Diagnosis Kanker Payudara Menggunakan Machine Learning Dengan Algoritma K-Nearest Neighbor

Chalifa Chazar[1], Indra Nursyamsi[2], Patah Herwanto[3]
[1,2,3] Program Studi Teknik Informatika, STMIK Indonesia Mandiri
Bandung, Indonesia
chalifa.chazar@gmail.com, Indra.syamsi95@gmail.com, pherwanto@gmail.com

ABSTRAK

Banyak anggapan bahwa kanker payudara sama seperti tumor, pada kenyataanya tumor yang muncul tidak berarti kanker payudara. Penyakit ini sulit didiagnosis pada tahap awal, hal ini menyebabkan banyak penderita baru mengetahui kondisinya setelah memasuki tahapan yang sulit disembuhkan. Biopsi adalah teknik pemeriksaan dengan mengambil cairan di payudara dengan menggunakan *Fine Needle Aspiration* (FNA), selanjutnya akan dilakukan diagnosis untuk mengetahui jenis kanker apakah termasuk pada kelompok jinak atau ganas. Diagnosis ini membutuhkan waktu yang relative lama. *Machine Learning* (ML) memiliki kemampuan untuk dapat meniru kecerdasan manusia dan terus melakukan pembelajaran berdasarkan data atau pengalaman di masa lalu. Semakin sering digunakan *Machine Learning* (ML) akan menghasilkan tingkat akurasi yang lebih tinggi. Untuk meningkatkan hasil akurasi dari predisksi yang dihasilkan digunakan algoritma K-*Nearest Neighbor* (K-NN). Penelitian ini bertujuan untuk membangun aplikasi *Machine Learning* (ML) untuk mendiagnosis jenis kanker payudara dengan menggunkan algoritma K-Nearest Neighbor. Hasil dari penelitian ini menunjukan bahwa algoritma K-*Nearest Neighbor* (K-NN) dapat menentukan jenis kanker payudara dengan menggunkan sedikit data atau pengalaman dengan hasil akhir yang mudah dipahami.

Kata kunci: Kanker payudara, K-Nearest Neighbor, Machine Learning

ABSTRACT

Many assume that breast cancer is the same as a tumor, in fact a tumor that appears does not mean breast cancer. This disease is difficult to diagnose at an early stage, this causes many sufferers to find out their condition only after entering a stage that is difficult to cure. Biopsy is an examination technique by taking fluid in the breast using Fine Needle Aspiration (FNA), then a diagnosis will be made to determine whether the type of cancer is in the benign or malignant group. This diagnosis takes a relatively long time. Machine Learning (ML) can

imitate human intelligence and continue to learn based on data or past experiences. The more frequently used Machine Learning (ML) will result in a higher level of accuracy. To improve the accuracy of the resulting predictions, the K-Nearest Neighbor (K-NN) algorithm is used. This study aims to build a Machine Learning (ML) application to diagnose the type of breast cancer using the K-Nearest Neighbor algorithm. The results of this study indicate that the K-Nearest Neighbor (K-NN) algorithm can determine the type of breast cancer by using a little data or experience with easy-to-understand results.

***Key words**: Breast Cancer, K-Nearest Neighbor, Machine Learning*

Pendahuluan

Kanker payudara adalah sebuah kondisi ketika sel kanker terbentuk di jaringan kelenjar yang menghasilkan susu (*lobulus*), atau di saluran (*duktus*) yang membawa air susu dari kelenjar ke puting payudara, kanker juga bisa terbentuk di jaringan lemak atau jaringan ikat di dalam payudara (Chazar dan Widhiaputra, 2020). *Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women* (Asri et al., 2016). Hal ini disebabkan karena penyakit ini sulit didiagnosis pada tahap awal, banyak penderita baru mengetahui kondisinya setelah memasuki tahap/fase dengan risiko kematian yang lebih tinggi. Tumor yang muncul tidak berarti kanker. Jenis tumor dapat dibedakan berdasarkan pertumbuhanya menjadi tumor ganas (*malignant*) dan tumor jinak (*benign*). Tumor ganas ini yang kemudian disebut sebagai kanker.

Machine Learning (ML) adalah pemograman komputer untuk mencapai kriteria/performa tertentu dengan menggunakan sekumpulan data training atau pengalaman di masa lalu (Primartha, 2018). Saat ini ML cukup popular digunakan, karena memiliki kemampuan pembelajaran secara mandiri berdasarkan data training yang disediakan tanpa perlu diprogram berulang-ulang. ML banyak diterapkan pada bidang medis, khususnya untuk memprediksi sebuah hasil atau diagnosis dari suatu penyakit tertentu. Hal ini dibuktikan dengan beberapa penelitian yang telah berhasil menerapkan ML untuk mendiagnosis suatu penyakit tertentu (Chazar dan Widhiaputra, 2020; Redjeki, 2013; Mahdi et al, 2011; Putra dan Laksita Akbar, 2016).

ML membutuhkan sebuah data yang dikenal dengan data training, berdasarkan data training ini kemudian akan ditentukan sebuah pola yang unik untuk menghasilkan suatu prediksi. Beberapa algoritma dan metode dapat digunakan untuk menentukan pola data dalam ML, diantaranya adalah K-*Nearest Neighbor* (K-NN). K-NN memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma lainnya. Berdasarkan penelitian yang telah dilakukan dengan membandingkan tingkat akurasi, algoritma K-NN memiliki tingkat akurasi sebesar

87,4% (Mahdi et al, 2011). Selain memiliki tingkat akurasi yang tinggi, K-NN juga merupakan algoritma yang cukup mudah untuk diimplementasikan karena tidak membutuhkan banyak data training untuk proses pembelajaran. K-NN adalah salah satu algoritma *machine learning* untuk melakukan klasifikasi terhadap objek baru berdasarkan sejumlah *k* tetangga terdekatnya (Primartha, 2018).

Penelitian ini mencoba untuk mendiagnosis kanker payudara apakah termasuk pada tipe ganas atau jinak dengan menggunakan ML untuk memperoleh hasil diagnosis dan juga algoritma K-NN yang memiliki tingkat akurasi yang tinggi, sehingga dapat menghasilkan prediksi yang akurat.

Materi dan Metode

1. Diagnosis Kanker Payudara

Kanker payudara adalah sebuah kondisi ketika sel kanker terbentuk di jaringan kelenjar yang menghasilkan susu (lobulus), atau di saluran (duktus) yang membawa air susu dari kelenjar ke puting payudara, kanker juga bisa terbentuk di jaringan lemak atau jaringan ikat di dalam payudara (Chazar dan Widhiaputra, 2020). Beberapa teknik yang dapat digunakan untuk mendiagnosis jenis kanker payudara, antara lain (1) Biopsi (*Biopsy*) dan (2) Mammografi (*Mammography*) (Chazar dan Widhiaputra, 2020). (1) Biopsi adalah merupakan teknik yang dilakukan dengan mengambil cairan pada payudara untuk dilihat dengan menggunakan *Fine Needle Aspiration* (FNA), selanjutnya hasil dari *Fine Needle Aspiration Biopsy* (FNAB) diperiksa berdasarkan beberapa indikator untuk dapat menentukan jenis kanker. (2) Mammografi adalah teknik pemeriksaan dengan menggunakan sinar *X-Ray* untuk menilai jaringan payudara. Teknik biopsi lebih banyak digunakan karena memiliki akurasi yang lebih tinggi tetapi membutuhkan waktu yang cukup lama.

2. Dataset Breast Cancer Wisconsin

Dataset merupakan himpunan dari data yang berasal berisi informasi-informasi dari masa lampau. Pada penelitian ini dataset yang digunakan adalah *Breast Cancer Wisconsin* Dataset yang dibuat oleh Dr. William, H. Wolberg, W. Nick Street dan Olvi L. Mangasarian. Dataset ini didapatkan dari hasil analisis citra digital massa payudara dengan menggunakan FNA (Chazar dan Widhiaputra, 2020). Dalam dataset ini terdapat 9 indikator utama untuk mendeteksi adanya sel kanker payudara. Penjelasan tentang informasi dataset *Breast Cancer Wisconsin* disajikan kedalam Tabel 1, sebagai berikut.

Table 1. Informasi dataset breast cancer Wisconsin (Wolberg *et al*, 1995)

#	Attribute	Domain
H1	Sample code number	id number

#	Attribute	Domain
H2	Clump Thickness	1 -10
H3	Uniformity of Cell Size	1 -10
H4	Uniformity of Cell Shape	1 -10
H5	Marginal Adhesion	1 -10
H6	Single Epithelial Cell Size	1 -10
H7	Bare Nuclei	1 -10
H8	Bland Chromatin	1 -10
H9	Normal Nucleoli	1 -10
H10	Mitoses	1 -10
H11	Class:	(2 for benign, 4 for malignant)

Breast Cancer Wisconsin memiliki 699 instances (benign: 458 dan malignant: 241), 2 kelas (65.5% ganas dan 34.5% jinak), dan 11 atribut bernilai integer (Asri *et al*., 2016). Dataset ini juga merupakan indikator yang dapat dilihat pada sel hidup untuk mendeteksi adanya kanker payudara, setiap record memiliki sembilan atribut indikator selain Sample Code Number dan Class yang sembilan atribut tersebut dinilai pada skala interval 1 sampai 10, dengan skala 10 merupakan penilaian keadaan paling abnormal, sehingga semakin dekat nilai dari masing-masing atribut sampai ke 10 maka semakin besar kemungkinan terdeteksi *malignant* (ganas) (Chazar dan Widhiaputra, 2020).

3. Algoritma K-NN

K-NN *is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it* (Primartha, 2018). K-NN merupakan algotitma untuk klasifikasi data pada *supervised learning*. K-NN merupakan *lazy learning algorithm*, karena tidak membutuhkan banyak data untuk proses pembelajarannya. Beberapa kelebihan dari algoritma K-NN yaitu:

- K-NN merupakan algoritma pembelajaran yang bersifat nonparametric, sehingga tidak memerlukan asumsi.
- Mudah diimplementasikan.
- Tahan terhadap training data yang bersifat noisy.
- Hasil akhir mudah dipahami atau diterjemahkan.

Langkah-langkah penyelesaian algoritma K-NN, dijelaskan dalam bentuk flowchart seperti yang diilustrasikan pada Gambar 1.

Gambar 1. Algoritma K-NN

Hasil dan Pembahasan

1. Proses Data *Training*

Langkah awal untuk membangun machine learning adalah dengan menentukan data training. Pada penelitian ini data *training* yang digunakan adalah Dataset Breast Cancer Wisconsin.

Tabel 2. Data Training

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
1	9	5	5	2	2	2	5	1	1	4
2	3	1	1	1	2	2	7	1	1	2
3	7	4	6	4	6	1	4	3	1	4
4	3	1	1	1	2	1	3	2	1	2
5	2	1	1	2	2	1	3	1	1	2
6	5	2	3	4	2	7	3	6	1	4
7	3	2	1	2	3	1	4	2	1	**?**

2. Menentukan Nilai *k*

Menentukan nilai k, yaitu jumlah tetangga terdekat yang digunakan sebagai pembanding. Jumlah data training terdiri dari 6 data training dan 1 test data. Untuk itu dipilih nilai $k = 3$.

3. Menghitung Jarak *d*

Sesuai dengan namanya K-NN melakukan klasifikasi berdasarkan faktor kedekatan dengan tetangganya. Perhitungan jarak terdekat (*d*) ini dapat dipecahkan dengan menggunakan metode *Euclidean distance*. Berikut ini adalah rumus dari *Euclidean distance*.

$$d(x,y) = \sqrt{\sum_{i=1}^{k} x_i - y_i{}^2}$$

Table 3, menggambarkan hasil dari perhitungan jarak terdekat (*d*) dengan metode *Euclidean distance*.

Table 3. Hasil Perhitungan Jarak Terdekat (*d*)

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	*d*
1	9	5	5	2	2	2	5	1	1	4	7.93
2	3	1	1	1	2	2	7	1	1	2	1.41
3	7	4	6	4	6	1	4	3	1	4	7.68
4	3	1	1	1	2	1	3	2	1	2	2
5	2	1	1	2	2	1	3	1	1	2	2.23
6	5	2	3	4	2	7	3	6	1	4	8.12

4. Mengurutkan nilai *d* mulai dari nilai terkecil ke nilai terbesar

Selanjutnya melakukan pengurutan (*sorting*) nilai *d* mulai dari nilai terkecil ke nilai terbesar yang bertujuan untuk menghasilkan *Euclidean distance* terdekat. Tabel 4, menggambarkan hasil dari perhitungan jarak terdekat (*d*) setelah diurutkan dari nilai terkecil ke nilai terbesar.

Tabel 4. Hasil Perhitungan Jarak Terdekat (*d*) Setelah Proses Sorting

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	*d*
2	3	1	1	1	2	2	7	1	1	2	1.41
4	3	1	1	1	2	1	3	2	1	2	2
5	2	1	1	2	2	1	3	1	1	2	2.23
3	7	4	6	4	6	1	4	3	1	4	7.68
1	9	5	5	2	2	2	5	1	1	4	7.93
6	5	2	3	4	2	7	3	6	1	4	8.12

5. Mengklasifikasikan Test Data

Berdasarkan nilai perhitungan jarak terdekat (*d*) yang telah di urutkan (*sorting*), selanjutnya dipilih sesuai dengan nilai *k* yang telah ditentukan di awal yaitu, *k* = 3. Table 5, menggambarkan proses klasifikasi berdasarkan nilai kedekatannya.

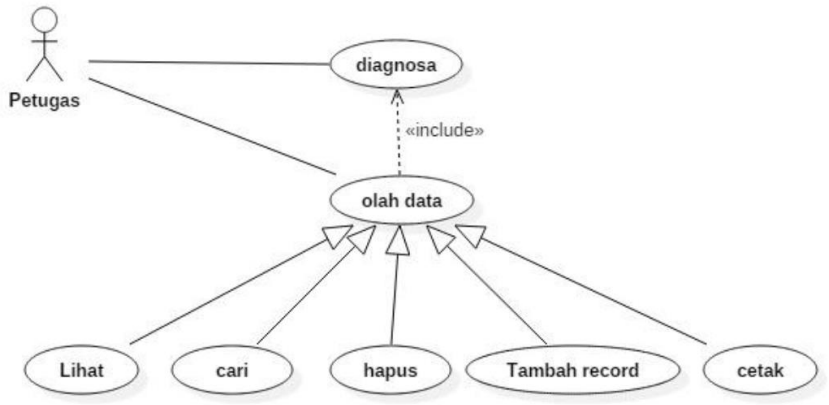
Tabel 5. Hasil Klasifikasi K-NN

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	*D*
2	3	1	1	1	2	2	7	1	1	2	1.41
4	3	1	1	1	2	1	3	2	1	2	2
5	2	1	1	2	2	1	3	1	1	2	2.23
7	3	2	1	2	3	1	4	2	1	**?**	

Dari Tabet 5, dapat disimpulkan bahwa terdapat 3 kelas bernilai 2 yaitu jinak (*benign*). Sehingga dapat disimpulkan bahwa hasil dari data test (index H1 bernilai 7) menunjukan jenis kanker pada kelas jinak (*benign*).

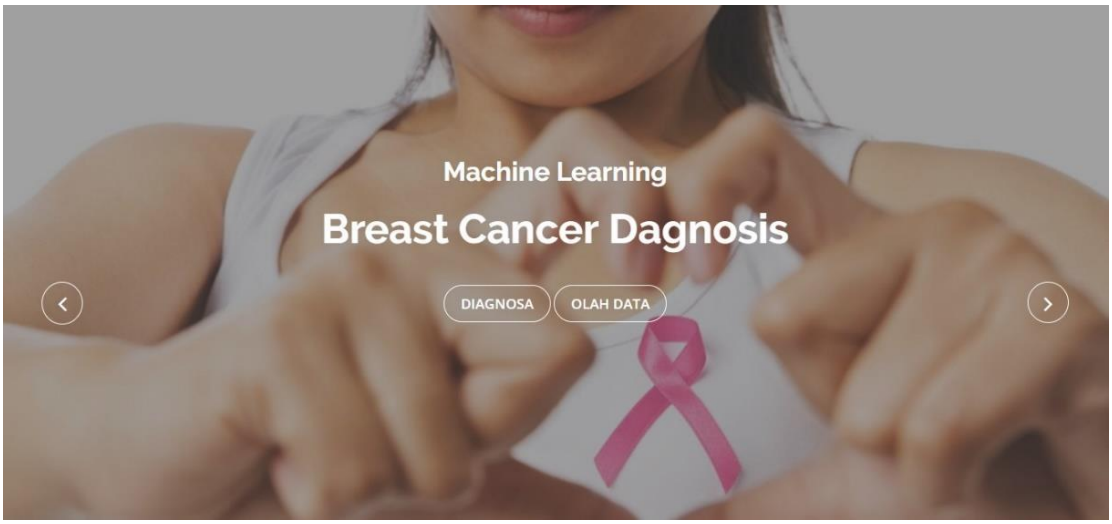
6. Perancangan dan Implementasi

Proses perancangan program menggunakan *Unified Modelling Language* (UML). UML adalah suatu metode dalam pemodelan secara visual sebagai sarana untuk menggambarkan perancangan sistem berorientasi objek. Sasaran pengguna aplikasi ini adalah orang yang berkerja di laboratorium, untuk menginputkan nilai indikator dari hasil FNAB. Gambar 2, mengambarkan *use case* diagram dari aplikasi diagnosis kanker payudara menggunakan ML dengan algoritma K-NN.

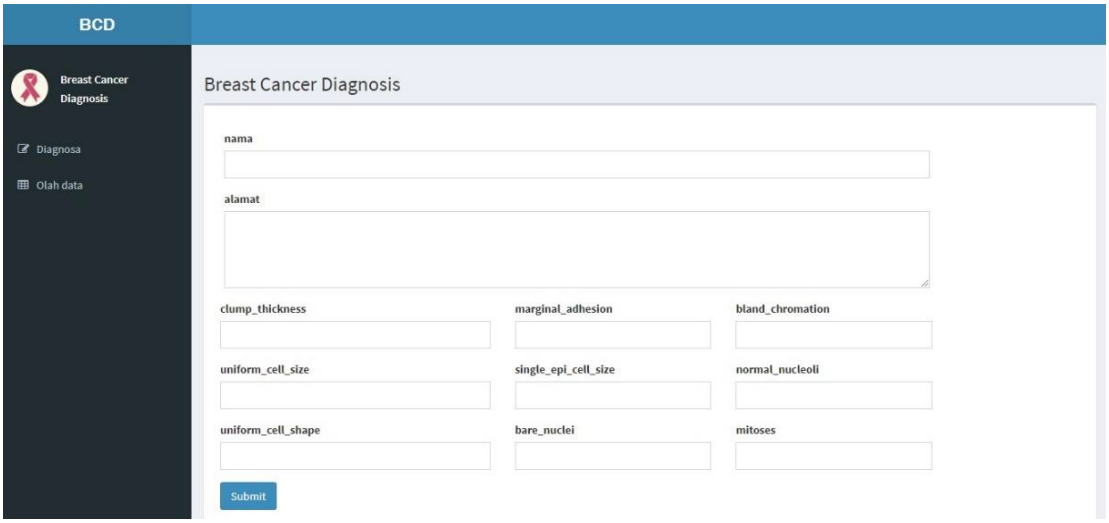


Gambar 2. *Use Case* Diagram

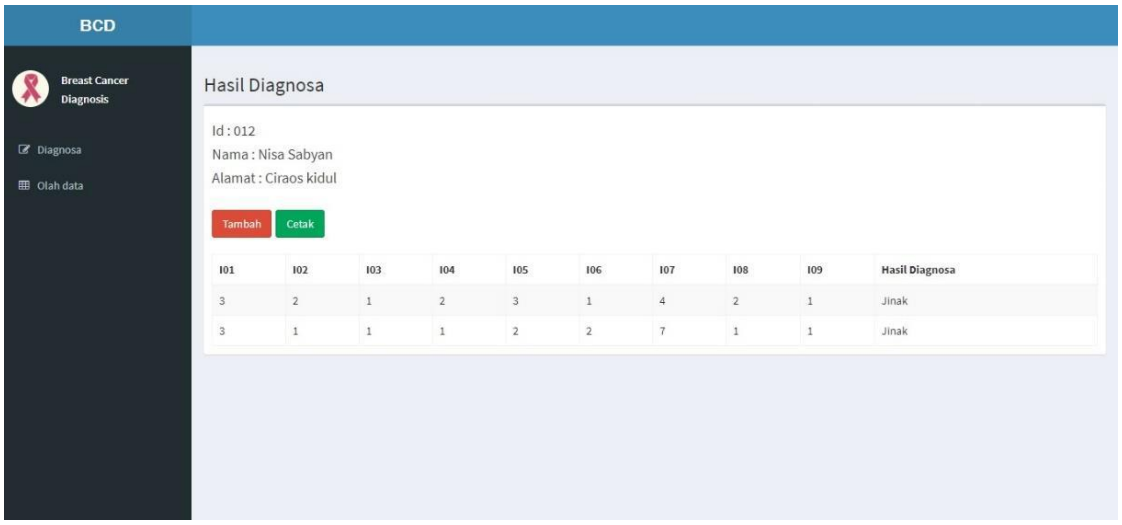
Hasil implementasi aplikasi diagnosis kanker payudara menggunakan ML dengan algoritma K-NN, diimplementasikan dengan menggunakan *website*. Gambar 3, menggambarkan hasil implementasi dari halaman utama aplikasi.



Gambar 3. Implementasi Halaman Utama



Gambar 4. Implementasi Halaman Diagnosis



Gambar 5. Implementasi Halaman Hasil Diagnosis

Kesimpulan

Hasil penelitian menunjukan bahwa aplikasi diagnosis kanker payudara menggunakan ML dapat menghasilkan 2 prediksi keputusan yaitu jenis kanker jinak atau ganas. Penerapan algoritma K-NN pada implementasi aplikasi ML memudahkan dalam proses pengumpulan data training, karena algoritma K-NN tidak membutuhkan banyak dataset untuk proses training. Hasil diagnosis dengan menggunakan algoritma K-NN menunjukan nilai akurasi yang cukup tinggi. Berdasarkan penelitian yang telah dilakukan, maka terdapat beberapa hal penting yang dapat berpengaruh terhadap keputusan dari aplikasi berbasis ML, yaitu (1) akurasi data training yang digunakan; (2) algoritma untuk mengklasifikasikan data. Algoritma K-NN memiliki banyak keunggulan, tetapi juga memiliki kelemahan antara lain, (1) sulitnya penentuan nilai *k* yang optimal; (2) metode perhitungan jarak ikut berperan dalam penentuan hasil yang optimal; dan (3) kebutuhan penyimpanan data yang besar untuk menyimpan dataset.

Daftar Pustaka

Asri, H., Mousannif, H., al Moatassime, H., & Noel, T. 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83 (Fams), 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

Chazar, C., & Widhiaputra, B. E. 2020. Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine. INFORMASI (Jurnal Informatika dan Sistem Informasi). Vol 12 Nomor 1/05/2020. http://ojs.stmik-im.ac.id/index.php/INFORMASI/article/view/48

Mahdi, A., Razali, A., & Alwakil, A. 2011. Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosis. https://www.researchgate.net/publication/266389722

Primartha, R (2018). Belajar Machine Learning. Informatika. Bandung.

Putra, J. A., & Laksita Akbar, A. 2016. Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine dan K-Nearest Neighbour. In Informatics Journal (Vol. 1, Issue 2). http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Redjeki, S. 2013. Perbandingan Algoritma Backpropagation dan K-Nearest Neighbour (K-NN) untuk Identifikasi Penyakit. Seminar Nasional Aplikasi Teknologi Informasi (SNATI). ISSN: 1907-5022

W. N. Street, O. L. Mangasarian, and W.H. Wolberg. Breast Cancer Wisconsin (Prognostic) Dataset. UCI. http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/