FLOBAMORA BIOLOGICAL JOURNAL
website:https://ejurnal-unisap.ac.id/index.php/flobijo/index
Volume (2) No. 2 (2023) | E-ISSN 2829-1840

KLASIFIKASI *SUPPORT VECTOR MACHINE* DAN *RANDOM FOREST* PADA DATA BIOMEDIS : APLIKASI DALAM ANALISIS DATA PENYAKIT DIABETES

Support Vector Machine and Random Forest Classification on Biomedical Data: Application in Diabetes Data Analysis

Sefri Imanuel Fallo[1], Moch. Anjas Aprihartha[2], Jus Prasetya [3]
Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas San Pedro [1],
Program Studi PJJ Informatika Fakultas Ilmu Komputer Universitas Dian Nuswantoro[2],
Alumni Magister Matematika Universitas Gadjah Mada[3]
fallosefriimanuel@gmail.com

Abstrak

Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam mengklasifikasikan data biomedis terkait penyakit diabetes. Data yang digunakan mencakup informasi tentang deteksi diabetes pada pasien, dengan variabel-variabel seperti Indeks Massa Tubuh (BMI), tingkat HbA1c, dan tingkat glukosa darah. Hasil penelitian menunjukkan bahwa *Random Forest* memberikan tingkat akurasi yang lebih tinggi dibandingkan SVM. Hasil prediksi random forest menghasilkan pasien terdeteksi diabetes sebanyak 73 kasus dan pasien tidak terdeteksi diabetes 1127 kasus dengan akurasi 97.17%. Evaluasi model menegaskan bahwa Random Forest mencapai nilai kappa sebesar 0.7967, menandakan kemampuannya dalam memprediksi penyakit diabetes dengan lebih efektif. Hasil ini menyiratkan bahwa *Random Forest* dapat menjadi pilihan yang lebih optimal dalam memodelkan prediksi penyakit diabetes, terutama ketika mempertimbangkan variabel-variabel yang relevan seperti BMI, tingkat HbA1c, dan tingkat glukosa darah dalam analisisnya.

Kata Kunci: *Diabetes*, *Biomedical*, *Support Vector Machine*, *dan Random Forest*

Abstract

The aim of this study is to compare the performance of the Support Vector Machine (SVM) and Random Forest algorithms in classifying biomedical data related to diabetes mellitus. The dataset comprises information concerning the detection of diabetes in patients, incorporating variables such as Body Mass Index (BMI), HbA1c levels, and blood glucose levels. The findings of the research demonstrate

that Random Forest exhibits a higher accuracy rate compared to SVM. The Random Forest prediction results indicate that 73 cases were detected with diabetes while 1127 cases were not, achieving an accuracy of 97.17%. Model evaluation further confirms that Random Forest achieves a kappa value of 0.7967, signifying its effectiveness in predicting diabetes mellitus. These results suggest that Random Forest may represent a more optimal choice for modeling diabetes prediction, particularly when considering relevant variables such as BMI, HbA1c levels, and blood glucose levels in the analysis.

***Keywords**: Diabetes, Biomedical, Support Vector Machine, dan Random Forest*

PENDAHULUAN

Penyakit diabetes merupakan salah satu masalah kesehatan global yang signifikan, dengan prevalensi yang terus meningkat di seluruh dunia. Diperkirakan bahwa sekitar 1.3 Milyar orang dewasa menderita diabetes pada tahun 2050, dan angka ini diperkirakan akan terus bertambah dalam beberapa dekade mendatang (Tural Buyuk et al. 2023). Diabetes, baik tipe 1 maupun tipe 2, memiliki dampak yang serius pada kesehatan individu, termasuk risiko tinggi terhadap komplikasi jangka panjang seperti penyakit jantung, stroke, gagal ginjal, dan kehilangan penglihatan. Diabetes melibatkan berbagai faktor yang kompleks dan saling terkait, termasuk faktor genetik, gaya hidup, dan lingkungan (Singh et al. 2024). Pengelolaan penyakit ini membutuhkan pendekatan yang holistik dan personal, yang memungkinkan pencegahan, deteksi dini, dan pengobatan yang tepat waktu. Dalam konteks ini, analisis data biomedis dapat membantu dalam pengembangan model prediktif yang dapat membantu dalam identifikasi individu yang rentan terhadap diabetes, serta dalam pemantauan dan pengelolaan penyakit pada tingkat individual (Agarwal et al. 2024).

Dalam upaya untuk memahami dan mengelola penyakit ini, analisis data biomedis telah menjadi semakin penting (Zhang et al. 2023). Data biomedis, yang mencakup informasi tentang parameter klinis, genetik, dan lingkungan, dapat memberikan wawasan yang berharga tentang faktor risiko, pola perkembangan penyakit, dan respons terhadap perawatan. Di antara berbagai metode analisis data, klasifikasi menggunakan teknik pembelajaran mesin telah menunjukkan keunggulan dalam mengidentifikasi pola kompleks dalam data biomedis (Zhang et al. 2023).

Dalam konteks ini, penelitian ini bertujuan untuk menerapkan dan membandingkan dua teknik klasifikasi yang umum digunakan, yaitu *Support Vector Machine* (SVM) dan *Random Forest*, pada data biomedis terkait diabetes (Febrian et al. 2022). SVM dan Random Forest adalah dua algoritma pembelajaran mesin yang kuat dan sering digunakan dalam berbagai aplikasi, termasuk analisis data kesehatan (Mujumdar and Vaidehi 2019). Teknik klasifikasi seperti *SVM* dan *Random Forest* telah terbukti efektif dalam menghadapi tantangan yang terkait dengan data biomedis yang kompleks dan berdimensi tinggi (Oikonomou and Khera

2023). SVM bekerja dengan memisahkan dua kelas dengan cara menemukan hyperplane yang paling memaksimalkan margin antara kelas-kelas tersebut dalam ruang fitur, sementara *Random Forest* menggabungkan prediksi dari beberapa pohon keputusan yang dihasilkan secara acak. Kedua metode ini telah digunakan secara luas dalam berbagai aplikasi, termasuk dalam analisis data kesehatan, karena kemampuan mereka untuk menangani data yang tidak linear, besar, dan berisiko tinggi (Harsa et al. 2023).

Melalui penerapan teknik-teknik ini, kami berharap untuk meningkatkan pemahaman tentang hubungan antara variabel biomedis dan diagnosis diabetes, serta membandingkan kinerja keduanya dalam konteks analisis data ini (Mujumdar and Vaidehi 2019). Penelitian ini dapat memberikan kontribusi signifikan untuk pengembangan metode diagnosis dan manajemen penyakit diabetes secara efektif dan efisien.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Kaggle.com. Dataset diabetes ini merupakan kumpulan data medis dan demografis dari pasien, beserta status diabetes mereka (positif atau negatif). Data tersebut mencakup fitur seperti usia, jenis kelamin, indeks massa tubuh (BMI), hipertensi, penyakit jantung, riwayat merokok, tingkat HbA1c, dan kadar glukosa darah. Dataset ini kemudian digunakan untuk membangun model machine learning untuk memprediksi diabetes pada pasien berdasarkan riwayat medis dan informasi demografis mereka. Teknik analisis yang digunakan dalam penelitian ini adalah *Machine Learning*. *Machine Learning*, sebuah subbidang dari kecerdasan buatan, memungkinkan sistem komputer untuk memperoleh pengetahuan dari data dan menghasilkan prediksi atau keputusan tanpa pemrograman eksplisit (Ponti 2018). Dengan mengenali pola dan tren dalam data, *Machine Learning* memberdayakan sistem komputer untuk belajar dan beradaptasi secara mandiri. Selama beberapa dekade terakhir, *Machine Learning* telah muncul sebagai teknologi yang menonjol dan berpengaruh (Tehrani et al. 2022).

1. *Support Vector Machine (SVM)*

   SVM adalah metodologi pembelajaran mesin yang didasarkan pada prinsip Minimalisasi Risiko Struktural (SRM)(Huang, Lu, and Ling 2003). Tujuan utamanya adalah untuk membedakan *hyperplane* optimal dalam ruang input, yang dapat efektif memisahkan kelas-kelas yang berbeda. Tantangan utama yang dihadapi oleh SVM terletak pada penentuan fungsi yang sesuai yang dapat secara akurat memisahkan kedua kelas, menggunakan informasi yang berasal dari data pelatihan yang tersedia(Pradeep and Naveen 2018). Algoritma SVM dapat diimplementasikan melalui langkah-langkah berikut:

   a. Membagi data menjadi data *testing* dan data *training*
   b. Menetukan *cost value*

c. Memntukan standar deviasi dan alpha
d. Mengimplementasikan RBF Kernel

$$K\left(x_i, x_j\right) = e^{-\frac{1}{2\sigma^2}\left\|x_i - x_j\right\|^2}$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left(x_i - x_j\right)^T\left(x_i - x_j\right)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left(x_i^T x_i - 2x_i^T x_j + x_i^T x_j\right)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left(x_i^T x_i + x_i^T x_j\right)\right] + \exp\left[x_i^T x_j\right]$$

e. Menentukan hasil prediksi dari RBF Kernel

$$f\left(x\right) = sign\left(\sum_{i,j}^{n} \alpha_i y_i \exp\left[-\frac{1}{2\sigma^2}\left(x_i^T x_i + x_i^T x_j\right)\right] + \exp\left[x_i^T x_j\right] + b\right)$$

2. *Random Forest*

Random Forest adalah algoritma pembelajaran mesin yang menggabungkan beberapa pohon keputusan untuk membuat prediksi (Tang et al. 2020). Setiap pohon dibangun pada subset acak dari data pelatihan dan fitur, dan algoritma memilih titik pemisahan terbaik berdasarkan kriteria tertentu. Prediksi akhir dibuat dengan menggabungkan prediksi dari semua pohon (Liu et al. 2022). *Random Forest* dikenal karena kemampuannya untuk menangani data berdimensi tinggi dan mengurangi *overfitting* (Adnan et al. 2023). Algoritma Random Forest dapat diimplementasikan melalui langkah-langkah berikut:

a. Pemilihan k sampel dari dataset D, secara acak dan dengan penggantian.
b. Memanfaatkan dataset D untuk membangun pohon keputusan ke-i.
c. Dalam membangun pohon keputusan ke-i, metodologi CART dapat digunakan. Metodologi CART menggunakan keuntungan informasi untuk menentukan setiap simpul dalam pohon (Kao et al. 2021). Perhitungan keuntungan informasi dapat dihitung menggunakan persamaan:

$$Gain\left(A\right) = Info\left(D\right) - Info_A\left(D\right)$$

Nilai dari $Info\left(D\right)$ dapat ditentukan dengan formula $Info_A\left(D\right)$ di bawah ini:

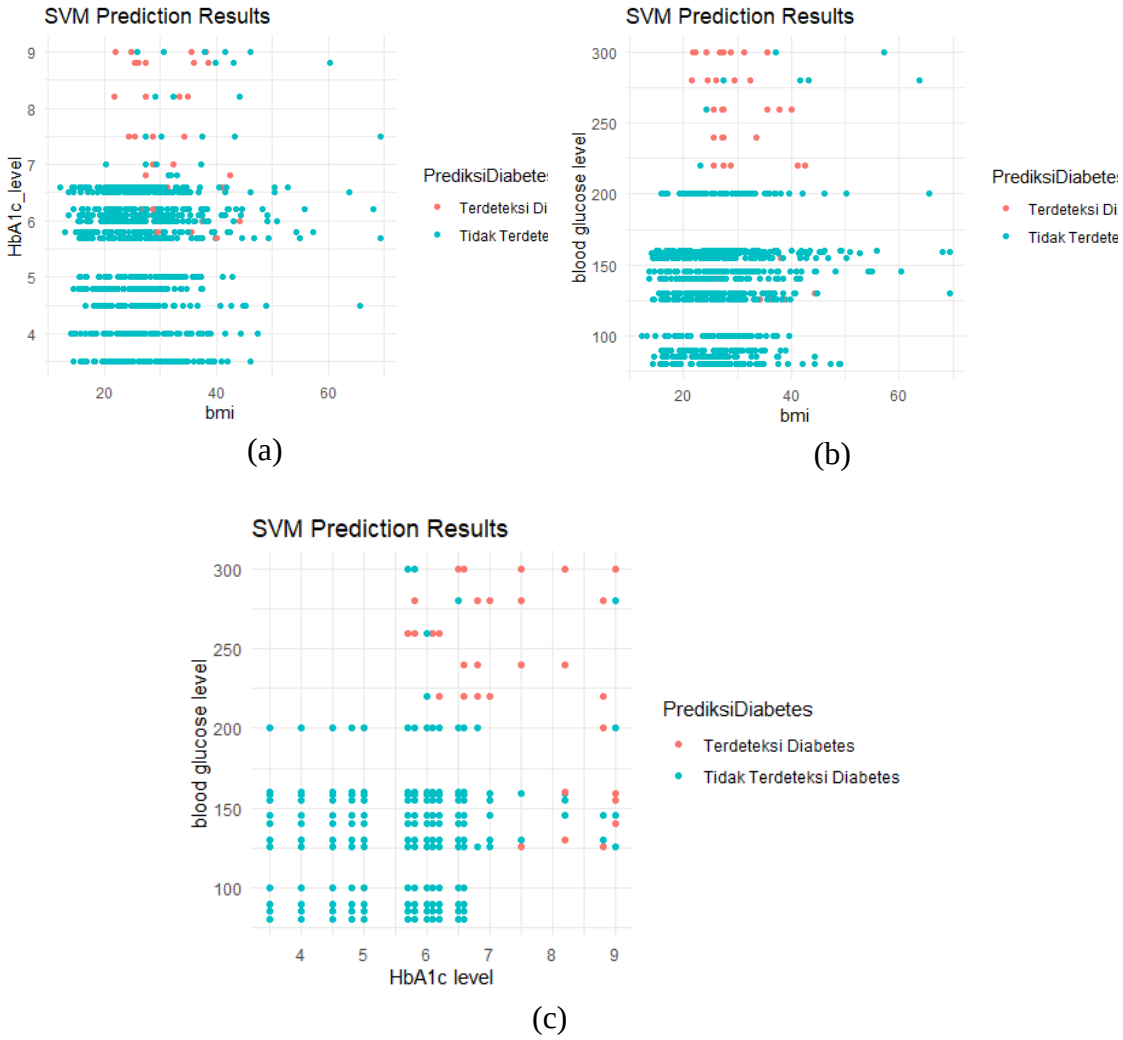
$$Info\left(D\right) = -\sum_{i=1}^{m} p_i \log_2\left(p_i\right)$$

HASIL DAN PEMBAHASAN

Hasil peneitian ini merupakan evaluasi model *machine learning* dalam memprediksi penyakit diabetes.

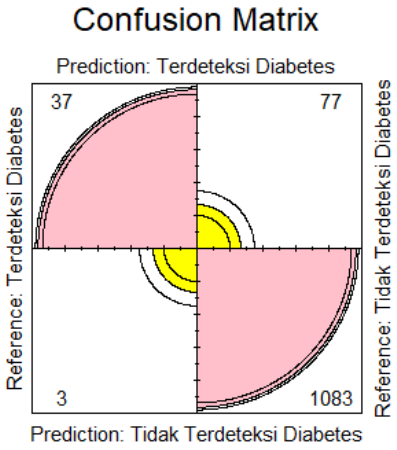
1. *Support Vector Machine*

SVM memprediksi data diabetes dengan kernel RBF, dengan hasil evaluasi model yang disajikan di bawah ini:



Gambar 1. (a)(b)(c) Visualisasi Grafik Hasil Prediksi SVM

Berdasarkan ilustrasi di atas diketahui bahwa hasil prediksi pasien terdeteksi diabetes sebanyak 40 kasus dan pasien tidak terdeteksi diabetes 1160 kasus.



Gambar 2. Confusion Matriks SVM

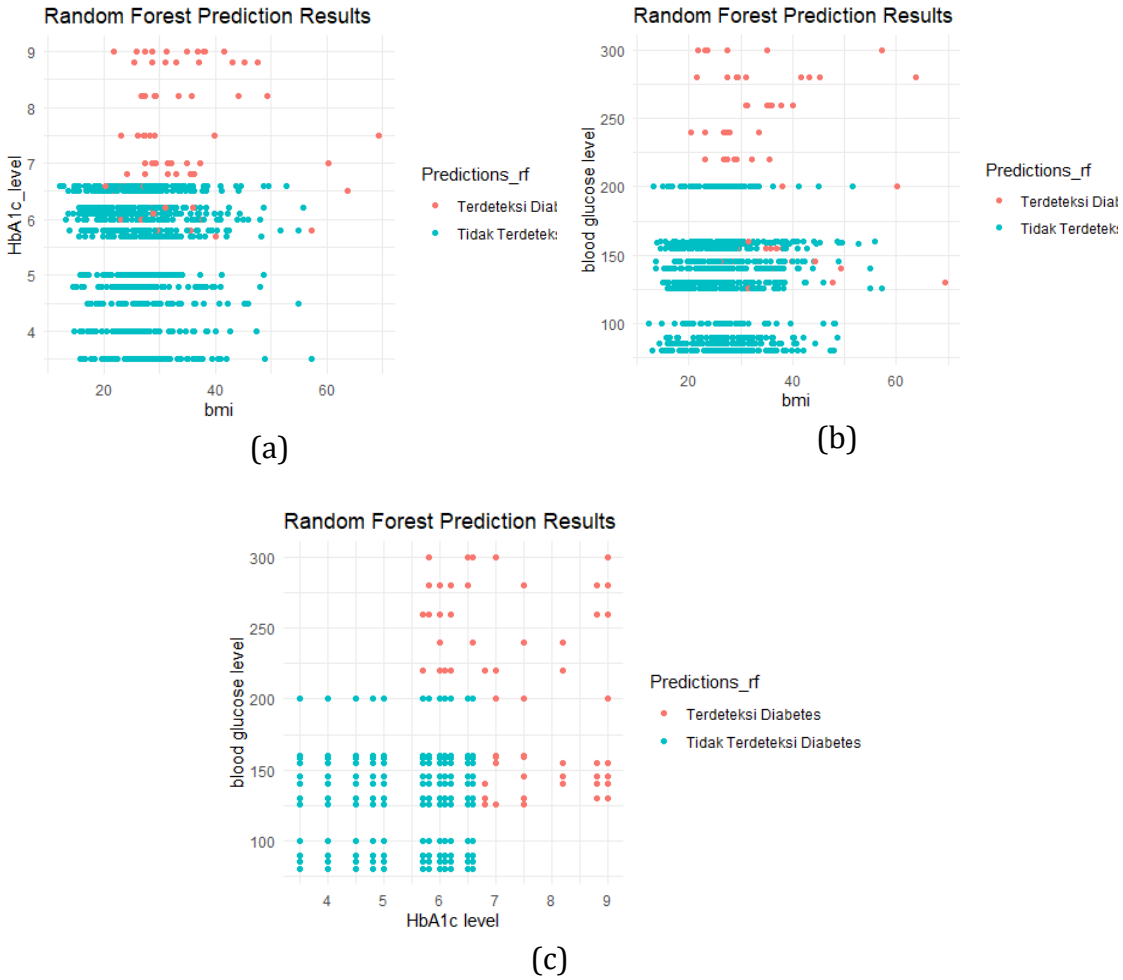
Hasil statistik komprehensif menggunakan algoritma SVM dijelaskan dalam tabel di bawah ini.

Tabel 1. Statistik Komprehensif SVM

Overall statistics	**Nilai**
Accuracy	0.9333
95% CI	(0.9177, 0.9468)
No Information Rate	0.9667
P-Value [Acc > NIR]	1
Kappa	0.4536

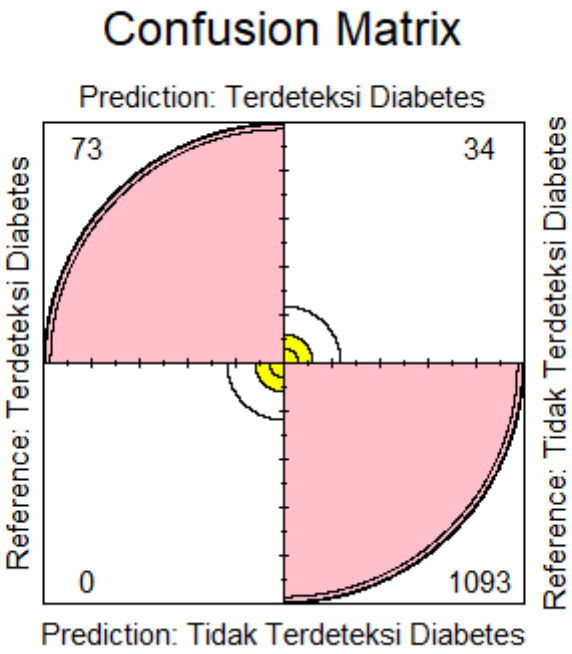
2. *Random Forest*

Random Foresr memprediksi data diabetes dengan model CART, dengan hasil evaluasi model yang disajikan di bawah ini:



Gambar 3. (a)(b)(c) Visualisasi Grafik Hasil Prediksi Random Forest

Berdasarkan ilustrasi di atas diketahui bahwa hasil prediksi pasien terdeteksi diabetes sebanyak 73 kasus dan pasien tidak terdeteksi diabetes 1127 kasus.



Gambar 4. Confusion Matriks Random Forest

Hasil statistik komprehensif menggunakan algoritma random forest dijelaskan dalam tabel di bawah ini.

Tabel 2. Statistik Komprehensif Random Forest

Overall statistics	**Nilai**
Accuracy	0.9717
95% CI	(0.9606, 0.9803)
No Information Rate	0.9392
P-Value [Acc > NIR]	1.404e-07
Kappa	0.7964

3. Evaluasi Model

Pada fase khusus ini, dilakukan pemeriksaan cermat untuk meneliti kinerja algoritma *machine learning* yang telah diterapkan. Setelah diterapkan untuk menilai penyakit diabetes untuk tujuan mendeteksi ada tidaknya diabetes pada pasien, evaluasi model yang dihasilkan dijelaskan sebagai berikut:

Tabel 3. Evaluasi Model

Algoritma Machine learning	Accuracy	Kappa
Support Vector Machine	0.9333	0.4536
Random Forest	0.9717	0.7964

SIMPULAN

Berdasarkan hasil penelitian yang dilakukan dengan menerapkan algoritma SVM dan *Random Forest* pada data biomedis mengenai penyakit diabetes, dapat disimpulkan sebagai berikut:

1. SVM dan Random Forest berhasil digunakan untuk melakukan klasifikasi pada data penyakit diabetes, dengan klasifikasi antara terdeteksi diabetes dan tidak terdeteksi diabetes.
2. Hasil prediksi dari SVM menunjukkan bahwa sebanyak 40 pasien terdeteksi diabetes dan 1160 pasien tidak terdeteksi diabetes, dengan tingkat akurasi mencapai 93.33%.
3. Sedangkan hasil prediksi dari *Random Forest* menunjukkan bahwa sebanyak 73 pasien terdeteksi diabetes dan 1127 pasien tidak terdeteksi diabetes, dengan tingkat akurasi mencapai 97.17%.
4. Evaluasi model akhir menunjukkan bahwa *Random Forest* memiliki akurasi yang lebih tinggi dibandingkan dengan SVM, dengan nilai kappa sebesar 0.7967. Hal ini menunjukkan bahwa *Random Forest* lebih efektif dalam memprediksi penyakit diabetes pada pasien dengan mempertimbangkan nilai BMI, HbA1c *level*, dan *blood glucose level*.

DAFTAR PUSTAKA

Adnan, Mohammed Sarfaraz Gani et al. 2023. "A Novel Framework for Addressing Uncertainties in Machine Learning-Based Geospatial Approaches for Flood Prediction." *Journal of Environmental Management* 326(PB): 116813. https://doi.org/10.1016/j.jenvman.2022.116813.

Agarwal, Manisha et al. 2024. "Diabetic Retinopathy Screening Guidelines for Physicians in India: Position Statement by the Research Society for the Study of Diabetes in India (RSSDI) and the Vitreoretinal Society of India (VRSI)-2023." *International Journal of Diabetes in Developing Countries* (0123456789): 1–8.

Febrian, Muhammad Exell et al. 2022. "Diabetes Prediction Using Supervised Machine Learning." *Procedia Computer Science* 216(2022): 21–30. https://doi.org/10.1016/j.procs.2022.12.107.

Harsa, Hastuadi et al. 2023. "Machine Learning and Artificial Intelligence Models Development in Rainfall-Induced Landslide Prediction." *IAES International*

Journal of Artificial Intelligence 12(1): 262–70.

Huang, Jin, Jingjing Lu, and Charles X. Ling. 2003. "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy." *Proceedings - IEEE International Conference on Data Mining, ICDM*: 553–56.

Kao, I. Feng, Jia Yi Liou, Meng Hsin Lee, and Fi John Chang. 2021. "Fusing Stacked Autoencoder and Long Short-Term Memory for Regional Multistep-Ahead Flood Inundation Forecasts." *Journal of Hydrology* 598(October 2020): 126371. https://doi.org/10.1016/j.jhydrol.2021.126371.

Liu, Qiang et al. 2022. "Discussion on the Tree-Based Machine Learning Model in the Study of Landslide Susceptibility." *Natural Hazards* 113(2): 887–911. https://doi.org/10.1007/s11069-022-05329-4.

Mujumdar, Aishwarya, and V. Vaidehi. 2019. "Diabetes Prediction Using Machine Learning Algorithms." *Procedia Computer Science* 165: 292–99. https://doi.org/10.1016/j.procs.2020.01.047.

Oikonomou, Evangelos K., and Rohan Khera. 2023. "Machine Learning in Precision Diabetes Care and Cardiovascular Risk Prediction." *Cardiovascular Diabetology* 22(1): 1–16. https://doi.org/10.1186/s12933-023-01985-3.

Ponti, Rodrigo Fernandes de Mello Moacir Antonelli. 2018. 45 Machine Learning *Machine Learning A Practical Approach on the Statistical Learning Theory*. Springer International Publishing AG. https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA.

Pradeep, K. R., and N. C. Naveen. 2018. "Lung Cancer Survivability Prediction Based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics." *Procedia Computer Science* 132: 412–20. https://doi.org/10.1016/j.procs.2018.05.162.

Singh, Awadhesh Kumar et al. 2024. "A Randomized, Double-Blind, Active-Controlled Trial Assessing the Efficacy and Safety of a Fixed-Dose Combination (FDC) of MEtformin Hydrochloride 1000 Mg ER, SItagliptin Phosphate 100 Mg, and DApagliflozin Propanediol 10 Mg in Indian Adults with Type 2 Diabetes: The MESIDA Trial." *International Journal of Diabetes in Developing Countries* (0123456789). https://doi.org/10.1007/s13410-024-01321-9.

Tang, Xianzhe et al. 2020. "Flood Susceptibility Assessment Based on a Novel Random Naïve Bayes Method: A Comparison between Different Factor Discretization Methods." *Catena* 190(March): 104536. https://doi.org/10.1016/j.catena.2020.104536.

Tehrani, Faraz S. et al. 2022. 114 Natural Hazards *Machine Learning and Landslide Studies: Recent Advances and Applications*. Springer Netherlands. https://doi.org/10.1007/s11069-022-05423-7.

Tural Buyuk, Esra, Hatice Uzsen, Merve Koyun, and Reyhan Dönertaş. 2023. "Parental Monitoring Status of the Children with Type 1 Diabetes Mellitus (DM)

and Their Quality of Life." *International Journal of Diabetes in Developing Countries* (Dm). https://doi.org/10.1007/s13410-023-01304-2.

Zhang, Lu et al. 2023. "Differences in Nutrition, Handgrip Strength, and Quality of Life in Patients with and without Diabetes on Maintenance Hemodialysis in Xi'an of China." *International Journal of Diabetes in Developing Countries* (0123456789). https://doi.org/10.1007/s13410-023-01282-5.