585 (0.495 - 0.677)	0.480 (0.326 - 0.600)	0.026 (0.023 - 0.030)
Case 2	0.857 (0.814 - 0.887)	0.516 (0.476 - 0.544)	0.229 (0.206 - 0.252)	0.156 (0.147 - 0.172)	0.771 (0.748 - 0.794)	0.947 (0.893 - 1.000)	0.085 (0.081 - 0.092)
Case 3	0.670 (0.632 - 0.713)	0.204 (0.129 - 0.273)	0.229 (0.194 - 0.265)	0.062 (0.038 - 0.076)	0.771 (0.737 - 0.802)	0.360 (0.206 - 0.474)	0.034 (0.021 - 0.041)
Case 4	0.875 (0.864 - 0.887)	0.532 (0.519 - 0.545)	0.220 (0.202 - 0.236)	0.165 (0.156 - 0.178)	0.780 (0.762 - 0.800)	0.973 (0.946 - 1.000)	0.091 (0.085 - 0.097)

This model behaviour highlights a series of considerations. It is well-known that body composition (e.g., lean body muscle mass and levels of adipose tissue) are linked to all-cause mortality and cancer-specific mortality [27-29]. The results obtained in Case 0 and Case 4 confirm mortality predictors reported in the clinical literature, thus highlighting even more the possible effects of adiposity and body composition on cancer mortality. It is also well-known that obesity and low levels of physical activity are associated with an increased risk of mortality and that, especially in older adults, exercise and increased fitness promote positive changes in body composition, reducing the risk for adverse events in the aging population [30]. Again, results in Case 2 confirmed this view and suggested that objective physical activity-related metrics obtainable from a wearable accelerometer worn by the study participants over 1 week in free-living environments, in conjunction with demographics data, may predict subsequent cancer-specific mortality in older adults. While wearables have been recently used in oncology trials to predict clinical outcomes in patients undergoing specific treatments [31], not many studies have investigated the possibility to predict cancer-related mortality in older people even several years before symptoms occur. In a recent paper (2020), Smirnova et al. [32] observed that objective accelerometry-derived physical activity measures outperformed traditional predictors of 5-year all-cause mortality. This paper, therefore, has confirmed the importance of wearable technology for providing unbiased health markers and has further acknowledged wearables' use in the possible accurate prediction of 2-7 year cancer-related mortality in older adults. Further analysis are required to investigate which features in the considered subsets are most predictive of cancer-related mortality. Moreover, further analysis are required to benchmark the performance of the developed ensemble model against standard approaches in literature.

IV. Conclusion

Cancer is an aggressive disease with a tremendous socio-economic burden on the community. In this paper, a dataset from a longitudinal study collected among 2291 70-year olds in Sweden has been analyzed to investigate the possibility for predicting 2-7 year cancer-specific mortality. The analysis suggests that a feature subset including demographics, self-report health and lifestyle data, and wearable-related data minimized the AUC-ROC loss against a full-feature model (0.882 vs 0.857), suggesting that clinicians could potentially rely exclusively on easy-to-use, easy-to-collect, and non-invasive data sources. This analysis confirmed the importance and usefulness of wearable technology for providing unbiased health markers and proved its ability to contribute to accurate prediction of future cancer-related mortality in older adults.

References

[1] WHO: https://www.who.int/health-topics/cancer#tab=tab_1

[2] J. Ferlay, et al., "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *Int J Cancer*, 144, 8, 1941-1953, 2019.

[3] G. Carioli, et al., "European cancer mortality predictions for the year 2020 with a focus on prostate cancer,"*Ann Oncol*, 31, 5, 2020.

[4] M. Litwin, et al., "The UCLA prostate cancer index: Development, reliability, and validity of a health-related quality of life measure," *Medical Care*, 36, 7, 1002-1012, 1998.

[5] J.S. Rhee, et al., "The skin cancer index: Clinical responsiveness and predictors of quality of life," *Laryngoscope*, 117, 3, 399-405, 2007.

[6] E.J. Guerard, et al., "Frailty index developed from a cancer-specific geriatric assessment and the association with mortality among older adults with cancer," *J Natl Compr Canc Netw*, 15, 7, 894-902, 2017.

[7] S.P. McKenna, "The limitations of patient-reported outcome measurement in oncology," *J Clin Pathways*, 2, 7, 37-46, 2016.

[8] M.A. Bookman, "Can we predict who lives long with ovarian cancer?," *Cancer*, 125, S24, 4578-4581, 2019.

[9] J.A.M. Sidey-Gibbons, et al., "Machine learning in medicine: A practical introduction," *BMC Med Res Method*, 19, 64, 2019.

[10] S.F. Weng, et al., "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, 12, 4, 2017.

[11] D.S. Komaris, et al., "Predicting three-dimensional ground reaction forces in running by using artificial neural networks and lower body kinematics," *IEEE Access,* 7, 156779-156786, 2019.

[12] S. Tedesco, et al., "Motion sensors-based machine learning approach for the identification of anterior cruciate ligament gait patterns in on-the-field activities in rugby players," *Sensors,* 2020.

[13] P. Gupta, et al., "Prediction of colon cancer stages and survival period with machine learning approach," *Cancers*, 11, 12, 2019.

[14] R.B. Parikh, et al., "Machine learning approaches to predict 6-month mortality among patients with cancer," *JAMA Netw Open*, 2, 2019.

[15] C.M. Lynch, et al., "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *Int J Med Inform*, 108, 1-8, 2017.

[16] S. Huang, et al., "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," *Cancer Letters*, 2020.

[17] G. Zhang, et al., "A machine learning approach for mortality prediction only using non-invasive parameters," *Med Biol Eng Comput*, 58, 10, 2195-2238, 2020.

[18] Healthy Ageing Initiative: https://www.livsmedicin.se/variabellista

[19] D.M. Williams, J. Jylhava, N.L. Pedersen, S. Hagg, "A frailty index for UK Biobank participants," *J Gerontol A*, 74, 4, 582-587, 2019.

[20] N.H. Kim, et al., "Predictive mortality index for community-dwelling elderly Koreans," *Medicine*, 95, 5, e2696, 2016.

[21] W. Tao, et al., "Gait analysis using wearable sensors," *Sensors*, 12, 2, 2255-2283, 2012.

[22] S. Tedesco, et al., "Experimental validation of the Tyndall portable lower-limb analysis system with wearable inertial sensors," *Procedia Eng*, 147, 2016.

[23] L. Puggini, S. McLoone, "Forward selection component analysis: Algorithms and applications," *IEEE Trans Pattern Anal Mach Intell*, 39, 12, 2017.

[24] J.F. Diez-Pastor, et al., "Random balance: ensembles of variable priors classifiers for imbalanced data," *Knowl Based Syst*, 85, 2015.

[25] D. Altman, "How to obtain the P value from a confidence interval," *BMJ*, 343:d2304, 2011.

[26] A.R. Nallabasannagari, et al. "All data inclusive, deep learning models to predict critical events in the medical information mart for intensive care III database (MIMIC-III)," arXiv:2009.01366

[27] B.J. Caan, et al., "The importance of body composition in explaining the overweight paradox in cancer," *Cancer Res*, 78, 8, 2018.

[28] N. Charette, et al., "Prognostic value of adipose tissue and muscle mass in advanced colorectal cancer: A post hoc analysis of two non-randomized phase II trials," *BMC Cancer*, 19, 134, 2019.

[29] D.H. Lee, et al., "Predicted lean body mass, fat mass, and all cause and cause specific mortality in men: Prospective US cohort study," *BMJ*, 362, 2018.

[30] A.S. Ryan, "Exercise in aging: Its important role in mortality, obesity and insulin resistance," *Aging Health*, 6, 5, 551-563, 2010.

[31] G. Gresham, et al., "Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients," *npj Digital Medicine*, 1, 27, 2018.

[32] E. Smirnova, et al., "The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National health and nutritional examination survey 2003-2006," *J Gerontol A*, 75, 2020