*Seminar I*

**IMPLEMENTASI CONVULUTIONAL NEURAL NETWORK PADA DATA HANDSPOSTURE**



***Oleh :***

**FATUR RAHMAN**

**H13116518**

**Pembimbing Utama : Dr. Amran, S.Si., M.Si.  
Pembimbing Pertama : Supri Bin Hj Amir, S.Si., M.Eng.  
Penguji : 1. Dr. Anna Islamiyati, S.Si, M.Si.  
 2. Nur Hilal A Syahrir, S.Si, M.Si.**

**PROGRAM STUDI ILMU KOMPUTER**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**2019**

# PENDAHULUAN

## Latar Belakang

Di era teknologi sekarang , perkembangan yang begitu pesat terjadi di semua bidang ilmu Pengetahuan , termsuk teknologi. Perkembangan teknologi berdampak pada pertumbuhan data mentah secara eksponensial []. World Economic Forum , data digital dunia akan mencapai 44 zettabyte , atau 44 triliun gigabyte pada tahun 2020. Jumlah ini akan terus berkembang hingga lebih dua kali lipat setiap dua tahunnya .

Di era teknologi sekarang, perkembangan ilmu Pengetahuan begitu pesat terjadi di semua lini kehidupan, termasuk teknologi. Perkembangan teknologi berdampak pada pertumbuhan data mentah secara eksponensial.[]. Data digital menurut World Economic Forum akan mencapai 44 Zettabyte, atau 44 triliun gigabyte pada tahun 2020. Jumlah ini akan terus berkembang hingga dua kali lipat setiap dua tahun.

Dari pertumubhan data tersebut , %% persen merupakan data citra , dimana data citra dapat berupa sekumpulan pixel hingga beberapa sekumpulan frame pixel (video). Menurut laporan dari … pada tahu.. permintaan gambar sangat berkembang...

Namun data citra dapat menjadi sebuah data yang tidak menyajikan sebuah Informasi ketika data tersebut tidak diketahui objeknya.. dalam suatu citra biasanya tidak hanya berupa haaya satu objek namun ada beberapa objek yang lain termasuk background.

Citra kedalaman merupakan citra satu kanal dengan nilai jarak sebagai nilai intensitasnya.Citra ini disebut juga sebagai citra 3D[1]. Dalam citra terdapat terdapat sebuah nilai yang merepresentasikan tiap pixel dalam citra. nilai tersebut ditampung dalam sebuah matriks yang besarnya sama dengan jumlah pixelnya. Nilai tersebut berasal dari konsep Red ,Gree, Blue(RGB) pada komputer yang nilainya dari 0 hingga 255.

Handsposture adalah Bahasa isyarat yang direpresentasikan oleh bentuk jari. Beragam pola jari dapat mewakili huruf yang ada sehingga dapat membantu seseorang untuk menyampaikan Informasi. Handsposture biasa digunakan untuk membantu seseorang yang tidak dapat mendengar agar dapat mengetahui sebuah informasi , seperti dalam acara berita terdapat seseorang yang menyampaikan berita menggunakan handsposture, maka dalam hal ini perlu dilakukan penelitian mengenai Klasifikasi data handsposture.

Pengenalan obyek dalam citra merupakan hal yang menantang dalam bidang computer vision. Bagaimana komputer dapat mengenali objek merupakan suatu hal yang dapat membantu manusia pada saat ini. Dataset handsposture mempunyai karakteristik yang membedakan terhadap dataset lainnya,seperti gambar yang mencakup berbagai macam tangan dengan menggunakan kondisi pengcahayaan yang berbeda(asl).

Convolution Neural Network merupakan metode yang trend pada saat ini untuk mengklasifikasikan gambar, dimana telah dilakukan penelitian terkait menggunakan CNN terbukti memiliki akurasi yang tinggi, seperti yang telah dilakukan oleh minae dan alvares yang menggunakan arsitektur VGG-Net dengan data CASIA-Iris-100 yang memiliki 20.000 gambar dan mendapatkan akurasai 99.4 % , CNN adalah teknik pembelajaran mesin yang berhasil mengklasifikasikan berbagai objek, karena CNN dapat meniru sifat korteks visual mamalia(hangat,2016).

Di dalam peneleitian ini metode CNN yang digunakan untuk mengklasifikasikan objek. Difokuskan pada pembuatan model dalam melakukan ekstraksi fitur dan Klasifikasi objek penelitian. Sehingga dilakukan penelitian mengenai implementasi Convulutional Neural Network pada data Handsposture.

## Rumusan Masalah

Berdasarkan latar belakang tersebut,berikut diberikan rumusan masalah pada

penelitian ini,yaitu :

1. Bagaimana implementasi Convutional Neural Network pada data handsposture?
2. Bagaimana tingkat akurasi yang didapatkan dari hasil Convulutional Neural Network?
3. Bagaimana menentukan model dan jumlah filter yang digunakan dalam convolutional neural network untuk data handposture?

## Batasan Masalah

Permasalahan yang dibahas pada penelitian ini memiliki ruang lingkup yang luas, sehingga diberikan batasan-batasan sebagai berikut:

1. Dataset yang digunakan adalah dataset handsposture .
2. Dataset memiliki 37 kelas.
3. Penelitian ini tidak membahas tentang penyakit yang diderita.
4. Data latih di ambil dari ASL

## Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan dari penelitian ini adalah:

1. Mengetahui implementasi Convolutional Neural Network pada data handsposture.
2. Mengetahui tingkat akurasi yang didapatkan dari hasil Convulutional Neural Network?
3. Mengetahui model dan jumlah filter yang efektif yang digunakan dalam convolutional neural network pada data handsposture.

## Manfaat Penelitian

Hasil penelitian ini diharapkan dapat bermanfaat:

1. Sebagai rujukan untuk mengatasi dataset dengan kedalaman gambar yang sangat sering dijumpai.
2. Menjadi sumber informasi mengenai performa Convolutional Neural Network.
3. Menjadi sumber Informasi penggunaan Convulutional Neural Network dengan data handsposture.

# TINJAUAN PUSTAKA

## Penelitian Terkait

## Dataset Handsposture

## Deep Learning

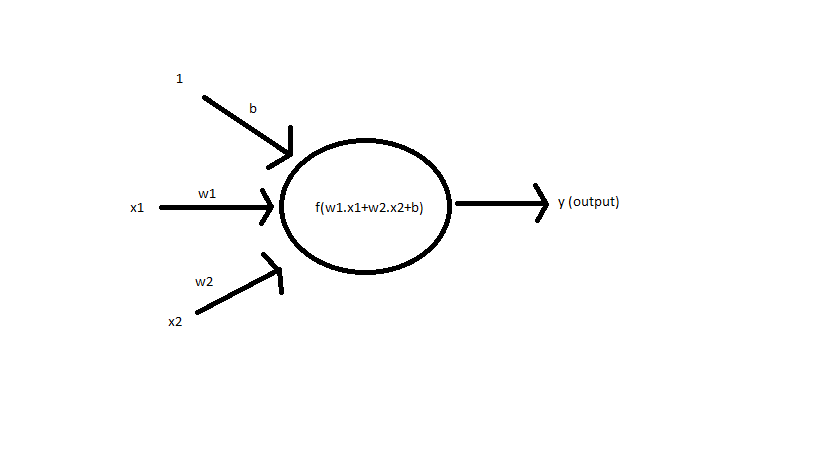
Deep Learning adalah bagian dari machine learning. Deep Learning banyak digunakan untuk kebutuhan analisis data yang mengkombinasikannya dengan kecerdasan buatan. Aplikasi dari penggunaan deep learning banyak seperti Klasifikasi gambar, mobil tanpa pengemudi, klasifikasi berita, dsb[tipus1] dan banyak digunakan oleh perusahaan besar , seperti Google, Microsof , Facebook , IBM , Baidu , Apple , Adobe, Netflix, NVDIA dan NEC [1].

Teknik Deep Learning memberikan model yang bagus untuk algoritma supervised learning. Dengan menambahkan data latih maka model pembelajaran tersebut dapat mewakili kelas yang lebih baik.

## Neural Network

Neural Networks atau jaringan saraf merupakan sebuah model dari jaringan saraf otak manusia yang ditiru oleh banyak peneliti untuk diadopsi cara kerjanya di berbagai bidang kajian seperti biologi, fisika, ilmu komputer, dll [5]. Dibidang ilmu komputer terdapat istilah Artificial Neural Networks (ANN). Konsep ANN diadopsi dari jaringan saraf sehingga ANN biasa disebut jaringan saraf tiruan. Model yang dibuat oleh ANN dapat untuk beradaptasi, belajar dan mengelompokkan berbasis pemrosesan parallel.

Adapun beberapa hal yang mendasar pada komputasi neural network adalah neuron, atau biasa disebut juga dengan node. Node dapat menerima input dari node lainnya atau dari sumber eksternal kemudian dihitung untuk mendapatkan sebuah output. Setiap input memiliki bobot(w) tersendiri, yang mana bobot ini diberikan dengan dasar hubungannya dengan input lainnya. Node tersebut menggunakan fungsi ke dalam input yang telah diberi bobot.



Gambar contoh arsitektur sebuah neuron

Sebuah neuron input berupa X1 dan X2 dengan bobotnya yaitu w1 dan w2. Selain dari input dan bobot juga terdapat inputan lain berupa 1 dengan bobot b(bias).

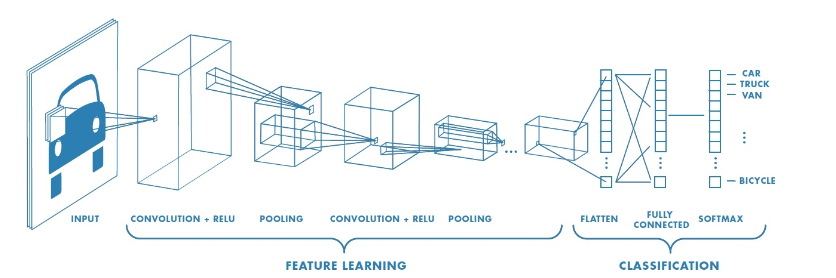
Fungsi utama dari bias adalah untuk menyediakan setiap node dengan sebuah nilai constant yang dapat dilatih( sebagai tambahan selain dari input normal yang diterima)

Output pada neuron tersebut yaitu Y yang merupakan hasil perhitungan dari sebuah fungsi non linear f yang disebut dengan fungsi aktivasi. Tujuan dari sebuah fungsi aktivasi adalah untuk menunjukkan non-linear terhadap output dari neurpn tersebut.

## Convulutional Neural Network

Convolutional Neural Network adalah salah satu metode machine learning dari pengembangan Multi Layer Perceptron (MLP) yang didesain untuk mengolah data dua dimensi. CNN termasuk dalam jenis Deep Neural Network karena dalamnya tingkat jaringan dan banyak dimplementasikan dalam data citra[4]. CNN dapat menggunakan operasi konvolusi sebagai dasar dari algoritma nya. Nama Konvolusi sendiri merupakan operasi aljabar linear yang mengkalikan matriks dari filter pada citra yang akan di proses, sehingga proses ini disebut lapisan konvulusi dan merupakan salah satu jenis layer yang dimiliki oleh CNN. Convolutional Neural Network memiliki input, input yang berupa kumpulan pixel yang berupa array.

Gambaran umum Convolutional Neural Network seperti berikut :



Credit Matlab

Lapisan lapisan pada gambar merupakan lapisan yang paling umum digunakan dalam arsitektur Convolutional Neural Network. Lapisan lapisan ini melakukan operasi yang mengubah data berdasarkan lapisan yang telah dilalui dengan maksud mempelajari fitur-fitur khusus pada data. Terdapat tiga lapisan yang umum yaitu konvolusi , fungsi aktivasi dan pooling.

### Lapisan Konvulusi

Lapisan Konvolusi menempatkan input yang berupa gambar melalui serangkaian filter konvolusional, yang masing-masing mengaktifkan fitur tertentu dari gambar(matlab). Pada lapisan konvolusi terdapat filter yang dapat di modifikasi sesuai kebutuhan peneliti. Pada saat input masuk ke lapisan konvolusi input memiliki beberapa parameter yaitu Width(W) , Height(H) dan D (Depth) serta yang menjadi inti berupa nilai dari setiap pixel input tersebut. Untuk mengatur sebuah ukuran output dari lapisan konvolusi terdapat parameter yang mempengaruhi output tersebut, biasanya dituliskan dengan rumus:

Output Layer = (W – F + 2P) / S + 1 dimana

W = Volume Size

F = Ukuran Filter

P = Zero Padding

S = Stride

Nilai 1 diambil dari nilai Bias.

Parameter diatas harus diatur untuk memenuhi output yang di inginkan, contoh :

Sebuah input yang berukuran 5 x 5 x 3 menggunakan zero padding 2 dikonvulusikan dengan Filter 3 x 3 dan menggunakan Stride 2 maka output dari layer tersebut adalah

Output Layer = (W – F + 2P) / s + 1 maka (5 – 3 + 2(2) ) /2 + 1 = 4

Jika kasus diganti dengan Kasus yang sama namun Stride yang digunakan adalah 3 dan padding hanya 1 maka filter ini tidak dapat di konvolusikan ke input. Perhatikan, sesuai rumus maka Output Layer = ( 5 – 3 +2(1) ) / 3 + 1 dimana (5 – 3 + 2 ) / 3 hasilnya bukan merupakan sebuah bilangan integer.

### Fungsi Aktivasi

Fungsi Aktivasi merupakan layer yang memungkinkan pelatihan lebih cepat dan lebih efektif dengan memetakan nilai berdasarkan fungsi aktivasi yang digunakan. Fungsi aktivasi yang sering digunakan adalah fungsi aktivasi ReLU atau Rectified linear unit(ReLU) yaitu fungsi aktivasi yang memetakan nilai negative menjadi nol dan mempertahankan nilai positif.

|  |  |  |  |
| --- | --- | --- | --- |
| 9 | 2 | -5 | 2 |
| 9 | -8 | 5 | -3 |
| -1 | 2 | 4 | 5 |
| 1 | 5 | -4 | 6 |

|  |  |  |  |
| --- | --- | --- | --- |
| 9 | 2 | 0 | 2 |
| 9 | 0 | 5 | 0 |
| 0 | 2 | 4 | 5 |
| 1 | 5 | 0 | 6 |

Contoh Fungsi Aktivasi ReLU

### Pooling

Pooling merupakan layer untuk menyederhanakan output yang sebelumnya telah di convolusikan. Menggunakan prinsip yang menggunakan sebuah filter dengan ukuran dan stride tertentu yang akan bergeser pada seluruh area feature map. Pooling yang biasa digunakan adalah Max Pooling.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 5 | 2 | 5 |
| 2 | 4 | 6 | 2 |
| 1 | 2 | 5 | 3 |
| 4 | 1 | 1 | 2 |

|  |  |
| --- | --- |
| 5 | 6 |
| 4 | 5 |

Contoh operasi Max Pooling 2 x2 dengan stride 2

### Fully Connected Layer

## Image Depth Data

## Akurasi

# METODE PENELITIAN

## Jenis Penelitian

Populasi yang digunakan dalam penelitian ini adalah data citra atau gambar American Sign Language(ASL).

## Waktu dan Tempat

Penelitian ini dilaksanakan dari bulan Oktober 2019 sampai dengan bulan Desember 2019. Lokasi penelitian dilakukan di Laboratorium Rekayasa Perangkat Lunak Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin

## Tahapan Penelitian

### Tahap Pra Penelitian(Persiapan)

Dalam tahapan ini , beberapa hal yang dilakukan oleh peneliti adalah melakukan perencanaan pelaksaan secara konseptual, operasional, persiapan instrument penelitian dan sebagainya. Peneliti juga menentukan tema penelitian , masalah yang terkait mengenai penelitian dan mengumpulkan referensi atau literatur seperti jurnal dan buku yang mendukung penelitian.

### Penelitian

Dalam tahapan ini, peneliti akan melewati beberapa tahap penelitian, yaitu

1. Eksplorasi data ,

Pada eksplorasi data, peneliti akan mencoba melihat karakteristik data yang dalam hal ini data citra.

1. Pembuatan Model

Pada tahap ini akan membuat model untuk mengklasifikasikan citra, penentuan arsitektur CNN yang digunakan akan mempengaruhi hasil yang diperoleh, sehingga pada tahap inilah pembuatan arsitektur CNN sangat penting. Arsitektur CNN sangat bervariasi , setiap layer yang digunakan akan mempengaruhi hasil. Tahap ini akan membuat peneliti mencoba membuat model yang mampu mengklasifikasikan citra yang mempunyai tingkat akurasi yang baik.

1. Evaluasi Model

Pada tahap ini akan diuji dengan citra handsposture dan menghasilkan parameter-parameter yang baik untuk model arsitektur CNN yang telah dibuat.

1. Analisis hasil

Pada tahap ini peniliti akan merangkum hasil dan menyimpulkan hasil dari penelitian ini.

## Deskripsi Data

Data diambil dari Website resmi Kaggle (kaggle.com), UCI Machine Learning Repository (archive.ics.uci.edu/ml/) dan KEEL (sci2s.ugr.es/keel/imbalanced.php). Data tersebut berupa tiga dataset, yaitu:

1. Credit Card Fraud Dataset (Kaggle), yang terdiri dari 30 kolom attribut dengan 1 kolom kelas, 284.807 baris. 284.315 jumlah sampel kelas mayoritas dan 492 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 577:1. Dataset ini merupakan dataset terpopuler di Kaggle sebab jumlah data yang besar dengan imbalanced ratio yang sangat tinggi.
2. Spambase Dataset (UCI), yang terdiri dari 57 kolom atribut dengan 1 kolom kelas, 4.601 baris. 2788 jumlah sampel kelas mayoritas dan 1813 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 1,5:1
3. Image Segmentation Dataset (KEEL), yang terdiri dari 19 kolom atribut dengan 1 kolom kelas, 2308 baris. 1962 jumlah sampel kelas mayoritas dan 346 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 6:1.

Seluruh dataset hanya memiliki atribut kontinu dengan label kelas biner yang sesuai dengan tema penelitian dan metode-metode yang digunakan.

## Alur Penelitian

Studi Literatur

Eksplorasi Data

Preprocessing

Fitting

Analisis dan Pembahasan

Kesimpulan

Tuning

Tuning

Menentukan Target dan Objek Penelitian

# Daftar Pustaka

|  |  |
| --- | --- |
| [dapus2] | Alpaydin, E. (2009). Introduction to Mechine Learning, Second Edition. London: MIT Press |
| [tipus1] | Putra Gotama W. J, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning*,* edisi. 1.2, 2018. |
| [tipus2] | C. Snijders, U. Matzat and U.-D. Reips, "'Big Data'. Big gaps of knowledge in the field of Internet," *International Journal of Internet Science,* vol. 7, pp. 1-5, 2012. |
| [tipus4] | C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. |
| [tipus3] | Sena, S. (2018, Mei 27). Pengenalan Deep Learning Part 7 : Convolutional Neural Network (CNN). Diambil kembali dari Medium: https://medium.com/@samuelse na/pengenalan-deep-learningpart-7-convolutional-neuralnetwork-cnn-b003b477dc94 . |
| [tipus5] | Setiawan Muhammad, "Klasifikasi Penyakit pada Citra Daun Menggunakan Convulutional Neural Networks*,* Institut Pertanian Bogor, 2018. |
| [7] | E. LeDell, M. Petersen and M. v. d. Laan, "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates," *Electronic Journal of Statistics,* vol. 9, pp. 1583-1607, 2015. |
| [8] | A. P. Bradley, "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern recognition,* vol. 30, no. 7, pp. 1145-1159, 1997. |
| [9] | A. Howard, Elementary Linear Algebra, Binder Ready Version: Applications Version, John Wiley & Sons., 2013. |
| [10] | A. Y.-c. Liu, "The effect of oversampling and undersampling on classifying imbalanced text datasets," *The University of Texas at Austin,* 2004. |
| [11] | A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir and A. Hawalah, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access,* vol. 4, pp. 7940-7957, 2016. |
| [12] | N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002. |
| [13] | H. Han, W. Y. Wang and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International conference on intelligent computing,* pp. 878-887, August 2005. |
| [14] | H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* pp. 1322-1328, June 2008. |
| [15] | S. J. Yen and Y. S. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," *Intelligent Control and Automation,* pp. 731-740, 2006. |
| [16] | A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048,* 2016. |
| [17] | I. Tomek, "Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics,* vol. 6, pp. 769-772, 1976. |
| [18] | I. Rish, "An empirical study of the naive Bayes classifier," *Watson Research Center,* 2011. |
| [19] | D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *European conference on machine learning,* pp. 4-15, 1998. |
| [20] | I. Rish, "An Empirical Study of the Naive Bayes Classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence,* pp. 41-46, 2001. |
| [21] | H. Zhang, "The Optimality of Naive Bayes," *American Association for Artificial Intelligence,* 2004. |
| [22] | O. F.Y., A. J.E.T., A. O., Hinmikaiye, Olakanmi and Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT),* vol. 48, no. 3, pp. 128-138, 2017. |
| [23] | J. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, pp. 81-106, 1986. |
| [24] | Q.-y. Dai, C.-p. Zhang and H. Wu, "Research of Decision Tree Classification Algorithm in Data Mining," *International Journal of Database Theory and Application,* vol. 9, no. 5, pp. 1-8, 2016. |
| [25] | T. Vafeiadis, K. Diamantaras, G. Sarigiannidis and K. Chatzisavvas, "A comparison of machine learning techniques for customer," *Simulation Modelling Practice and Theory,* vol. 55, no. 1, pp. 1-9, 55. |
| [26] | N. Guenther and M. Schonlau, "Support vector machines," *The Stata Journal,* vol. 16, no. 4, pp. 917-937, 2016. |
| [27] | C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," *Proceedings of the 17th International Conference on Pattern Recognition,* vol. 3, pp. 32-36, 2004. |
| [28] | V. Vapnik and C. Cortes, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995. |
| [29] | N. K. Ahmed, A. F. Atiya, N. E. Gayar and H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting," *Econometric Reviews,* vol. 29, no. 5-6, pp. 594-621, 2010. |
| [30] | J. M. Zurada, Introduction to artificial neural systems, St. Paul: West publishing company, 1992. |
| [31] | S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall PTR, 1994. |
| [32] | Z. H. Zhou and X. Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," *IEEE Transactions on Knowledge & Data Engineering,* no. 1, pp. 63-77, 2006. |
| [33] | C. M. Bishop, Neural Networks for Pattern Recognition, Oxford university press, 1995. |
| [34] | A. N. N. (. MULTILAYER, "Artificial Neural Networks (The Multilayer Perceptron) - A Review of Applications in the Atmospheric Sciences," *Atmospheric Environment,* vol. 32, no. 14-15, pp. 2627-2636, 1998. |
| [35] | H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge & Data Engineering,* no. 9, pp. 1263-1284, 2008. |
| [36] | J. Desjardins, "How much data is generated each day?," World Economic Forum, 2019. |
| [37] | M. Chen, S. Mao and Y. & Liu, "Big Data: A survey," *Mobile networks and applications,* vol. 19, no. 2, pp. 171-209, 2014. |
| [38] | S. Lohr, "The age of big data," New York Times, New York, 2012. |
| [39] | N. Elgendy and A. Elragal, "Big data analytics: a literature review paper," *Industrial Conference on Data Mining,* pp. 214-227, 2014. |
| [40] | LinkedIn Economic Graph Team, "Linkedin 2018 Emerging Jobs Report," LinkedIn, 2018. |
| [41] | A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil and D. Barton, "Big data: the management revolution," *Harvard business review,* vol. 90, no. 10, pp. 60-68, 2012. |
| [42] | V. Beal, "Unstructured Data," 2019. [Online]. Available: https://www.webopedia.com/TERM/U/unstructured\_data.html. [Accessed 20 6 2019]. |
| [43] | H. Baars and H. G. Kemper, "Management support with structured and unstructured data—an integrated business intelligence framework," *Information Systems Management,* vol. 25, no. 2, pp. 132-148, 2008. |
| [44] | G. Weglarz, "Two Worlds of Data - Unstructured and Structured," *DM Review,* vol. 14, pp. 19-23, 2004. |
| [45] | F. Berman, R. Rutenbar, B. Hailpern, H. Christensen, S. Davidson, D. Estrin, M. Franklin, M. Martonosi, P. Raghavan, V. Stodden and A. S. Szalay, "Realizing the Potential of Data Science," *Communications Of The Acm,* vol. 61, no. 4, pp. 67-72, 2018. |
| [46] | V. Dhar, "Data Science and Prediction," *Communications of the ACM,* vol. 56, no. 12, pp. 64-73, 2012. |
| [47] | F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big data,* vol. 1, no. 1, pp. 51-59, 2013. |
| [48] | F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Muller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and M. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research 12,* 2011. |
| [49] | M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 42, no. 4, pp. 463-484, 2011. |
| [50] | S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, no. 1, pp. 25-36, 2006. |
| [51] | A. Kumar and H. Sheshadri, "On the Classification of Imbalanced Datasets," *International Journal of Computer Applications (0975 – 8887),* vol. 44, no. 8, pp. 1-7, 2012. |
| [52] | S. Visa and A. Ralescu, "Issues in Mining Imbalanced Data Sets - A Review Paper," *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference,* vol. 2005, pp. 67-73, April 2005. |
| [53] | F. Provost, "Machine Learning from Imbalanced Data Sets 101," *Proceedings of the AAAI’2000 workshop on imbalanced data sets,* vol. 68, pp. 1-3, July 2000. |
| [54] | M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *International Journal of Machine Learning and Computing,* vol. 3, no. 2, p. 224, 2013. |
| [55] | Statistics Solution, "Statistics Solution," 2016. [Online]. Available: https://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification-resampling/. [Accessed 14 August 2019]. |
| [56] | E. Burnaev, P. Erofeev and A. Papanov, "Influence of Resampling on Accuracy of Imbalanced Classification," *In Eighth International Conference on Machine Vision (ICMV 2015),* vol. 9875, p. 987521, 2015. |
| [57] | A. Anand, G. Pugalenthi, G. B. Fogel and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino acids,* vol. 39, no. 5, pp. 1385-1391, 2010. |
| [58] | B. Diri and S. Albayrak, "Visualization and analysis of classifiers performance in multi-class medical data," *Expert Systems with Applications,* vol. 34, no. 1, pp. 628-634, 2008. |
| [59] | G. E. Batista, R. C. Prati and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD explorations newsletter,* vol. 6, no. 1, pp. 20-29, 2004. |
| [60] | P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," *Catena,* vol. 145, pp. 164-179, 2016. |
| [61] | D. W. Hosmer and S. Lemeshow, Applied Logistic Regression (Second Edition), Canada: Wiley-Interscience Publication, 2000. |