Demonstrating Machine Learning for Cancer Diagnostics

Paul Walsh[1], Jennifer Lynch[1], Brian Kelly[1], Cintia Palu[1], Onofrio Gigliotta[2], Raffaele Di Fuccio[2]

[1] NSilico Life Science, Dublin, Ireland
[2] University of Naples, Italy
Corresponding Author: jennifer.lynch@nsilico.com

Abstract. This paper describes how machine learning systems can be explained and demystified for non-technical audiences through the use of an online simulation. This research is the result of a European Union funded project, SageCare, which focuses on developing machine learning systems to classify clinical and genomic data. In disseminating the use of machine learning to non-specialists we often encounter resistance or suspicion on the veracity of approach. Hence, we present artificial intelligence/machine learning for non-specialists and present a case study and an interactive simulation on how machine learning can be used in cancer diagnostics. The simulation system serves as a basis for both informing clinical practitioners how machine learning can be used to build diagnostic models and describes how feedback from users will be gathered and analyzed to assess how machine learning is viewed in such an application.

Keywords: First Keyword, Second Keyword, Third Keyword.

1 Introduction

The SageCare Project [1] tackles the important area of personalized medicine, by addressing health informatics in a holistic way by creating a platform that interlinks spatially distributed clinical care information sources, EHRs and associated genomic sequences, thereby allowing clinicians to make reasoned queries using machine learning over vast knowledge bases of health and research data. This requires a number of disciplines and skills to be brought together in order to achieve success, including clinicians active in the diagnosis and treatment of cancer. To gauge the effectiveness of machine learning in the domain of cancer diagnostics, a JavaScript simulation, based on a simulator developed by [2], is configured to build a machine learning model using a real cancer data set. The simulation serves as a basis of explaining the dynamics of machine learning to potential end users.

2 Cancer Diagnostics

Cancer is one of the diseases which has a huge impact on patients and their families, so understanding how artificial intelligence can be leveraged to aid diagnosis is important in order to help find ways to alleviate the prevalence of this disease. This paper outlines how machine learning driven artificial intelligence (AI) can be used to aid diagnosis of cancer by building a model that assesses visual input features of cell nuclei. It also serves as a useful example to non-specialists interested in AI to help them understand the dynamics of machine learning algorithms and to understand how to assess their performance.

Breast cancer is one of the most commonly occurring cancers, with over 2 million new cases diagnosed globally every year [3]. While around 5% to 10% of cases are due to inherited genes, such as variants of BRCA [4], there is a higher risk of developing this form of cancer linked to lifestyle factors such as alcohol consumption and obesity [5]. For example, overweight women have an increased invasive breast cancer risk versus women of normal weight [5]. However, the major risks associated with this disease are age, due to likelihood of mutations caused by cell division, and gender, as breast cancer mainly affects women. Breast cancer frequently occurs in the cells lining the milk ducts, where it is referred to as ductal carcinomas, and the tissue that produces the milk supplied to these ductal carcinomas, where it is referred to as lobular carcinomas [7]. Diagnosing such carcinomas involves taking a biopsy of cells from the site in question, which may be deep within the breast tissue. Early diagnosis is key to the effective treatment of such cancers, as studies have shown increases in cancer survival due to advances in early detection and treatment [8], so performing an effective assessment is critical. X-rays of the breast known as mammograms are frequently used as a screening method to identify potential cancerous growths, along with physical contact examination to determine if there is a need for further investigation. Suspect tissue is often biopsied using a fine needle aspiration, whereby a narrow hollow needle is inserted into the tissue to collect a sample of cells [9]. An example image of an invasive ductal carcinoma biopsy is given in Figure 1.

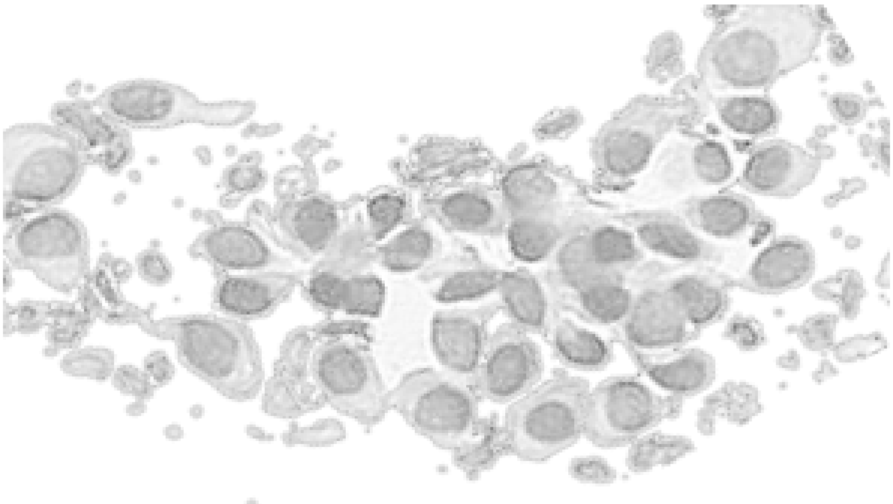


Fig. 1. Image capture of cell features of invasive ductal carcinoma.

These cells are then prepared for examination by a pathologist who examines the characteristics of individual cells, as many different cell features are thought be highly correlated with malignancy [10]. Malignant cells tend to be irregular compared to normal cells, so larger values for features related to shape, such as symmetry, fractal and concavity tend to indicate that the cells are cancerous. It is possible to use machine vision to detect such cell features from biopsies via a digital microscope. This is the basis of the widely studies Wisconsin breast cancer dataset, where 569 biopsies were collected and the following ten geometric features calculated for cells in each of the samples [11]:

1. The radius of the nucleus.
2. The perimeter of the cell nucleus.
3. The area of the cell nucleus.
4. The perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula:

$$\frac{perimiter^2}{area}$$

Cell nuclei that have an irregular shape will have a higher measure of compactness.

5. The smoothness as measured by the difference between the radii across the cell nucleus.
6. The number and severity of concave features around the cell nucleus.
7. The number of concave points around the cell nucleus.
8. A measure of symmetry, sampled at points around the cell nucleus.
9. A measure of the fractal dimension along the cell.
10. The texture of the cell nucleus by measuring the grayscale intensity variation across pixels within the cell nucleus.

The mean, max and standard error of each feature are computed for each image to give a total of 30 input features per sample, which are suitable for machine learning.

3 Machine Learning Approach

Machine learning is a computational approach to AI that uses algorithms that iterate over datasets to build statistical models [12]. Machine learning techniques can be broadly classified as supervised, which use labelled input data to train a model, or unsupervised algorithms that cluster data into related groups. The power of supervised machine learning is the ability to generalise to correctly classify unseen data, based on models built using training data. We use a Support Vector Machine (SVM) to build a machine learning model for the Wisconsin breast cancer dataset, using a portion of the data (80%) for training and the rest for testing the model (20%).

The SVM is a supervised learning algorithm that has been shown to have good performance as a classifier [13]. The SVM Algorithm trains by iterating over a set of labeled samples, which in this case are entries from the Wisconsin breast cancer dataset, which are labelled as either benign or malignant. A good way to explain the operation of machine learning is to use a two-dimensional input feature space as this allows us to more easily visualize the decision boundary that the algorithm produces. Figure 3 shows a number of examples from Wisconsin breast cancer dataset plotting the radius feature on the x-axis against the texture feature on the y-axis. An SVM algorithm finds an optimal decision boundary by finding data points, known as support vectors that maximise the separation between classes.

One approach to gauging the performance of the classifier is to compute the F1 score, which is a useful measure of the level of precision and recall in a machine learning system [14]. Precision is the portion of instances among the classified instances that are relevant, while recall or sensitivity is the fraction of correctly classified relevant instances that have been retrieved over the total amount of relevant instances. An algorithm with high precision over a data set will return more relevant results than irrelevant ones. For cancer diagnosis, this is critical as both false positives and false negative errors should be avoided. In particular, a false negative result should be avoided as the impact could result in missed life-saving treatment. Precision can be thought of as the ratio of correctly classified true positives t_p, over the sum of true positives t_p and falsely classified positive f_p:

$$Precision = \frac{t_p}{t_p + f_p}$$

An algorithm with high recall will classify most of the relevant data correctly and can be thought of as the ratio of correctly classified true positives t_p, over the sum of true positives t_p and false negatives f_n (the number of instances falsely classified as negative instances):

$$Recall = \frac{t_p}{t_p + f_n}$$

There is a trade-off between precision and recall as it is possible to have an algorithm with high precision but low recall and vice versa. For example, the algorithm may be precise by correctly classifying a subset of malignant breast cancer cases, however it could achieve this by being stringent in its classification and could exclude many other malignant cases, which would give it a low recall.

The balance between precision and recall can be captured using an F1 score which is the harmonic mean of the precision and recall scores, where a score of 1 indicates perfect precision and recall [15].

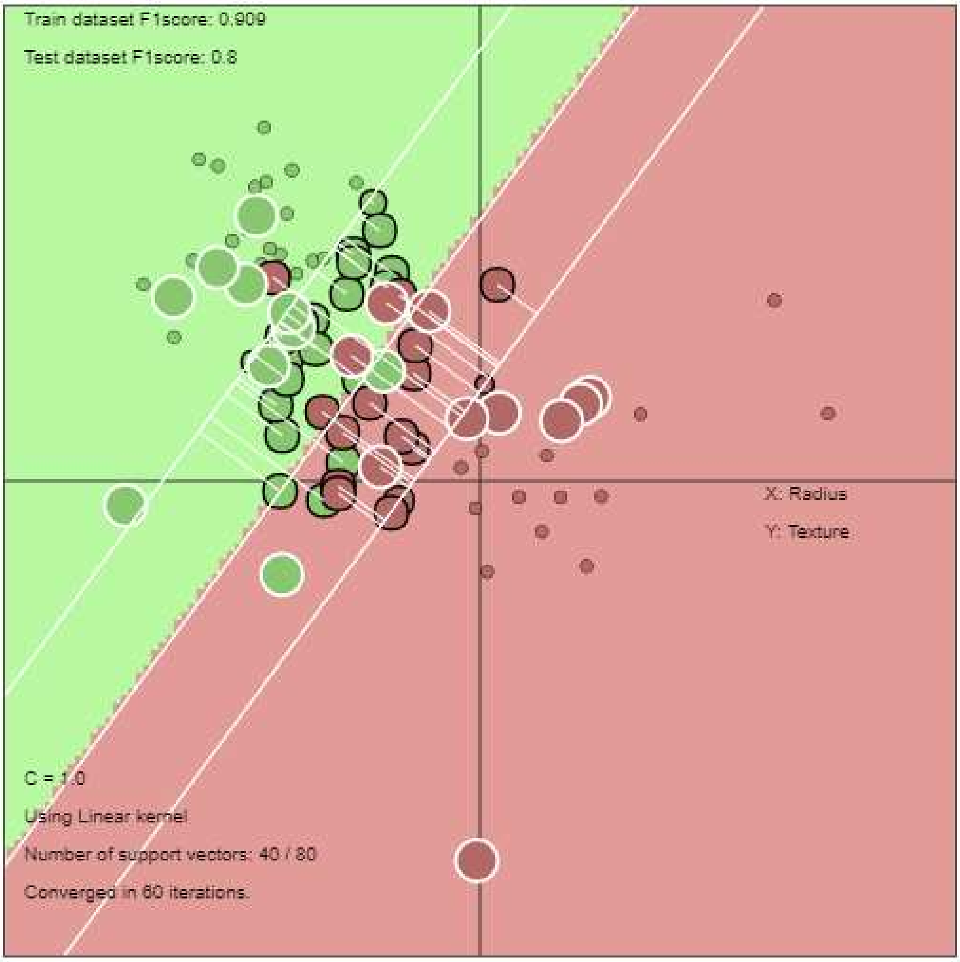
$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

The machine learning model should be trained in such a way that the algorithm does not overfit, which occurs when the algorithm fits a decision boundary tightly to all of the data, including any noise in the training data, so that it generalises poorly to any unseen input. To avoid over-estimation of model performance, a test data set is held back and is used as the final unbiased measure of the algorithm’s performance on the training data. A model that produces a high score on the training set but a low score on the test set will overfit, while a model that produces a high score on the training set and a high score on the test set should provide good classifications. A model that underfits, by failing to find any useful decision boundary will perform poorly on both data sets.

The simulation in Figure 3 shows the F1 scores for the algorithm on the training set and the test set, thereby allowing users to gauge the performance of the algorithm. This also challenges the user to investigate how tuning the hyper-parameters for a machine learning algorithm affects its performance and so will enhance their understanding of the dynamics of the problem.

In the example demonstrated to the non-specialist audience a two-dimensional feature space was presented. The simulator developed by Karpathy was enhanced to allow users to select which feature they would like to evaluate. Figure 3 shows use of a linear kernel when trying to find the ideal separation, and the F1 score for both the training set and test set is shown. The choice of using a kernel is an important machine learning hyperparameter; practitioners needs to consider if the data set is linearly separable or not. This simulation is presented as a game to the users, where the goal is to reach a perfect F1 score of 1 on the test data set.

The use of a nonlinear kernel is shown in Figure 3. Choosing a non-linear kernel for a linear data set will tend to cause the model to over fit the data, which will reduce its ability to generalize as indicated by a poor performance on the test data set F1 score. SVMs use a technique known as the kernel trick which maps data points to a higher dimensional space where a linear separation may be found [16].



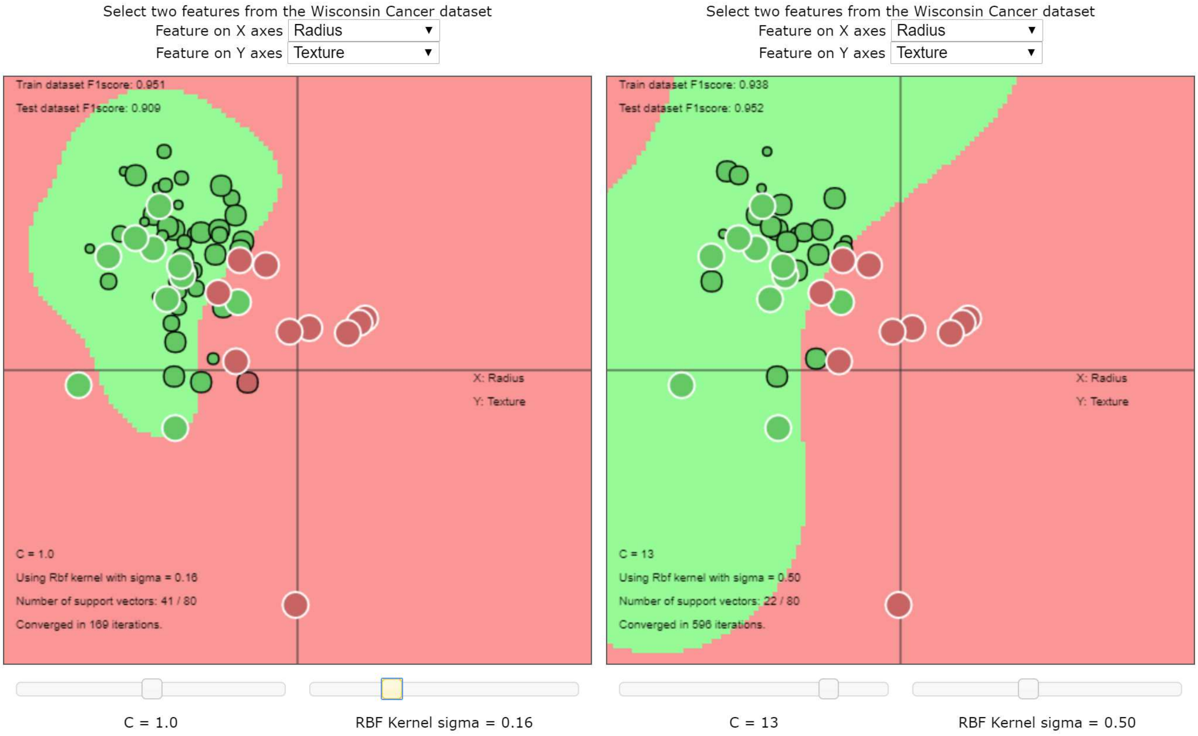
```
Train dataset F1score: 0.909
Test dataset F1score: 0.8
```

```
C = 1.0
Using Linear kernel
Number of support vectors: 40 / 80
Converged in 60 iterations.
```

Fig. 2. An illustration of the performance of a SVM on the Radius and Texture features of the WBCD using a linear mode, based on a fork of an online SVM simulator [2].

A support vector machine can be tuned via a cost function, denoted C, which penalises the algorithm for points that fall within the margin. A small value of C, imposes a low penalty for misclassification, thereby allowing a "soft margin", which promotes better generalisation at the risk of lower precision. A large value of C imposes a high cost of misclassification, thereby producing a “hard margin", which promotes higher precision but poorer generalisation and recall. The JavaScript framework [2] allows users to modify the cost function C and are challenged to find a balance that maximises the F1 score.

For non-linear kernel the Karpathy SVM JavaScript framework uses a Gaussian radial basis function, which allows the SVM algorithm to fit the maximum margin separating hyperplane in a transformed input feature space. The radial basis function is controlled by the parameter sigma (σ), which determines the influence that feature vectors have on the kernel mapping. Intuitively, low values of sigma narrow the region of influence of the kernel for vectors in the feature space, which can cause the SVM to overfit the data. High values of sigma widen the region of influence, making the algorithm better at generalizing at the expense of losing precision. Users can experiment with features, kernels and hyper-parameters as shown in the adaption of Karpathy’s software, see Figure 3 (b). Communicating the effect of σ and other parameters, to non-technical audiences in a visual manner supports the objective of this study; to investigate how an interactive tool enhances their understanding and user of machine learning tools.



Select two features from the Wisconsin Cancer dataset
Feature on X axes Radius
Feature on Y axes Texture

Train dataset F1score: 0.951
Test dataset F1score: 0.909

C = 1.0
Using Rbf kernel with sigma = 0.16
Number of support vectors: 41 / 80
Converged in 169 iterations.

C = 1.0
RBF Kernel sigma = 0.16

Select two features from the Wisconsin Cancer dataset
Feature on X axes Radius
Feature on Y axes Texture

Train dataset F1score: 0.938
Test dataset F1score: 0.952

C = 13
Using Rbf kernel with sigma = 0.50
Number of support vectors: 22 / 80
Converged in 596 iterations.

C = 13
RBF Kernel sigma = 0.50

Fig. 3. Gaussian radial basis function non-linear kernel with (a) sigma = 0.16, C=1.0, causing overfitting and (b) with sigma = 0.5, C=13 improving generalization[1].

[1] Available at http://www.nac.unina.it/svm2/sage_care_svm_two_screen.html

4 Understanding AI

In the last decade, we have witnessed a growing interest on AI applications. Numbers of commercials on a variety of everyday objects (e.g. mobile phones, vacuum cleaners, thermostats etc.), present AI as an important added value. Indeed, it is. With AI powered cameras we can get better photographs, with an autonomous vacuum cleaner we can gain more spare time and with a smart thermostat we can save a lot of money on heating bills.

However, despite this growing interest, very often AI is perceived, by the general public, like a sort of magic, or, to put it in the words used by Arthur C. Clarke, an advanced technology indistinguishable from magic. But such perception has a problem in that it presents a technology's inner mechanisms as being incomprehensible.
Rather, AI is a powerful tool that can be harnessed for personal or professional purposes, even by lay people thanks to off-the-shelf software packages in which users are requested only to add their data. Adding data alone, however, could simplify the process too much, preventing users from grasping the inner mechanics of what they are using, leading to potential mistakes or misuse.

AI, undoubtedly, represents an effective tool to solve or to simplify many relevant problems in our daily lives. For this reason, we should disclose as much as possible about how certain algorithms work. Such an operation can be beneficial to improve a basic understanding of a particular algorithm, whether a user is just willing to better grasp a topic or where a user is interested in using that kind of technology with and increased awareness for his own purposes. This can provide the following benefits of explaining the technology to potential users as they can:

1) make better use of the tools,
2) better understand the problems it can and cannot solve and
3) make a more informed assessment and evaluation of the produced solution

5 Exploratory focus group

In order to understand which kind of information should be conveyed about an intelligent technology in general and in particular to a classifier system such as that presented in this paper, we organized a focus group, held in Rome in December 2018, with a small number of participants. Focus groups have a long tradition in behavioral sciences where have been used to understand how an issue or a product is perceived by a group of people [17].

Seven participants, 1 woman and 6 men, with an average age of 37.57 (SD= 6.23) with a background in AI research were identified within an Italian research center. For the aim of this exploratory work, participants were chosen for their unfamiliarity with support vector machines although well-versed in other AI techniques such as neural networks, genetic algorithms etc. The focus group was organized with the following structure:

1) Short introduction on the topic of the focus group
2) Brief presentation of the participants

3) Discussion of the topic:
- Question 1: How do you think people perceive AI?
- Question 2: Do you think it is necessary to explain how AI works?
- Question 3: Which features of a classifier system should be stressed for educational purposes?

4) Presentation of SVMs through the software we developed
5) Request of feedback on the software as an educational tool

6 Qualitative results

The focus group lasted three hours and stimulated a very interesting discussion on the general importance of AI in our lives and the features that should be shared in order to increase people's awareness on specific algorithms.

The first question raised a dualistic view on AI. To the participants, people seem to see AI either as the ultimate evil or magic. Both polarized views, however, lack a realistic perspective and all the participants agreed with the fact that AI at the moment is an inflated word due to marketing purposes.

The second question divided the participants. Three of them underlined that in order to understand many AI algorithms a strong background in maths is needed, hence it is impossible to provide such kinds of concepts to a general public, who likely lack specific hard skills. The rest of the group in different ways highlighted the need to explain how algorithms work. In particular, two proposals emerged on how best to explain how AI works to members of the general public: a) using a very simple language without referring to mathematical jargon; and b) demonstrating algorithm with micro-educational software in which users can manipulate data and parameters.

The final question firstly collected a series of answers related to the fact that the outcome of classifiers system, regardless of the algorithm that is being used, is strictly connected to the data we put in. Secondly, although sometimes very complicated math is required to understand the specific aspects of an algorithm, an extremely simple formula can often be used to evaluate the outcome of a classifier (see for example precision and recall).

After the discussion raised by the first three questions, we presented our software (by explaining the objectives and the algorithm behind it) to the participants and asked them if it was, in their opinion, viable for educational purposes.

Participants appreciated two aspects of the software: 1) that is web-based and it is able to seamlessly run on mobile phones without issues, and 2) the two windows easily allow for seeing what happens to the outcome when different parameters are applied to the underlying algorithm. Less appreciated was the graphical aspect. Participants suggested to improve the graphics in order to make the training and the testing sets more visible.

An overall positive consideration emerged about the possible use of the software as an educational tool. Although not experts in SVMs, they understood how this type of classifier works in a relatively simplified way (here we remember that the participants shared an AI background)

7 Future research

AI and its applied arm, machine learning, are becoming an important part of our daily life. Our mobile phone recognizes our vocal commands and the AI powered cameras can take pictures with a professional quality. Applications, however, are not limited to our spare time. AI can also be added to the toolkit of our professions: a biologist or a psychologist, for example, could exploit a machine learning solution for their own purposes. In particular, a biologist could use a classifier such as the one described in this paper to classify his/her own data points or re-run the algorithm with new collected data. In order to do that it is not required to be a data scientist but just to be able to use an off-the-shelf solution with a proper awareness.

Qualitative data collected in an exploratory focus group seems to suggest that our approach goes in this direction, however, in order to evaluate its effectiveness we need quantitative data. Gathering this quantitative data will be the objective of the next step of this research.

Acknowledgements

PW is supported by funding through Science Foundation Ireland (SFI) Multi-Gene Assay Cloud Computing Platform - (16/IFA/4342), BK, JL, CP, OG and RDF are funded under H2020 RISE project SemAntically integrating Genomics with Electronic health records for Cancer CARE (SageCare), grant number 644186.

References

1 S. Consortium, "Periodic Reporting for period 1 - SAGE-CARE (SemAntically integrating Genomics with Electronic health records for Cancer CARE)," H2020 CORDIS European Commission, 2014.

2 A. Karpathy, "Support Vector Machine in Javascript," [Online]. Available: https://cs.stanford.edu/people/karpathy/svmjs/demo/. [Accessed August 2018].

3 T. Vos and e. al., "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.," *The Lancet,* vol. 388, no. 10053, pp. 1545-1602, 2016.

4 M.-C. King, J. H. Marks and J. B. Mandell, "Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2," *Science ,* vol. 302, no. 5645, pp. 643-646, 2003.

5 L. M. Morimoto, E. White, Z. Chen, R. T. Chlebowski, J. Hays, L. Kuller, A. M. Lopez, J. Manson, K. L. Margolis, P. C. Muti, M. L. Stefanick and A. McTiernan, "Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States)," *Cancer Causes & Control,* vol. 13, no. 8, p. 741–751, 2002.

6 P. R. Marian L. Neuhouser, M. Aaron K. Aragaki, P. Ross L. Prentice and e. al, "Overweight, Obesity, and Postmenopausal Invasive Breast Cancer Risk," *JAMA oncology,* vol. 1, no. 5, pp. 611-621, 2015.

7 P. A. T. E. B. P. B. C. T. B. M. N. C. I. NCI, "Breast Cancer Treatment (PDQ®)–Patient Version," 2019. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq. [Accessed 2 2 2019].

8 K. D. M. MPH, R. L. S. MPH, M. Chun Chieh Lin PhD, A. B. M. PhD, J. L. K. MD, J. H. R. PhD, K. D. S. PhD, R. A. MD and P. Ahmedin Jemal DVM, "Cancer treatment and survivorship statistics, 2016," *CA: A Cancer Journal for Clinicians,* vol. 66, no. 4, pp. 271-289, 2016.

9 M. Wu and D. E. Burstein, "Fine needle aspiration.," *Cancer investigation,* vol. 22, no. 4, pp. 620-628, 2004.

10 W. Nick Street, W. H. Wolberg and O. L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," Biomedical Image Processing and Biomedical Visualization, 1993.

11 W. N. e. a. Street, "Breast Cancer Wisconsin (Diagnostic) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) . [Accessed August 2018].

12 A. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development,* vol. 3, no. 3, p. 210–229, 1959.

13 C. Cortes and V. N. Vapnik, "Support-vector networls.," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

14 D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies,* vol. 2, no. 1, pp. 37-63, 2011.

15 Y. Sasaki, "The truth of the F-measure.," *Teach Tutor mater 1,* vol. 1, no. 5, pp. 1-5, 2007.

16 B. E. Boser, I. M. Guyon and V. N. Vapnik, "A training alorithm for optimal margin classifiers," Proceedings of the fifth annual workshop on Computational learning theory. , 1992.

17 D. W. Stewart and P. N. Shamdasani, Focus Groups: Theory and Practice, 3rd ed., Sage, 2015.

18 M. C. Pike, D. V. Spicer, L. Dahmoush and M. F. Press, "Estrogens progestogens normal breast cell proliferation and breast cancer risk.," *Epidemiologic reviews,* vol. 15, no. 1, pp. 17-35, 1993.