ISSN: 2962-3545
Prosiding SAINTEK: Sains dan Teknologi Vol.1 No.1 Tahun 2022
Call for papers dan Seminar Nasional Sains dan Teknologi Ke-1 2022
Fakultas Teknik, Universitas Pelita Bangsa, Juli 2022

Klasifikasi Curah Hujan di Kota Semarang Menggunakan *Machine Learning*

Rainfall Classification in the Semarang City Using Machine Learning

Carissa Devina Usman[1], Usman Sudibyo[2]
[1]Departemen Matematika, Fakultas Sains dan Matematika, Universitas Diponegoro
[2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[1]carissadvn@students.undip.ac.id, [2]usman.sudibyo@dsn.dinus.ac.id

Abstract

The erratic distribution of rainfall greatly affects people's daily activities, especially in the Semarang City. Therefore, it is necessary to predict rainfall in Semarang City. Correct prediction of rainfall can improve community preparedness in dealing with various natural disasters caused by rain. Machine learning algorithms and data mining have been widely used in research for rainfall data in various regions. The main purpose of this study is to obtain predictions of rainfall in the city of Semarang using machine learning algorithms and to find out the best algorithm for classifying. The dataset used was obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) which is the daily rainfall data in Semarang City. From the dataset, three machine learning algorithms will be classified, namely Logistic Regression, Random Forest, and Gradient Boosting. To measure the performance of the machine learning algorithm, the classification accuracy of each algorithm is measured. From the research results, the performance of the Gradient Boosting algorithm is better than other algorithms, with an accuracy value of 71.6%.

Keywords: *Machine Learning, Logistic Regression, Random Forest, Gradient Boosting, Rainfall Prediction*

Abstrak

Distribusi curah hujan yang tidak menentu sangat mempengaruhi aktivitas sehari-hari masyarakat, terutama di Kota Semarang. Oleh karena itu, prediksi curah hujan di Kota Semarang perlu dilakukan. Prediksi curah hujan yang tepat dapat meningkatkan kesiapsiagaan masyarakat dalam mengatasi berbagai bencana alam yang disebabkan oleh hujan. Algoritma machine learning dan data mining telah banyak digunakan dalam penelitian untuk data curah hujan di berbagai wilayah. Tujuan utama dari penelitian ini adalah untuk memperoleh prediksi curah hujan di Kota Semarang menggunakan Algoritma machine learning dan untuk mengetahui Algoritma terbaik dalam melakukan klasifikasi. Dataset yang digunakan diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) yang merupakan data curah hujan harian di Kota Semarang. Dari dataset tersebut, akan dilakukan klasifikasi dengan tiga algoritma machine learning, yaitu Regresi Logistik, *Random Forest*, dan *Gradient Boosting*. Untuk mengukur kinerja dari Algoritma machine learning, dilakukan pengukuran akurasi klasifikasi dari masing – masing algoritma. Dari hasil penelitian, kinerja algoritma Gradient Boosting lebih baik daripada algoritma lainnya, dengan nilai akurasi sebesar 73,1%.

Kata kunci: *Machine Learning*, Regresi Logistik, *Random Forest, Gradient Boosting*, Prediksi Curah Hujan

Pendahuluan

Curah hujan sangat mempengaruhi aktivitas sehari-hari masyarakat. Terlebih hujan ekstrem yang dapat menyebabkan berbagai bencana alam yang tentunya dapat merugikan masyarakat. Berdasarkan data bencana dari Badan Penanggulangan Bencana Daerah (BPBD) Kota Semarang tahun 2018, 2019, dan 2020, hujan ekstrem meyebabkan berbagai bencana alam seperti banjir, tanah longsor, pohon tumbang, dan rumah roboh. Hal itu pasti merugikan masyarakat dari segi ekonomi maupun kesehatan. Oleh karena itu melakukan prediksi jumlah curah hujan sangat penting untuk meningkatkan kesiapsiagaan masyarakat terhadap bencana, khususnya di daerah Kota Semarang, Indonesia.

Menurut Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), hujan berdasarkan intensitasnya dikategorikan menjadi enam, yaitu berawan, hujan ringan, hujan sedang, hujan lebat, hujan sangat lebat, dan

hujan ekstrem. Intensitas dari masing-masing kategori hujan adalah 0 – 0.5 mm/hari untuk kondisi berawan, 0.5 – 20 mm/hari untuk hujan ringan, 20 – 50 mm/hari untuk hujan sedang, 50 – 100 mm/hari untuk hujan lebat, 100 – 150 mm/hari untuk hujan sangat lebat, dan lebih dari 150 mm/hari untuk hujan ekstrem. Maka apabila pada suatu hari intensitas curah hujan diketahui, maka kategori hujan yang akan terjadi dapat diketahui pula. Intensitas curah hujan dapat diketahui dari beberapa variabel yang meliputi suhu minimal, suhu maksimal, suhu rata-rata, kelembaban, lama sinar matahari, kecepatan angin maksimal, kecepatan angin rata-rata, kecepatan arah angin, dan arah angin terbanyak. Beberapa variabel tersebut baik yang memiliki dampak langsung atau tidak langsung pada curah hujan harus dipelajari untuk memprediksi keberadaan dan intensitas curah hujan. Hal itu dapat memudahkan masyarakat untuk mengetahui cuaca yang akan terjadi sehingga dapat meminimalisir kerugian yang disebabkan oleh hujan ekstrem.

Prediksi curah hujan dilakukan dengan menggunakan algoritma *machine learning*. *Machine learning* menyelidiki bagaimana komputer dapat belajar berdasarkan data. Tujuannya adalah supaya program komputer secara otomatis belajar mengenali pola kompleks dan membuat keputusan cerdas berdasarkan data [1]. Ada beberapa Algoritma yang dimiliki oleh *machine learning*, namun pada umumnya *machine learning* memiliki dua Algoritma dasar yaitu *supervised* dan *unsupervised*[2],[3]. Data curah hujan dapat diolah dengan Algoritma pembelajaran *supervised*, sebab memiliki variabel target yaitu curah hujan. Sedangkan prediksi dilakukan dengan melakukan klasifikasi hujan berdasarkan intensitasnya.

Penelitian sebelumnya tentang prediksi curah hujan dibahas pada [4], [5], dan [6]. Pada [4] tidak hanya memprediksi apakah hujan atau tidak, tetapi juga menunjukkan prediksi jumlah curah hujan harian dengan algoritma Multivariate Linear Regression, Random Forest, dan Extreme Gradient Boost. Namun penelitian tersebut hanya fokus membahas perbandingan kinerja algoritma daripada pembahasan output dari masing-masing metode. Sedangkan pada [5] dan [6], model prediksi dengan machine learning telah dikembangkan, yaitu Bayesian Quantile Regression dan ANFIS untuk memperoleh prediksi yang akurat. Namun perbandingan kinerja dengan algoritma machine learning yang lain tidak dilakukan.

Dalam penelitian ini akan dibahas mengenai prediksi curah hujan harian di Kota Semarang dengan metode regresi pada machine learning, yaitu Regresi Logistik, *Random Forest*, dan *Gradient Boosting* yang meliputi metode pengumpulan data, preprocessing data, machine learning, pengukuran performa, serta hasil dan diskusi dari penelitian yang telah dilakukan.

Metode Penelitian

Penelitian ini menggunakan data curah hujan harian Kota Semarang pada tanggal 1 Januari 2018 hingga 30 Maret 2021, yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Data tersebut terdiri dari 9 variabel yang mempengaruhi jumlah curah hujan, yaitu suhu minimal, suhu maksimal, suhu rata-rata, kelembaban, lama sinar matahari, kecepatan angin maksimal, kecepatan angin rata-rata, kecepatan arah angin, dan arah angin terbanyak serta memiliki target berupa kategori hujan. Kumpulan data curah hujan yang berjumlah 1.096 data dicatat dalam format tabel file Microsoft Excel.

Secara umum, preprocessing data terdiri dari lima tugas utama, yaitu pembersihan data, reduksi data, penskalaan data, transformasi data, dan partisi data (Cheng Fan et al, 2021). Pada penelitian ini dilakukan pembersihan data dengan mempertimbangkan missing value serta dilakukan transformasi data. Untuk menangani missing value dalam membangun data operasional dilakukan dengan cara membuang sampel data dengan missing value, karena sebagian besar algoritma data mining tidak dapat menangani data dengan missing value. Metode ini dipilih ketika nilai label kelas tidak ada, dan digunakan ketika tupel memiliki beberapa atribut dengan nilai kosong (Suad A. Alasadi, 2017).

Metode tersebut hanya dapat diterapkan jika proporsi missing value tidak signifikan (Cheng Fan et al, 2021). Transformasi data dilakukan untuk mengubah data menjadi format yang dapat digunakan dalam proses data mining. Hal ini dilakukan dengan cara normalisasi data. Pada dasarnya proses ini digunakan untuk menskalakan data atribut, sehingga dihasilkan data dengan rentang yang lebih kecil, yang berguna untuk algoritma klasifikasi (Ashish P. Joshi, 2020). Pada penelitian ini, metode yang digunakan untuk normalisasi data yaitu normalisasi min-max. sesuai dengan namanya, metode ini menggunakan nilai minimum dan maksimum untuk melakukan konversi data secara linier [7].

Beberapa algoritma *machine learning* telah ditinjau untuk mempelajari prediksi jumlah curah hujan dengan melakukan klasifikasi. Algoritma yang akan digunakan antara lain Regresi Logistik, *Random Forest*, dan *Gradient Boosting*, yang nantinya akan dibandingkan untuk mengetahui algoritma terbaik untuk memprediksi jumlah curah hujan.

Dalam regresi logistik, variabel dependen ditampilkan sebagai variabel biner [8]. Pada prediksi curah hujan, variabel dependen berjumlah lebih dari dua, sehingga regresi logistik multinomial digunakan. Seperti dalam regresi logistik biner, regresi logistik multinomial menggunakan estimasi *maximum likelihood* untuk mengevaluasi probabilitas keanggotaan kategoris. Dengan demikian, jenis model ini memungkinkan untuk mengkarakterisasi probabilitas keputusan responden untuk pilihan diskrit multinomial tertentu, tergantung pada nilai-nilai variabel penjelas [9]. Setelah model regresi multinomial dibuat, parameter digunakan untuk membuat prediksi tentang probabilitas suatu peristiwa yang terjadi dibandingkan dengan kategori referensi.

Model *Random Forest* biasanya memiliki kinerja yang baik pada berbagai masalah, termasuk fitur dengan hubungan non-linear [3]. *Random Forest* adalah algoritma *supervised machine learning* yang menggunakan metode pembelajaran *ensemble*. *Random forest* dioperasikan dengan membangun banyak *decision tree* pada waktu pelatihan dan mengeluarkan kelas yang merupakan model prediksi rata-rata. Menurut [4] algoritma *Random Forest* efisien untuk kumpulan data yang besar dan hasil eksperimen yang baik diperoleh dengan menggunakan kumpulan data besar yang memiliki sebagian besar *missing value*.

Untuk metode *Gradient Boosting* konvensional, pohon prediktif yang sederhana namun lemah dihasilkan secara berulang dan ditambahkan ke mesin prediksi hingga prediksi mendekati kebenaran [12]. Metode CatBoost diterapkan pada penelitian ini. CatBoost adalah implementasi dari *Gradient Boosting*, yang menggunakan *Decision Tree* biner sebagai prediktor dasar [10]. Algoritma ini menggunakan skema baru untuk menghitung nilai daun pada saat memilih struktur pohon yang dapat membantu mengurangi *overfitting* [11]-[13].

Hasil dan Pembahasan

Pada penelitian ini digunakan data curah hujan yang berjumlah 1.096 data dan dilakukan klasifikasi data berdasarkan kategori hujan. Berdasarkan intensitas curah hujan, kategori hujan dibagi menjadi 6 kategori, yaitu berawan, hujan ringan, hujan sedang, hujan lebat, hujan sangat lebat, dan hujan ekstrem. Namun karena kategori hujan sangat lebat dan hujan ekstrem memiliki jumlah data yang sangat sedikit, maka keduanya dikategorikan sebagai hujan lebat. Hal itu karena apabila dilakukan klasifikasi maka akan terjadi masalah pada *cross validation*.

Algoritma *machine learning* dilatih dengan menggunakan *k-fold cross validation*, dengan memilih nilai *folds* = 10. Pada algoritma Regresi Logistik, digunakan tipe regulasi Ridge L2 dan dengan kekuatan C=19. Sedangkan pada algoritma *Random Forest* dilakukan kontrol pertumbuhan dengan 18 pohon, batas kedalaman pohon yang ditanam adalah 3, dan batas split subset kurang dari 5. Sementara untuk algoritma *Gradient Boosting* digunakan metode catboost dengan jumlah pohon sebanyak 100, learning rate 0.3, regularisasi lambda:3, batas kedalaman pohon yang ditanam adalah 6, dan pecahan fitur untuk setiap pohon bernilai 1. Berikut diberikan tabel evaluasi hasil klasifikasi data curah hujan menggunakan regresi logistik, *random forest*, dan *gradient boosting*.

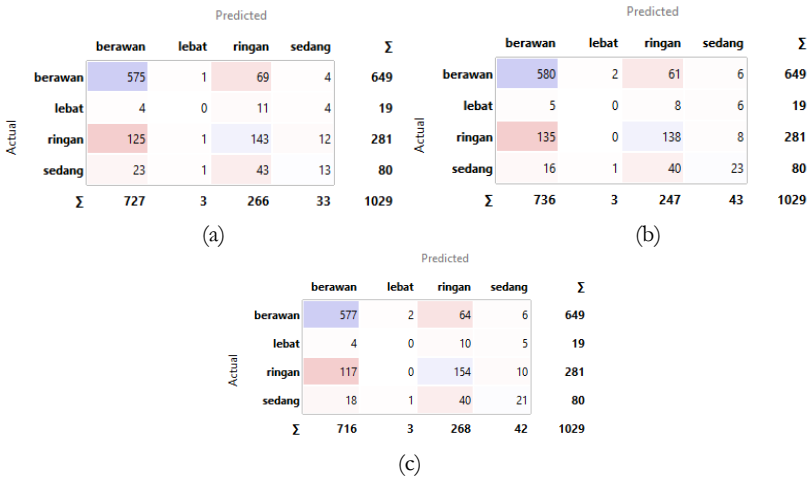
Tabel 1 Evaluasi Hasil Klasifikasi Data Curah Hujan

Model	AUC	CA	F1	Precision
Regresi Logistik	0.809	0.698	0.673	0.663
Random Forest	0.822	0.720	0.700	0.691
Gradient Boosting	0.844	0.731	0.713	0.704

Pada Tabel 1, kolom *Classification Accuracy* (CA) menyatakan akurasi dari masing – masing model. Akurasi tertinggi yaitu sebesar 73,1% yang diperoleh dengan model Gradient Boosting, sedangkan akurasi terendah sebesar 69,8% yang diperoleh dengan model Regresi Logistik. Ini berarti bahwa pada *confusion matrix* untuk *gradient boosting*, nilai yang diklasifikasikan dengan benar lebih banyak dari 2 algoritma lain.

Kolom *Area Under ROC Curve* (AUC) menunjukkan nilai skalar tunggal yang mengukur kinerja keseluruhan dari pengklasifikasi biner. Nilai AUC berada dalam kisaran [0.5–1.0], dimana nilai minimum adalah kinerja pengklasifikasi acak dan nilai maksimum adalah pengklasifikasi sempurna [14],[15]. Pada Tabel 1 terlihat

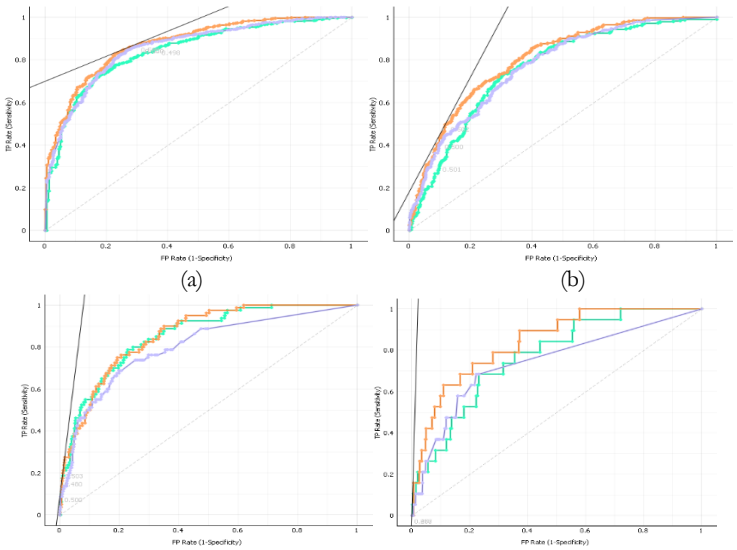
bahwa semua algoritma pengklasifikasi memiliki nilai AUC berada dalam kisaran [0.5–1.0], sehingga ketiga algoritma klasifikasi memiliki kinerja yang baik. AUC biasanya dihitung dengan menambahkan area trapesium berturut-turut di bawah kurva ROC [16]. Dalam kasus ini, pengklasifikasi *Gradient Boosting* memiliki nilai AUC yang lebih besar daripada pengklasifikasi lainnya, sehingga kinerja klasifikasinya adalah yang terbaik. Metrik lain seperti F1 dan Precision juga ditunjukkan pada Tabel 1.



Gambar 1 *Confusion Matrix* Algoritma (a) Regresi Logistik, (b) *Random Forest*, dan (c) *Gradient Boosting*

Confusion matrix berisi informasi tentang klasifikasi aktual dan prediksi dari pengklasifikasi. Kinerja pengklasifikasi seperti itu biasanya dievaluasi menggunakan data dalam matriks [13]. Gambar 1 menunjukkan *confusion matrix* untuk pengklasifikasi (Regresi logistik, *Random forest, Gradient boosting*).

Untuk kelas berawan dan hujan sedang, kesesuaian data aktual dan prediksi yang terbanyak dihasilkan oleh algoritma *random forest*. Pada kelas hujan lebat, tidak ada satupun algoritma yang dapat mengklasifikasi dengan benar. Sedangkan untuk kelas hujan ringan, kesesuaian data aktual dan prediksi yang terbanyak dihasilkan oleh algoritma *gradient boosting*. Oleh karena itu algoritma *gradient boosting* memiliki ketepatan prediksi yang terbaik, sebab jumlah data yang terprediksi dengan tepat lebih banyak dibandingkan algoritma regresi logistik dan *random forest*.



Gambar 2 Kurva *Receiver Operating Characteristic* (ROC) pada Kelas (a) Berawan, (b) Hujan Ringan, (c) Hujan Sedang, (d) Hujan Lebat

Gambar 2 menunjukkan kurva ROC untuk tiga pengklasifikasi, yaitu garis hijau sebagai kurva regresi logistik, garis biru sebagai kurva *random forest*, dan garis jingga sebagai kurva *gradient boosting*. Sumbu y adalah *true*

positive rate dan sumbu x adalah *false positive rate*. Semakin luas area dibawah kurva (AUC) maka semakin baik algoritmanya. Dengan kata lain, semakin kurva algoritma mendekati garis hitam dan menjauhi garis abu-abu, maka semakin baik algoritma tersebut. Oleh sebab itu pada Gambar 2, baik pada kelas berawan, hujan ringan, hujan sedang, maupun hujan lebat, kurva algoritma *gradient boosting* adalah yang paling baik.

Kesimpulan

Prediksi curah hujan di Kota Semarang adalah hal yang penting untuk meningkatkan kesiapsiagaan masyarakat Kota Semarang dalam menghadapi berbagai jenis hujan, terutama hujan ekstrem yang dapat menyebabkan bencana alam. Penelitian ini menganalisis berbagai algoritma *machine learning* untuk prediksi curah hujan. Algoritma *machine learning* seperti Regresi Logistik, *Random Forest*, dan *Gradient Boosting* disajikan dan diuji menggunakan data yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika Kota Semarang. Fitur lingkungan yang cocok untuk prediksi curah hujan dipilih dan digunakan sebagai variabel input untuk model machine learning. Perbandingan akurasi antara tiga algoritma telah diperoleh. Hasilnya menunjukkan bahwa *Gradient Boosting* adalah algoritma *machine learning* yang lebih cocok untuk prediksi jumlah curah hujan harian dengan nilai akurasi sebesar 73,1%. Pada penelitian selanjutnya diharapkan penelitian untuk prediksi jumlah curah hujan dapat berkembang untuk meningkatkan akurasi pada data dan kinerja algoritma.

Daftar Rujukan

[1] Han, J., Kamber, M. and Pei, J., (2012), Data Mining: Concepts and Techniques, 3rd ed., *USA: Elsevier Inc.,* doi: 10.1016/C2009-0-61819-5

[2] Pramana, Setia, (2018), Data Mining dengan R: Konsep Serta Implementasi, *Bogor: In Media.*

[3] Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(1), 187-192.

[4] C. M. Liyew and H. A. Melese, (2021), "Machine learning techniques to predict daily rainfall amount," *J. Big Data,* vol. 8, no. 1, doi: 10.1186/s40537-021-00545-4.

[5] W. Suparta and A. A. Samah, (2020), "Rainfall prediction by using ANFIS times series technique in South Tangerang, Indonesia," *Geod. Geodyn.,* vol. 11, no. 6, pp. 411–417, doi: 10.1016/j.geog.2020.08.001.

[6] R. N. Rachmawati, I. Sungkawa, and A. Rahayu, (2019), "Extreme rainfall prediction using Bayesian quantile regression in statistical downscaling modeling," *Procedia Comput. Sci.,* vol. 157, pp. 406–413, doi: 10.1016/j.procs.2019.08.232.

[7] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, (2021), "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.,* vol. 9, no. March, pp. 1–17, doi: 10.3389/fenrg.2021.652801.

[8] S. A. Alasadi and W. S. Bhaya, (2017), "Review of Data Preprocessing Techniques in Data Mining," *Journal of Engineering and Applie Sciences,* vol. 12, no. 16. pp. 4102–4107.

[9] A. P. Joshi and B. V. Patel, (2021), "Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Orient. J. Comput. Sci. Technol.,* vol. 13, no. 0203, pp. 78–81, doi: 10.13005/ojcst13.0203.03.

[10] Suyanto, (2017), Data Mining untuk Klasifikasi dan Klasterisasi Data, 1st ed. *Bandung: Informatika.*

[11] T. Ciu and R. S. Oetama, (2020), "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (International J. New Media Technol.,* vol. 7, no. 1, pp. 33–38, doi: 10.31937/ijnmt.v7i1.1340.

[12] B. Umaña-Hermosilla, H. de la Fuente-Mella, C. Elórtegui-Gómez, and M. Fonseca-Fuentes, (2020), "Multinomial logistic regression to estimate and predict the perceptions of individuals and companies in the face of the covid-19 pandemic in the Ñuble region, Chile," *Sustain.,* vol. 12, no. 22, pp. 1–20, doi: 10.3390/su12229553.

[13] Y. Xie et al., (2019), "Use of Gradient Boosting Machine Learning to Predict Patient Outcome in Acute Ischemic Stroke on the Basis of Imaging, Demographic, and Clinical Information," no. January, pp. 1–7, 2019.

[14] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, (2020), "Comparison of the CatBoost Classifier with other Machine Learning Methods," *Int. J. Adv. Comput. Sci. Appl.,* vol. 11, no. 11, pp. 738–748, doi: 10.14569/IJACSA.2020.0111190.

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.,* vol. 2018-December, no. Section 4, pp. 6638–6648.

[16] V. Brahmachari and S. Jain, (2013), *Encyclopedia of Systems Biology.*